Molecular Dynamics with the United-Residue Model of Polypeptide Chains. II. Langevin and Berendsen-Bath Dynamics and Tests on Model α-Helical Systems

Mey Khalili[†], **Adam Liwo**[†,‡], **Anna Jagielska**[†], and **Harold A. Scheraga**[*,†]

[†] *Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301*
[‡] *Faculty of Chemistry, University of Gda sk, Sobieskiego Str. 18, 80-952 Gda sk, Poland*

Abstract

The implementation of molecular dynamics (MD) with our physics-based protein united-residue (UNRES) force field, described in the accompanying paper (Khalili et al. *J. Phys. Chem. B* **2005**, *109*, 13785), was extended to Langevin dynamics. The equations of motion are integrated by using a simplified stochastic velocity Verlet algorithm. To compare the results to those with all-atom simulations with implicit solvent in which no explicit stochastic and friction forces are present, we alternatively introduced the Berendsen thermostat. Test simulations on the Ala_{10} polypeptide demonstrated that the average kinetic energy is stable with about a 5 fs time step. To determine the correspondence between the UNRES time step and the time step of all-atom molecular dynamics, all-atom simulations with the AMBER 99 force field and explicit solvent and also with implicit solvent taken into account within the framework of the generalized Born/surface area (GBSA) model were carried out on the unblocked Ala_{10} polypeptide. We found that the UNRES time scale is 4 times longer than that of all-atom MD simulations because the degrees of freedom corresponding to the fastest motions in UNRES are averaged out. When the reduction of the computational cost for evaluation of the UNRES energy function is also taken into account, UNRES (with hydration included implicitly in the side chain–side chain interaction potential) offers about at least a 4000-fold speed up of computations relative to all-atom simulations with explicit solvent and at least a 65-fold speed up relative to all-atom simulations with implicit solvent. To carry out an initial full-blown test of the UNRES/MD approach, we ran Berendsen-bath and Langevin dynamics simulations of the 46-residue B-domain of staphylococcal protein A. We were able to determine the folding temperature at which all trajectories converged to nativelike structures with both approaches. For comparison, we carried out ab initio folding simulations of this protein at the AMBER 99/GBSA level. The average CPU time for folding protein A by UNRES molecular dynamics was 30 min with a single Alpha processor, compared to about 152 h for all-atom simulations with implicit solvent. It can be concluded that the UNRES/MD approach will enable us to carry out microsecond and, possibly, millisecond simulations of protein folding and, consequently, of the folding process of proteins in real time.

1. Introduction

In the accompanying paper,[1] we derived the Lagrange equations of motion for the united-residue (UNRES) force field for protein simulations.[2–16] We applied the velocity Verlet algorithm to integrate the equations of motion and tested the symplectic properties of the algorithm by running microcanonical simulations on the Ala_{10} polypeptide. The algorithm

* Author to whom correspondence should be addressed. Phone: (607) 255 4034. Fax: (607) 254 4700. E-mail: has5@cornell.edu.

appears to be stable up to about a 1 fs time step; however, some energy drift appears because of instability of the gradient of the present side-chain-rotamer (U_{rot}) contribution to UNRES; this component must ultimately be replaced with one leading to stable gradients. To reduce the energy drift and, thereby, to increase the time step further, we developed a variable-time-step algorithm in which the time step is scaled down if the change in acceleration exceeds a cutoff value. With this modification, the algorithm is stable up to about a 2.5 fs time step.

In this paper, we introduce Langevin dynamics with the UNRES force field to take into account the dynamic effects of the solvent and also to couple the system under study to a thermal bath. To compare the results of the simulations with those of all-atom studies reported here and in the literature,[17] which were carried out at the all-atom level with implicit solvent, we also introduce the Berendsen thermostat, where the friction and random forces are present only implicitly. We carry out test simulations on the Ala_{10} polypeptide and the B-domain of staphylococcal protein A, a 46-residue protein with a three-helix-bundle topology, and compare the results with those of all-atom simulations to determine the relation between the UNRES and the all-atom time scale.

2. Methods

2.1. Langevin Dynamics with the UNRES Force Field

The UNRES force field is outlined in the accompanying paper[1] and also described in detail in our earlier work.[2–16] Briefly, a polypeptide chain is represented as a sequence of α-carbon atoms (C^{α}'s). The C^{α}'s are linked together by virtual bonds (designated as dCs), which constitute the backbone. The interaction sites are the united peptide groups (p's) in the middle of the dCs, and the united side chains at the ends of virtual bonds are designated as dXs. The p centers represent only the C′, O, N, and H atoms of the peptide groups, while the C^{α} atoms are included in the SC centers. Consequently, the positions of the C^{α} atoms are only geometric points. The energy of a united-residue chain is derived as a restricted free energy (RFE) function (or potential of mean force (PMF)) obtained by computing the free energy corresponding to a given coarse-grain conformation by integrating out those degrees of freedom that are not considered in the UNRES model.[7,13] The force field has been optimized by our novel hierarchical-optimization method[11,14–16] in which the energy of a training protein is required to decrease with native likeness increasing in an ordered way that corresponds to the experimentally observed folding pathway, if such information is available from experiment. In this work, we use the force field derived using 1IGD (a 61-residue $\alpha + \beta$-protein);[15] this force field is referred to as the F2 force field.

In the accompanying paper,[1] we derived the Lagrange equations of motion for the UNRES model. We chose the components of the $C^{\alpha}\cdots C^{\alpha}$ and $C^{\alpha}\cdots SC$ virtual-bond vectors as the generalized coordinates, the vector of generalized coordinates being defined as $\mathbf{q} = (\mathbf{dC}_o,$ $\mathbf{dC}_s, ..., \mathbf{dC}_{e-1}, \mathbf{dX}_{i_1}, ..., \mathbf{dX}_{i_m})^{\mathbf{T}}$, where $\mathbf{dC}_i$ denotes the vector pointing from C_i^{α} to C_{i+1}^{α} and $\mathbf{dX}$ denotes the vector pointing from C_i^{α} to SC_i with $\mathbf{dC}_o$ defining the absolute position of the first C^{α} atom, $i_1, i_2, ..., i_m$ denote the numbers (m) of non-glycine residues, while s and e are the number of the first and the last nondummy C^{α} atom, respectively. The entries corresponding to glycine residues are omitted from the list of $\mathbf{dX}$s because, for a glycine residue, the position of its "side chain" coincides with that of the respective C^{α} atom.[2] The "dummy" C^{α} atoms are introduced if the first (last) residue is not glycine and if the chain is unblocked;[2] this is necessary to define the local reference frame for the first (last) side chain. If the first residue is a dummy, then $s = 2$; otherwise $s = 1$. If the last residue is a dummy, then $e = n - 1$; otherwise $e = n$. We assumed that the virtual bonds are flexible rods with uniformly distributed masses. The number of $\mathbf{dX}$ vectors is equal to the number of residues minus the number of glycine residues in a chain. The Cartesian coordinates of the interacting sites $\mathbf{x} = (\mathbf{x}_{p_s}, \mathbf{x}_{p_s+1}, ...\mathbf{x}_{p_e}-1, \mathbf{x}_{SC_s},$

$\mathbf{x}_{SC_s+1}, \ldots, \mathbf{x}_{SC_e})^{\mathrm{T}}$ and their velocities $\mathbf{v} = (\mathbf{v}_{p_s}, \mathbf{v}_{p_s+1}, \ldots \mathbf{v}_{p_e-1}, \mathbf{v}_{SC_s}, \mathbf{v}_{SC_s+1}, \ldots, \mathbf{v}_{SC_e})^{\mathbf{T}}$ are related to the generalized coordinates $\mathbf{q}$ and generalized velocities $\dot{\mathbf{q}}$ by eqs 1 and 2, respectively.[1] It should be noted that all side chains (including the "side chains" of glycine residues that are centered at the C^α atoms) are included in the vectors $\mathbf{x}$ and $\mathbf{v}$.

$$x = Aq \tag{1}$$

$$v = A\dot{q} \tag{2}$$

with $\mathbf{A}$ being the transformation matrix from the generalized coordinates ($\mathbf{q}$) to the Cartesian coordinates ($\mathbf{x}$) of the sites. The matrix $\mathbf{A}$ is defined by eq 27 of the accompanying paper.[1]

On the basis of eq 9 and eqs 19–25 of the accompanying paper,[1] we can write the Langevin equations for the UNRES model as eq 3.

$$G\ddot{q} = - \nabla_q U(q) + f^{fric} + f^{rand} \tag{3}$$

with the inertia matrix $\mathbf{G}$ defined by eq 4.

$$G = A^{T}MA + H \tag{4}$$

where U is the UNRES potential energy, $\mathbf{f}^{fric}$ and $\mathbf{f}^{rand}$ are the friction and random forces, respectively, $\mathbf{M}$ (defined by eq 28 of the accompanying paper[1]) is the diagonal matrix of the masses of the interacting sites, and $\mathbf{H}$ (defined by eq 29 of the accompanying paper[1]) is the part of the inertia matrix that corresponds to the internal (stretching) motions of the virtual bonds.

We use Stokes law[18] to compute the friction forces. In this work, we assume that each interacting site behaves as a sphere moving through water. Of course, a more refined treatment should take into account at least the ellipsoidal shape of a site, but we leave this to further work. Further, we assume that the friction force is proportional to the solvent-accessible surface area of a site, following the suggestion of other authors.[19] Thus, the friction force acting on site z is expressed by eq 5.

$$f_z = - \gamma_z v_z \tag{5}$$

with

$$\gamma_z = 6\pi(r_z + r_{wat})\eta_{wat} \max\{S_z/(4\pi r_z^2), 0.1\}\alpha \tag{6}$$

where γ_z is the friction coefficient of site z, $\mathbf{v}_z$ is the velocity vector of site z, r_z is the effective radius of site z, r_{wat} is the effective radius of water (taken in this study as 1.4 Å), η_{wat} is the viscosity of water equal to 0.8904 cpoise at 298 K, S_z is the solvent-accessible surface area of site z, and α is a scaling factor. We used the radii computed by Levitt[20] for the radii of the side chains, while the effective radius of the peptide group was set at 5 Å, based on the van der Waals radius of the peptide group estimated from high-level ab initio quantum-mechanical calculations of the energy surfaces of interacting peptide groups.[13] We adapted the algorithm[21,22] from the TINKER package[23,24] to calculate the surface area. Because the surface area of the UNRES sites often happens to decrease to zero, we set a lower limit of 0.1 on the ratio of the solvent-accessible surface area of a site to its total surface area (eq 6). The scaling factor α is introduced to reduce the friction and random forces, which would otherwise be more than an order of magnitude higher than the forces of the potential. From the work of

other authors (see, e.g., the Appendix in [ref 25]), α should range from 0.001 (low-friction limit) to 0.1 (overdamped limit).

On the basis of eq 1, we can convert the forces expressed in the Cartesian coordinates of the interacting sites, $\mathbf{f_x}$, into forces expressed in the generalized coordinates, $\mathbf{f_q}$, by using eq 7.

$$f_q = A^T f_x \tag{7}$$

Through the use of eqs 2 and 7, we can convert the friction forces expressed in the Cartesian coordinates of the sites (eq 5) into the friction forces expressed in generalized coordinates

$$f^{fric} = -A^T \Gamma A \dot{q} \tag{8}$$

The friction matrix Γ is a diagonal matrix (eq 9) whose elements are calculated from Stokes law[18] modified by the dependence of the friction coefficients on the solvent-accessible surface area[19] according to eq 6.

$$\Gamma = \begin{pmatrix} \gamma_p 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \gamma_p 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_{SC_s} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \gamma_{SC_{s+1}} 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & \gamma_{SC_s} 1 \end{pmatrix} \tag{9}$$

where $\mathbf{1}$ is a 3×3 identity matrix and $\mathbf{0}$ is a 3×3 matrix composed of zeros. Thus, even though the friction acting on individual sites is assumed to be isotropic, the friction matrix ($\mathbf{A^T \Gamma A}$) in the generalized coordinates $\mathbf{q}$ is not diagonal.

The random forces acting on interaction sites are δ-correlated forces, i.e., $\text{cov}[f^{rand}(t_1), f^{rand}(t_2)] \propto \delta(t_2 - t_1)$ at times t_1 and t_2 of the trajectory (where cov denotes the covariance and δ the Kronecker δ function), which are expressed at every integration step by eq 10.[25–27]

$$f_z^{rand} = \sqrt{\frac{2\gamma_z RT}{\delta t}} N(0,1) \tag{10}$$

where f_z^{rand} is the vector of random forces acting on site z, γ_z is the friction coefficient of this site (defined by eq 6), R is the universal gas constant, T is the absolute temperature, δt is the integration time step, and $\mathbf{N}(0,1)$ is the three-dimensional vector whose components are sampled independently from a normal distribution with zero mean and unit variance. Together, the stochastic and friction forces constitute a thermostat that maintains the average temperature at the preset value.

Again, by using eq 7, we can transform the random forces from the Cartesian coordinates of the interacting sites to the generalized coordinates, which yields eq 11.

$$f^{rand} = \sqrt{\frac{2RT}{\delta t}} A^T \Gamma^{1/2} N(0,1) \tag{11}$$

2.2. Integrating the Langevin Equations of Motion by the Velocity Verlet Algorithm

A variety of stochastic integrators have been developed for Langevin dynamics, most of them being stochastic versions of of the leapfrog[28–31] and the velocity Verlet[32–34] algorithm.

These algorithms are based on stochastic differential equation theory and perform analytical integration of the friction and the stochastic forces in a single time step. An elegant way of deriving the algorithms for Langevin dynamics makes use of the Liouville propagator and the Suzuki formula[35] for time-ordered exponentials; this is the basis of the algorithms developed by Ciccotti and co-workers.[36,37] Their formulas are slightly different from those developed earlier.

In the Appendix, we describe in detail our adaptation of the procedure from the TINKER package[23,24] described in refs 32 and 33. However, we note that the stochastic integrator from TINKER as well as all stochastic integrators described in the papers cited above involve a tremendous increase of computational cost compared to microcanonical calculations with UNRES. The reason for this is that the exponentials of the friction matrix must be calculated, as opposed to the case of diagonal equations of motion when exponentials of the friction coefficients (scalars) are calculated. For not too large friction coefficients and not too large time steps, which is the case of our work, the matrix exponentials can be expanded into a power series, and only the first-order terms can be considered without loss of accuracy. If this is done and if all terms with δt^2 and higher are ignored in the expressions for new velocities and all terms with δt^3 and higher in the expressions for new coordinates, then the algorithm from TINKER[32,33] and the algorithm of Ciccotti et al.[36,37] reduce to the velocity Verlet algorithm with stochastic and friction forces included. This algorithm can be summarized as follows:

Step 1 (updating coordinates):

$$q(t+\delta t) = q(t) + \dot{q}(t)\,\delta t + \frac{1}{2}[\ddot{q}\prime(t) + \ddot{q}^{\mathrm{fric}}(t) + \ddot{q}_x^{\mathrm{rand}}(t)]\delta t^2 \tag{12}$$

Step 2 (updating velocities):

$$\dot{q}(t+\delta t) = \dot{q}(t) + \left\{\frac{1}{2}[\ddot{q}\prime(t) + \ddot{q}\prime(t+\delta t)] + \ddot{q}^{\mathrm{fric}}(t) + \ddot{q}_v^{\mathrm{rand}}(t)\right\}\delta t \tag{13}$$

with

$$\ddot{q}\prime(t) = -G^{-1}\nabla_q U[q(t)] \tag{14}$$

$$\ddot{q}\prime(t+\delta t) = -G^{-1}\nabla_q U[q(t+\delta t)] \tag{15}$$

$$\ddot{q}^{\mathrm{fric}}(t) = -G^{-1}A^T\Gamma A\dot{q}(t) \tag{16}$$

$$\ddot{q}^{\mathrm{rand}}(t) = -G^{-1}f^{\mathrm{rand}}(t) \tag{17}$$

and the subscripts "**x**" and "**v**" at $\ddot{q}^{\mathrm{rand}}$ indicating that the random forces are sampled independently to compute the new coordinates and velocities, respectively.

The simplified stochastic velocity Verlet algorithm (eqs 2 and 13) does not involve the computation of matrix exponentials and is, therefore, much more practical than the original integrators for the case of nondiagonal equations of motion implied in UNRES dynamics. We found that the simplified algorithm does not differ in performance (i.e., maintaining the average temperature, the value of the variance of the potential and the kinetic energy, and the velocity autocorrelation function) from the original algorithm from TINKER[32,33] and the algorithm introduced by Ciccotti.[36,37]

2.3. Berendsen's Thermostat

To carry out canonical simulations in the simplest possible manner, we implemented Berendsen's method to couple the system to a thermal bath.[38] This method is based on the Langevin equations, which are solved approximately at a given time step. Instead of using the simplified stochastic velocity Verlet algorithm (eqs 12−17), the velocities computed in the second step of the regular velocity Verlet algorithm (eq 31 of [ref 1]) are scaled by the factor λ defined by eq 18.

$$\lambda = \left[1 + \frac{\delta t}{\tau_T} \left(\frac{T_o}{T(t)} - 1 \right) \right]^{1/2}$$

(18)

where δt is the time step, T_o is the desired temperature, and $T(t)$ is the instantaneous temperature of the system at time t, which is calculated from eq 19.

$$T(t) = \frac{2K(t)}{RN_f}$$

(19)

where $K(t)$ is the total kinetic energy of the system at time t, R is the universal gas constant, and N_f is the number of degrees of freedom in the system, which is equal to $3n_p + 3n_{SC} - 3n_{Gly}$ where n_p is the number of peptide groups, n_{SC} is the number of side chains, and n_{Gly} is the number of glycine residues in the system. The parameter τ_T is a measure of the strength of the coupling to the thermal bath. For values of τ_T much larger than δt, the simulation is reduced to a microcanonical ensemble. It was shown[38] that, for small values of τ_T relative to δt, the fluctuations in the kinetic energy are reduced at the expense of an increase in the fluctuations of the potential and total energy. Although small values of τ_T influence fluctuations, average thermodynamic quantities are not disturbed even for very small τ_T values.[38] However, the fluctuations of global properties are strongly influenced for small coupling time constants, and their intensity cannot be used to derive thermodynamic properties.

At the end of every full velocity Verlet step, the velocities are scaled, following eq 20.

$$\dot{q}(t+\delta t) \leftarrow \lambda \dot{q}(t+\delta t)$$

(20)

The advantage of the Berendsen algorithm is that the coupling strength can be made as weak as possible to reduce the disturbance of the system. This algorithm is numerically stable, and truncation errors will not develop undesired deviations that would require further correction. This is especially important for long runs and is also fairly easy to implement. A disadvantage of the Berendsen algorithm is that it does not always give correct values of thermodynamic quantities.

An undesirable feature of the Berendsen-type runs is that, as the molecule relaxes (i.e., loses its potential energy), the kinetic energy is converted into net rotation of the whole molecule. To overcome this problem, we reset the angular momentum of the molecule to zero every 1000 time steps.

2.4. All-Atom Molecular Dynamics Simulations of the Ala₁₀ and Protein A System

For ultimate comparison of the time step in all-atom and UNRES calculations, we carried out all-atom molecular dynamics (MD) simulations of the unblocked Ala10 polypeptide with the AMBER 99 force field.[39,40] A series of nine simulations was carried out in a periodic box of water with 3155 water molecules in the *NPT* ensemble (constant number of particles, pressure, and temperature) at $T = 298$ K and $P = 1$ atm. An 8-Å cutoff was applied to nonbonded interactions, and the particle-mesh Ewald summation was applied to evaluate electrostatic

interactions. The TIP3P[41] model of water was used. The SHAKE[42,43] algorithm was used to maintain the valence geometry of the water molecules.

To compare the time scale of UNRES/MD with that of implicit solvent all-atom MD, we ran a series of 10 MD all-atom simulations on the unblocked Ala_{10} polypeptide with the AMBER 99 force field with the solvent treated with the generalized Born model.[44,45] The system was coupled to a thermal bath by using the Berendsen algorithm[38] with the coupling parameter $\tau_T = 0.1$ ps. The simulations were carried out at a temperature $T = 400$ K, and the integration time step was $\delta t = 2$ fs. A distance cutoff of 20 Å was applied to evaluate all nonbonded interactions in both series of all-atom simulations.

As a second set of calculations to characterize the time step with the UNRES model, we carried out all-atom MD simulations of the 46-residue N-terminal fragment of the B-domain of staphylococcal protein A[46] with the sequence Ac-Gln-Gln-Asn-Ala-Phe-Tyr-Glu-Ile-Leu-His-Leu-Pro-Asn-Leu-Asn-Glu-Glu-Gln-Arg-Asn-Gly-Phe-Ile-Gln-Ser-Leu-Lys-Asp-Asp-Pro-Ser-Gln-Ser-Ala-Asn-Leu-Leu-Ala-Glu-Ala-Lys-Lys-Leu-Asn-Asp-Ala-NHMe; the AMBER 99/GBSA force field at temperatures $T = 300, 320, 350, 400, 420, 450,$ and 500 K, with a coupling parameter $\tau_T = 1$ ps, was used. The other settings were the same as for all-atom simulations of Ala_{10} with implicit solvent. Five trajectories, starting from a fully extended chain, were run for each temperature, and the duration of a run was 20 ns.

3. Results and Discussion

3.1. Equilibrium and Folding Simulations of the Ala_{10} Polypeptide

To carry out initial tests of the UNRES/MD approach to protein folding, we ran simulations on the Ala_{10} polypeptide, using the Berendsen thermostat and full-blown Langevin equations, respectively. To run simulations with the Berendsen thermostat, we had to choose the value of the coupling parameter τ_T. In all-atom simulations, the coupling parameter τ_T of eq 18 usually ranges from 0.5 to 5 ps, and even values from 0.1 ps and higher result in stable trajectories[38] and reliable values of thermodynamic properties. However, this information cannot be implemented directly in the UNRES/MD approach because the time scale of a united-residue chain need not be the same as for an all-atom chain. We, therefore, ran a series of test simulations on the Ala_{10} polypeptide to determine the plausible value of τ_T. The global minimum of the UNRES potential energy with the F2 force field of [ref 15] is a full right-handed α-helix with potential energy equal to −44.8 kcal/mol. We wanted to satisfy three criteria: (i) the fluctuations of the kinetic, potential, and total energy should be reasonably small, (ii) the folding time should be short, and (iii) the native structure should be reasonably stable. Criteria i and iii somehow overlap because, with increasing energy fluctuations, the system must leave the native structure frequently and visit different regions of conformational space. Criterion ii ensures that the coupling is strong enough because, for large values of τ_T, we approach the microcanonical mode in which Ala_{10} does not fold fully because of accumulation of too much kinetic energy as the potential energy decreases.

The first simulations were started from the α-helical conformation of Ala_{10} (with $\theta = 90°$, $\gamma = 45°$, $\alpha = 120°$, and $\beta = -120°$ for all residues). We ran simulations at $T = 300$ K with the time steps 0.01 mtu (0.489 fs), 0.02 mtu (0.978 fs), 0.04 mtu (1.956 fs), and 0.1 mtu (4.89 fs). The duration of a simulation was 4000 mtu (196 ps). Ten trajectories were run for each value of τ_T and each time step. The standard deviations of the potential, kinetic, and total energy were computed after discarding the initial 10 ps equilibration period. The standard deviations (σ) of the energies as well as the actual temperature averaged over all trajectories for a given time step and a given τ_T are plotted versus τ_T in Figure 1. Figure 1 shows that, for small values of τ_T, the fluctuations in the kinetic energy are small, while those in the potential and total energy are large. For larger values of τ_T, the fluctuations in the kinetic energy increase while those of

the potential energy reach minima at τ_T between 5 and 10 mtu (Figure 1A). The depth of these minima increases with the time step, which strongly suggests that they are caused by energy drift due to integration errors; this energy drift is remarkable when the coupling is weak, and consequently, the system approaches the behavior of a microcanonical system. Therefore, τ_T larger than 10 mtu should not be used with larger time steps. It can also be noted that, for the largest time step (0.1 mtu), the average temperature deviates significantly from the set value for $\tau_T > 5$ mtu (Figure 1B); therefore, with this large time step, one should be careful to introduce tight coupling (i.e., set a small value of τ). The fluctuations of the kinetic, potential, and the total energy become comparable for $\tau_T = 5$ mtu (0.2445 ps); for all-atom simulations, this is a good criterion for selection of τ_T, because one can obtain reliable values of observables and their fluctuations.[38] However, we must bear in mind that there is no explicit solvent in UNRES, and therefore, in reality, the potential and the total energy can fluctuate more because the molecule exchanges energy with the surrounding solvent, which has to be simulated by the Berendsen thermostat.

The next series of simulations to estimate a plausible value of τ_T were the folding simulations of the Ala$_{10}$ polypeptide, starting from the fully extended chain ($\theta = 110°$, $\gamma = 180°$, $\alpha = 110°$, and $\beta = -120°$ for all residues). We ran 1 000 000-step (1956 ns) simulations at $T = 300$ K with $\delta t = 0.04$ mtu (1.956 fs) and for $\tau_T = 0.1, 0.5, 1.0, 2.0,$ and 5.0 mtu. Ten trajectories were run for each τ_T value. To monitor the progress of folding, we computed the C^α root-mean-square deviation (rmsd) from that of the Ala$_{10}$ all-atom chain in an ideal α-helix conformation minimized with the ECEPP/3 all-atom force field.[47] Because the ends of the chain are subject to much greater fluctuations than those of the inner residues, only residues from 2 to 9 were superposed in computing the rmsd; we found that this provided a much clearer analysis of the folding trajectory compared to the rmsd over all residues. The analyzed rmsd will hereafter be referred to as ρ_α.

To characterize the folding of Ala$_{10}$, we followed our earlier work on force-field optimization[14] and, for trajectory i, defined the folding time, τ_f^i, as the time when ρ_α first decreased below 1 Å and the stability time, τ_s^i, as the total length of those parts of a trajectory after τ_f where ρ_α was not larger than 2 Å. For trajectories in which the peptide did not fold (i.e., ρ_α did not decrease below 1 Å), τ_f was considered to take on the length of a trajectory. We also defined the stability (τ_s/t) as the ratio of τ_s to the length of the trajectory (t).

The average folding (τ_f) and stability (τ_s) times (over 10 trajectories) are plotted against the coupling constant τ_T in Figure 2, while sample plots of the potential energy, temperature, and ρ_α for five representative trajectories corresponding to $\tau_T = 1$ mtu are shown in Figure 3. It can be seen that the potential energy is always higher by about 15 kcal/mol than that of the global minimum found by the conformational space annealing (CSA) method[48] with the force field used here;[15] this occurs because the system is at finite temperature and cannot attain the lowest potential energy because of thermal motions introduced through coupling to the Berendsen thermostat.

It can be seen (Figure 2) that τ_f increases rapidly with τ_T after $\tau_T = 1$ mtu and decreases slowly with τ_T for smaller τ_T. It can, therefore, be concluded that the coupling is too weak for $\tau_T > 1$ mtu for the system to fold quickly because of too slow exchange of the potential energy with the environment. The stability time varies with τ_T in the opposite direction, decreasing significantly below $\tau_T = 1$ mtu. The decrease of the stability time for smaller times is caused by too pronounced coupling to the heat bath and, consequently, too large fluctuations of the potential energy. Therefore, the value of τ_T at 1 mtu (0.0489 ps) appears to be the optimal one, and we used it in further simulations reported in this paper.

To estimate the plausible values of the scaling factor of the friction coefficient in Langevin simulations, we ran the folding simulations of Ala_{10} with various scaling factors. By carrying out a similar analysis as for the Berendsen-type simulations, we found that the safe values of the scaling factor for the time step of 0.1 mtu (4.89 fs) range from 0.01 to 0.20. For smaller and higher values, the average temperature starts to deviate by more than 5 K from T_o, the value specified in the simulation. This is understandable because, with low friction coefficients, the coupling is too weak (as with too large τ_T values in Berendsen-type simulations), while too large friction results in instability of the integration algorithm because of too large friction forces.[27] Sample plots of the potential energy, temperature, and ρ_α for a representative trajectory corresponding to $\tau_T = 1$ mtu are shown in Figure 4.

Figure 5 shows the dependence of the folding time on the scaling factor of the friction coefficient. It can be seen that the folding time stays between 200 and 400 ps until the scaling factor reaches 0.1, then the folding time starts to grow approximately linearly with the scaling factor, as observed by Cieplak and co-workers.[27,49] It should be noted that, for each α, τ_f and τ_s sum up to about 4.9 ns (the length of the simulation). This means that the peptide stays folded after it folds for the first time, as opposed to the results of Berendsen-bath simulations (Figure 2). It should be noted (Table 1) that the average folding time (411 ps) is greater than for simulations with the Berendsen thermostat (52 ps), which is caused by the explicit presence of the friction forces opposing folding.

3.2. Relationship between the UNRES Time Scale and the Time Scale of All-Atom Molecular Dynamics

As established in the accompanying paper,[1] the safe value of the time step (δt) in integrating the equations of motion for microcanonical simulations is 0.02 mtu with a fixed time step or 0.05 mtu with a variable time step; for canonical simulations, the time step can be increased to 0.10 mtu. Even this larger time step corresponds to about 5 fs. However, because fast motions of the secondary degrees of freedom are averaged out in UNRES, it is certain that the time scale of molecular dynamics with UNRES is longer than that of all-atom molecular dynamics.

To estimate the relationship between the time scale of all-atom molecular dynamics and the time scale of molecular dynamics with UNRES, we carried out two series of folding simulations on the Ala_{10} polypeptide using all-atom dynamics with the AMBER 99 force field[39,40] and with UNRES dynamics. The all-atom simulations were carried out with both explicit and implicit solvent for 2 ns. Ten independent trajectories were run for implicit solvent simulations and seven for the simulations with explicit water molecules. The simulations run with explicit solvent were compared with the UNRES Langevin dynamics simulations (in both cases the friction is present either by inclusion of explicit solvent molecules or by considering friction forces), while the all-atom GBSA simulations were compared with UNRES simulations with the Berendsen thermostat (no explicit friction forces being present in both cases). To provide a fair comparison when all-atom GBSA is used, we determined the optimal folding temperature and τ_T value in all-atom simulations with implicit solvent. These were $T = 400$ K and $\tau_T = 0.1$ ps, respectively. The temperature was 300 K in both UNRES Langevin and Berendsen thermostat simulations. We set the scaling factor of the friction coefficient at 0.05 in UNRES Langevin simulations; this value appears in the interval where the folding time varies little with the friction coefficient (Figure 5).

Snapshots from representative all-atom simulations with explicit solvent, with solvent treated at the GBSA level, and UNRES (Langevin and Berendsen) trajectories are shown in Figures 6A–D. The maximum, minimum, and average values of τ_f and τ_s for all four approaches are summarized in Table 1. It can be seen that Ala_{10} folds to the α-helical conformation after about 1.8 ns of all-atom simulations with explicit water and about 0.4 ns of UNRES Langevin simulations on average. Similarly, all-atom implicit solvent simulations required about 0.2 ns

on average, while UNRES/MD with the Berendsen thermostat required about 0.05 ns on average. This suggests that the unit of the time step of UNRES dynamics corresponds to about 4 units of all-atom molecular dynamics with explicit solvent. As mentioned in the accompanying paper,[1] this is a result of averaging out the fast-moving degrees of freedom when passing to the UNRES from the all-atom representation. The mismatch between the UNRES and all-atom time scales can also be linked to the internal friction,[26] which has long been known in polymer physics. The presence of internal friction slows down the internal motion of a polymer chain more than expected based on the viscosity of the surrounding solvent. Part of internal friction is present even when a polypeptide chain is modeled on the all-atom level and the solvent on the mean-field level, because the forces acting on a polymer chain in a solvent cannot be neatly separated into the conservative forces acting within the polypeptide chain and the dissipative forces arising from the bulk of the solvent.[50] However, when the motions of an all-atom chain and an UNRES chain are compared, the main difference is likely to arise from the fact that the secondary degrees of freedom (not considered in UNRES) move faster than the coarse-grain (UNRES) degrees of freedom. The coupling of the fast motions to the slow motions of the UNRES degrees of freedom caused the latter to acquire a quasi-random component. (We should bear in mind that an "UNRES chain" physically corresponds to a collection of all-atom chains with the same coarse-grain geometry and velocities of the UNRES sites.) This coupling will, therefore, slow the net motion of the coarse-grain degrees of freedom. As stated in the accompanying paper,[1] the coupling of the motion of the secondary to that of the primary degrees of freedom is not taken into account in our present approach, which explains the observed mismatch between the UNRES and the all-atom time scales.

Even more favorable is the ratio of the computer time required to fold the Ala_{10} polypeptide with the all-atom and the UNRES representations: 1 ns of all-atom folding simulations (necessary to fold Ala_{10}) required 10 000 CPU min per single Alpha processor with explicit water molecules and 150 CPU min with implicit solvent. In contrast to this, 1 ns of UNRES simulations (both Langevin and Berendsen) takes 9.3 CPU min. Because, as demonstrated earlier in this section, the UNRES time scale is 4-fold longer than the all-atom time scale, this means that molecular dynamics simulations with the UNRES force field (which is based fully on the physics of interactions in proteins) offer a $4 \times (10\,000/9.3) \approx 4000$-fold and $4 \times (152/9.3) \approx 65$-fold speedup relative to the all-atom MD simulations with explicit and implicit solvent, respectively. Therefore, microsecond if not millisecond times now seem to be within reach.

3.3. Simulations of the Folding of the 46-Residue N-Terminal Fragment of the B-Domain of Staphylococcal Protein A

The Ala_{10} system is a good test case because computations in real time can be run even in the all-atom scheme with explicit water molecules. However, we are able to study only secondary-structure formation for this decapeptide. Assembling larger fragments rather than formation of secondary structure is a bottleneck of the folding process. We, therefore, ran canonical simulations on protein A, a 46-residue system. Recent work of Jang et al.[17] demonstrated that this system folds using an all-atom approach with implicit solvent. We, therefore, studied protein A at both the all-atom (AMBER 99 + GBSA) and the UNRES levels, the UNRES simulations being carried out with both the Berendsen thermostat and the Langevin dynamics.

Jang et al.[17] carried out their runs at 400 K and claimed that this temperature was above the folding temperature of the system. We do not share their view for a number of reasons. First, the Berendsen thermostat was used by Jang et al.[17] rather than the regular Langevin approach. The stochastic forces that are supposed to kick the system out of a low-energy region in which it is stuck were present only in an implicit manner through temperature scaling. The resulting equations are in fact deterministic, and the stochastic factor is introduced only once when

generating the initial velocities. It is, therefore, not clear that a given temperature will correspond to the forces due to the kicks from the solvent molecules characteristic of that temperature. Second, the energy scale of a force field might be different from the physical energy scale. This is especially true for UNRES, which, in its present version, was ultimately optimized to maximize the energy gaps between the conformations with increasing degree of native likeness.[14] No information from folding thermodynamics has yet been introduced in optimization except that the initial parameters of the potential of side-chain interactions ($U_{SC_iSC_j}$) were scaled[3] to reproduce the partition coefficients of the amino acids between water and n-octanol.[51] However, this constraint was not retained in the optimization of the force field. We, therefore, determined the folding temperature of protein A as a temperature with minimum τ_f for both the all-atom and the UNRES simulations. We ran five all-atom trajectories and 10 UNRES trajectories per temperature. The settings of the all-atom simulations are described in section 2.4, while, for UNRES, $\delta t = 0.10$ mtu (4.89 fs), $\tau_T = 1.0$ mtu (0.0489 ps), and the duration of a run was 20 ns.

We defined τ_f and τ_s as in section 3.1. For UNRES, we used the F2 force field of [ref 15]; with this force field, the lowest-energy structure of protein A has a 3.2 Å rmsd from the experimental structure (PDB code: 1BDD[46]). The rmsd was computed from all C^α atoms of the experimental structure of protein A, and the rmsd cutoff to consider a structure folded was 4 Å; the structure was considered to remain in the native basin if the rmsd did not exceed 5 Å.

The plots of the folding and stability versus temperature for both all-atom and UNRES runs are shown in Figure 7. It can be seen that the folding temperature is 420 K for the all-atom (Figure 7B) and 500 K for the UNRES (Figure 7A) approach. Plots of the potential energy, temperature, and rmsd for a representative UNRES trajectory with the Berendsen thermostat at the folding temperature are shown in Figure 8, while snapshots from a representative UNRES trajectory are shown in Figure 9. The corresponding plots obtained in UNRES Langevin dynamics at 500 K with the 0.05 scaling factor of the friction forces are shown in Figures 10 and 11, respectively. The average folding time over 10 trajectories is 4.2 ns for UNRES Langevin dynamics simulations. It can be seen that, regardless of the mode of simulation, the C-terminal α-helix is formed first as found in recent theoretical simulations[52–54] and which also is in agreement with experimental studies of the relative stability of helices.[55,56] The C-terminal α-helix appears stable throughout a trajectory. It should be noted, though, that earlier theoretical simulations[57] suggested that the N-terminal and the middle α-helices are formed first and recent analysis of experimental Φ values suggests that the C-terminal α-helix is poorly formed in the transition state and that the folding starts from the middle α-helix.[58] The difference between Berendsen and Langevin simulations is that, in the latter, packing of helices occurs more slowly; this is an explicit effect of introducing friction, which opposes motion more as the larger parts of a polymer chain move simultaneously in a given direction.[26]

It can be seen in Figure 8B that the molecule folds and unfolds several times during a 20-ns run in the UNRES Berendsen simulations. The average folding time at the folding temperature ($\tau_f = 0.74$ ns; Figure 7A) is 7.4 times shorter for UNRES (with the Berendsen thermostat) compared to the all-atom simulations with implicit solvent ($\tau_f = 5.5$ ns; Figure 7B); this ratio is similar to that derived from Ala_{10} simulations. The average folding time obtained in Langevin simulations (4.2 ns) is longer than that obtained in Berendsen simulations (0.74 ns). Thus, as for Ala_{10}, the presence of explicit friction forces slows down the folding. The average CPU time to fold protein A at 420 K with the all-atom approach is 9136 min/processor (152 h/processor), while that for UNRES (with the Berendsen thermostat) is 30 min/processor. This offers a 305-fold speedup. The time step used in UNRES simulations of protein A (4.89 fs) was larger than that used in the corresponding all-atom simulations (2 fs); it should be noted however that, even when the same time step was used for UNRES as for all-atom dynamics, UNRES still offers about a 100-fold speedup.

It can be seen from Figure 8 that the lowest potential energy obtained during MD runs at the folding temperature is by about 120 kcal/mol higher than the lowest potential energy found in CSA runs. This also occurred for the Ala_{10} polypeptide, but the difference was smaller because of the smaller size of the system. This suggests that the UNRES force field optimized for protein-structure prediction based on a global-minimum search need not necessarily be good for molecular dynamics, and reoptimization based on MD rather than CSA-generated ensembles is required.

Because the purpose of this study was to perform initial tests of the UNRES/MD approach and to compare the all-atom and the UNRES time scales, we will refrain from further analysis of the folding trajectories here. The all-atom results for protein A will be reported in detail in a separate paper.[59] More tests of the UNRES/MD approach on a larger number of proteins have appeared in a separate paper.[60]

It should be noted that UNRES molecular dynamics can be extended to calculate all-atom folding pathways. The coarse-grained structures from selected points of the pathway can be converted into an all-atom representation using the energy-based method developed in our laboratory,[61,62] and limited all-atom MD simulations can subsequently be carried out for each of them. The milestoning method developed recently by Faradjian and Elber[63] appears to be very appropriate for the last task.

4. Conclusions

We extended the Lagrange equations for the UNRES model of the polypeptide chain derived in the accompanying paper[1] to treat canonical ensembles. We implemented both the Langevin equation with explicit inclusion of the friction and stochastic forces and the Berendsen thermostat in which the friction and stochastic forces are included implicitly. We noted that, owing to the coupling to the thermal bath, which partially removes the energy drift caused by the error inherent in numerical integration, the time step can be increased to 0.1 mtu (4.89 fs), compared to 0.03 mtu (1.494 fs) for the microcanonical mode.[1] Comparison of the folding times for the Ala_{10} polypeptide (obtained with the AMBER 99 force field and explicit water molecules and with the AMBER 99 force field and implicit water treated at the GBSA level) with those obtained with the UNRES Langevin and UNRES Berendsen dynamics, respectively, reveals that the ratio of the UNRES to the all-atom time step is on the order of 4. A 1:7 ratio was obtained when comparing the folding times of protein A (with a three-helix bundle in the native structure) obtained from all-atom simulations with implicit solvent and UNRES simulations with the Berendsen thermostat. The dependence of the ratio of the UNRES to all-atom time scale on the system suggests that the effect of averaging out the internal degrees of freedom (the internal friction[26]) is system-dependent, and the equations of motion must be derived rigorously to establish the relationship between the UNRES and the all-atom and, ultimately, between the UNRES and the experimental time scales (not just based on the folding events as is done in this work). This can probably be accomplished by using the perturbation theory developed by van Kampen.[64] Research in this direction is currently underway in our laboratory.

Simulations of protein A have also demonstrated that molecular dynamics with UNRES is applicable to folding proteins and not only model polypeptides, using both the Berendsen thermostat and the full-blown Langevin dynamics. Simulations with the Berendsen thermostat result in quicker folding but are less realistic compared to Langevin dynamics simulations because of the absence of explicit friction in the former. Analysis of the CPU times required for folding demonstrates that UNRES offers at least a 4000-fold speedup relative to the all-atom approach with explicit water molecules and a 65-fold speedup relative to the all-atom approach with implicit water; the folding time of protein A using the Berendsen thermostat is

30 min on a single Alpha processor compared to about 152 h required for all-atom simulations with implicit solvent. This enables us to target a microsecond, if not a millisecond, time scale for protein folding.

It is interesting to note that, based on the experimental values of the folding times of α-helices, which are of the order of $0.6\,\mu s$,[65] one can conclude that both all-atom and UNRES time scales are underestimated. Even taking into account the fact that Ala_{10} is a faster folder than the model helical peptides used to estimate helix folding rates from experiment,[65] the time for folding this peptide should amount to at least 100 ns while, even with the all-atom approach with explicit water molecules, we obtained a 1-ns folding time. The answer to this is rather simple: Even when the solvent is treated at the detailed atomistic level, the vibrations of its bonds are suppressed by the SHAKE algorithm.[42,43] Without doing this, one observes primarily only collective vibrations of the solvent molecules, while the solute molecule remains highly intact. This means that a large part of the fast motions are removed by SHAKE even when the detailed approach is used. For protein A, the average folding time with UNRES dynamics was 4.5 ns, while the shortest folding times measured experimentally are on the order of microseconds even for very fast folding proteins.[65] Consequently, a unit of the UNRES time step corresponds to about 1000 units of the physical time scale.

It could be suspected that the short folding times obtained in our study are caused by the fact that UNRES is biased toward reproducing folded states relative to unfolded ones, because part of this force field consists of statistical potentials derived from the PDB. It should be noted, however, that the early versions of UNRES in which the bulk of the force field consisted of statistical potentials[2–4] were not capable of ab initio folding and required constraining the virtual-bond angles at 90°[2] or were only good for threading.[3,4] The capability of spontaneous secondary-structure formation and, thereby, ab initio folding was achieved only after the cumulant-based correlation terms were introduced.[5,7,13] These terms do not contain any information from the PDB statistics, are entirely physics-based, and were parametrized using ab initio quantum-mechanical energy surfaces of model systems.[13] The force field used in this work was ultimately parametrized by hierarchical optimization using 1IGD as a training protein; however, this is an $α + β$ protein as opposed to protein A studied in this work, which is an α protein. Thus, the force field is biased toward the native structure of 1IGD but not protein A. An interesting fact is that, although the force field was parametrized using 1IGD (but energy-minimized and not with MD decoys), MD simulations of this protein result in non-native structures, although the nativelike structure is the global minimum in the potential-energy surface.[60] The reason for this is the neglect of a large portion of configurational entropy when using energy-minimized decoys, while full configurational entropy is present in MD. Therefore, structures higher in energy but occupying a larger portion of conformational space will dominate in MD, while they will be rejected when using a global optimizer based on the potential energy as a scoring function.[60] Therefore, the fact that decoys from energy minimization and not from MD simulations were used in parametrization is a disadvantage as far as attraction toward native structure in MD simulations is concerned.

Acknowledgements

Appendix: Adaptation of the Integration Algorithm from TINKER to Molecular Dynamics with UNRES

Let us transform eq 3 to a system of equations with a diagonal friction matrix. For this purpose, we first introduce new variables $\mathbf{z}$ related to $\mathbf{q}$ by eq A1.

$$z = G^{1/2} q \tag{A1}$$

The Langevin equations can now be written as eq A2.

$$
\begin{aligned}
\ddot{z} &= -G^{-1/2} A^T \Gamma A G^{-1/2} \dot{z} - G^{-1/2} \nabla_q U + G^{-1/2} f^{rand} \\
&= -T^{-1} \dot{z} - \nabla_z U + f_z^{rand}
\end{aligned} \tag{A2}
$$

with

$$T = G^{-1/2} (A^T \Gamma A)^{-1} G^{-1/2} \tag{A3}$$

$$\nabla_z U = G^{-1/2} \nabla_q U \tag{A4}$$

$$f_z^{rand} = G^{-1/2} f^{rand} \tag{A5}$$

Now, by definition of the new variables

$$\xi = V^T z = V^T G^{-1/2} q \tag{A6}$$

where $\mathbf{V}$ is the matrix of the eigenvectors of $\mathbf{T}$, we can transform eq A2 to a system of uncoupled differential equations

$$\ddot{\xi}_i = -\tau_i^{-1} \dot{\xi}_i - \frac{\partial U}{\partial \xi_i} + f_{\xi_i}^{rand} \quad i = 1, 2, \dots, N \tag{A7}$$

where τ_i is the ith eigenvalue of matrix $\mathbf{T}$.

The stochastic integrator developed in refs [32] and [33] can now be implemented to eq A7, and the new quantities ξ and $\dot{\xi}$ can be transformed back to the original space. Steps 1 and 2 of the stochastic velocity Verlet algorithm are now the following:

Step 1:

$$q(t+\delta t) = q(t) + S(t,\delta t)\dot{q}(t) + T(t,\delta t)\ddot{q}\prime(t) + U(t,\delta t)\ddot{q}_x^{rand} \tag{A8}$$

Step 2:

$$
\begin{aligned}
\dot{q}(t+\delta t) = W(t,\delta t)\dot{q}(t) + \tfrac{1}{2} S(t,\delta t) \left[\ddot{q}\prime(t) + \ddot{q}\prime(t+\delta t) \right] + \\
X(t,\delta t)\ddot{q}_x^{rand} + Y(t,\delta t)\ddot{q}_v^{rand}
\end{aligned} \tag{A9}
$$

with

$$S(t,\delta t) = G^{1/2} V s(t,\delta t) V^T G^{-1/2} \tag{A10}$$

$$T(t,\delta t) = G^{1/2} V t(t,\delta t) V^T G^{-1/2} \tag{A11}$$

$$U(t,\delta t)=G^{1/2}Vu(t,\delta t)V^T G^{-1/2} \tag{A12}$$

$$W(t,\delta t)=G^{1/2}Vw(t,\delta t)V^T G^{-1/2} \tag{A13}$$

$$X(t,\delta t)=G^{1/2}Vx(t,\delta t)V^T G^{-1/2} \tag{A14}$$

$$Y(t,\delta t)=G^{1/2}Vy(t,\delta t)V^T G^{-1/2} \tag{A15}$$

where $\mathbf{s}$, $\mathbf{t}$, $\mathbf{u}$, $\mathbf{w}$, $\mathbf{x}$, and $\mathbf{y}$ are diagonal matrices with the diagonal elements defined by eqs A16–A21, respectively.

$$s_{ii}(t,\delta t)=\tau_i(t)\{1-\exp[-\tau_i^{-1}(t)\delta t]\} \tag{A16}$$

$$t_{ii}(t,\delta t)=\tau_i(t)[1-s_{ii}(t,\delta t)] \tag{A17}$$

$$u_{ii}(t,\delta t)=\tau_i(t)\left[\frac{1}{2}\tau_i^3(t)\delta t 2p_i\right]^{1/2} \tag{A18}$$

$$w_{ii}(t,\delta t)=\exp[-\tau_i^{-1}(t)\delta t] \tag{A19}$$

$$x_{ii}(t,\delta t)=\left[\frac{1}{2}\tau_i(t)v_i(t,\delta t)\delta t\right]^{1/2}\rho_i(t,\delta t) \tag{A20}$$

$$y_{ii}(t,\delta t)=\left[\frac{1}{2}\tau_i(t)v_i(t,\delta t)\delta t\right]^{1/2}[1-\rho_i(t,\delta t)] \tag{A21}$$

with

$$p_i(t,\delta t)=2\tau_i^{-1}(t)\delta t-3+\\ \{4-\exp[-\tau_i^{-1}(t)\delta t]\}\exp[-\tau_i^{-1}(t)\delta t] \tag{A22}$$

$$v_i(t,\delta t)=1-\exp[-2\tau_i^{-1}(t)\delta t] \tag{A23}$$

$$\rho_i(t,\delta t)=\frac{1-\exp[-\tau_i^{-1}(t)\delta t]}{\sqrt{p_i(t,\delta t)v_i(t,\delta t)}} \tag{A24}$$

This algorithm is more expensive compared to the simplified stochastic velocity Verlet algorithm described in section 2.2 because of the need to perform many matrix multiplications and to diagonalize the $\mathbf{A^T \Gamma A}$ matrix, if the friction coefficients (which depend on solvent-accessible surface area and, consequently, on conformation) are evaluated frequently. The last feature was not a big issue in our study, because we evaluate the friction coefficients only every 1000 MD steps. However, 7 multiplications of vectors by matrices must be carried out at each integration step, compared to 3 when applying the simplified algorithms (in the latter case those involve computation of the friction forces and the computation of accelerations from forces). Moreover, with our variable-time-step algorithm described in the accompanying paper,[1] the matrices $\mathbf{S}$, $\mathbf{T}$, $\mathbf{U}$, $\mathbf{W}$, $\mathbf{X}$, and $\mathbf{Y}$ must be reevaluated every time the time step is changed; this involves 12 matrix multiplications. We reduced this cost by storing the matrices for each time step, because the time step in our algorithm is always scaled down by integer powers of 2 (eq 42 of [ref 1]). By carrying out test calculations, we found that, given the values of the time step

and the friction coefficients used in this study, the factors $1 - \exp(-\tau_i^{-1}\delta t)$ are approximately $\tau_i^{-1}\delta t$ and the matrices $\mathbf{S}, \mathbf{T}, \mathbf{U}, \mathbf{W}, \mathbf{X}$, and $\mathbf{Y}$ are almost diagonal. With the additional assumption that the correlation coefficients ρ between velocities and positions are almost equal to zero, the TINKER algorithm reduces to the simplified stochastic velocity Verlet algorithm described in section 2.2. Consequently, we used the simplified algorithm in all tests reported in the paper.

We also adapted the stochastic integrator introduced by Ciccotti and co-workers[36,37] to the Langevin equations for UNRES. The derivation of the pertinent equations is similar to those for TINKER.

References and Notes

1. Khalili M, Liwo A, Rakowski F, Grochowski P, Scheraga HA. J Phys Chem B 2005;109:13785. [PubMed: 16852727]

2. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. Protein Sci 1993;2:1715. [PubMed: 8251944]

3. Liwo A, Ołdziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. J Comput Chem 1997;18:849.

4. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Ołdziej S, Scheraga HA. J Comput Chem 1997;18:874.

5. Liwo A, Kaźmierkiewicz R, Czaplewski C, Groth M, Ołdziej S, Wawak RJ, Rackovsky S, Pincus MR, Scheraga HA. J Comput Chem 1998;19:259.

6. Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA. Proc Natl Acad Sci USA 1999;96:5482. [PubMed: 10318909]

7. Liwo A, Czaplewski C, Pillardy J, Scheraga HA. J Chem Phys 2001;115:2323.

8. Lee J, Ripoll DR, Czaplewski C, Pillardy J, Wedemeyer WJ, Scheraga HA. J Phys Chem B 2001;105:7291.

9. Pillardy J, Czaplewski C, Liwo A, Wedemeyer WJ, Lee J, Ripoll DR, Arłukowicz P, Ołdziej S, Arnautova YA, Scheraga HA. J Phys Chem B 2001;105:7299.

10. Pillardy J, Czaplewski C, Liwo A, Lee J, Ripoll DR, Kaźmierkiewicz R, Ołdziej S, Wedemeyer WJ, Gibson KD, Arnautova YA, Saunders J, Ye YJ, Scheraga HA. Proc Natl Acad Sci USA 2001;98:2329. [PubMed: 11226239]

11. Liwo A, Arłukowicz P, Czaplewski C, Ołdziej S, Pillardy J, Scheraga HA. Proc Natl Acad Sci USA 2002;99:1937. [PubMed: 11854494]

12. Ołdziej S, Kozłowska U, Liwo A, Scheraga HA. J Phys Chem A 2003;107:8035.

13. Liwo A, Ołdziej S, Czaplewski C, Kozłowska U, Scheraga HA. J Phys Chem B 2004;108:9421.

14. Liwo A, Arłukowicz P, Ołdziej S, Czaplewski C, Makowski M, Scheraga HA. J Phys Chem B 2004;108:16918.

15. Ołdziej S, Liwo A, Czaplewski C, Pillardy J, Scheraga HA. J Phys Chem B 2004;108:16934.

16. Ołdziej S, Łągiewka J, Liwo A, Czaplewski C, Chinchio M, Nanias M, Scheraga HA. J Phys Chem B 2004;108:16950.

17. Jang S, Kim E, Shin S, Pak Y. J Am Chem Soc 2003;125:14841. [PubMed: 14640661]

18. Levy RM, Karplus M, McCammon JA. Chem Phys Lett 1979;65:4.

19. Yun-Yu S, Lu W, van Gunsteren WF. Mol Simul 1988;1:369.

20. Levitt M. J Mol Biol 1976;104:59. [PubMed: 957439]

21. Richmond TJ. J Mol Biol 1984;178:63. [PubMed: 6548264]

22. Wesson L, Eisenberg D. Protein Sci 1992;1:227. [PubMed: 1304905]

23. http://dasher.wustl.edu/tinker/.

24. Ren P, Ponder JW. J Phys Chem B 2003;107:5933.

25. Veitshans T, Klimov D, Thirumalai D. Folding Des 1996;2:1.

26. de Gennes, PG. Scaling Concepts in Polymer Physics. Chapter VI. Cornell University Press; Ithaca, NY: 1979.

27. Cieplak M, Hoang TX, Robbins MO. Proteins: Struct, Funct, Genet 2002;49:104. [PubMed: 12211020]

28. van Gunsteren WF, Berendsen HJC. Mol Phys 1982;45:637.

29. van Gunsteren WF, Berendsen HJC. Mol Simul 1988;1:173.

30. van Gunsteren, WF. Molecular dynamics and stochastic dynamics: A primer. In: van Gunsteren, WF.; Weiner, PK.; Wilkinson, AJ., editors. Computer Simulation of Biomolecular Systems. ESCOM; Leiden, The Netherlands: 1993. p. 3-36.

31. Zhang G, Schlick T. Mol Phys 1995;84:1077.

32. Allen MP. Mol Phys 1980;40:1073.

33. Guarnieri F, Still WC. J Comput Chem 1994;15:1302.

34. Paterlini MG, Ferguson DM. Chem Phys 1998;236:243.

35. Suzuki M. Proc Jpn Acad, Ser B 1993;69:161.

36. Ricci A, Ciccotti G. Mol Phys 2003;101:1927.

37. Ciccotti G, Kalibaeva G. Philos Trans R Soc London, Ser A 2004;362:1583.

38. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. J Chem Phys 1984;81:3684.

39. Weiner SJ, Kollman PA, Nguyen DT, Case DA. J Comput Chem 1986;7:230.

40. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE III, DeBolt S, Ferguson D, Seibel G, Kollman P. Comput Phys Commun 1995;91:11.

41. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. J Chem Phys 1983;79:926.

42. Ryckaert JP, Ciccotti G, Berendsen HJC. J Comput Phys 1977;23:327.

43. van Gunsteren WF, Berendsen HJC. Mol Phys 1977;34:1311.

44. Cramer CJ, Truhlar DG. Chem Rev 1999;99:2161. [PubMed: 11849023]

45. Bashford D, Case DA. Annu Rev Phys Chem 2000;51:129. [PubMed: 11031278]

46. Gouda H, Torigoe H, Saito A, Sato M, Arata Y, Shimada I. Biochemistry 1992;31:9665. [PubMed: 1390743]

47. Némethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA. J Phys Chem 1992;96:6472.

48. Lee J, Liwo A, Scheraga HA. Proc Natl Acad Sci USA 1999;96:2025. [PubMed: 10051588]

49. Hoang TX, Cieplak M. J Chem Phys 2000;112:6851.

50. Qiu L, Hagen SJ. Chem Phys 2004;307:243.

51. Fauchere JL, Pliška V. Eur J Med Chem 1983;18:369.

52. Alonso DOV, Daggett V. Proc Natl Acad Sci USA 2000;97:133. [PubMed: 10618383]

53. Takada S. Proteins: Struct, Funct, Genet 2001;42:85. [PubMed: 11093263]

54. Ghosh A, Elber R, Scheraga HA. Proc Natl Acad Sci USA 2002;99:10394. [PubMed: 12140363]

55. Bottomley SP, Popplewell AG, Scawen M, Wan T, Sutton BJ, Gore MG. Protein Eng 1994;7:1463. [PubMed: 7716157]

56. Bai YW, Karimi A, Dyson HJ, Wright PE. Protein Sci 1997;6:1449. [PubMed: 9232646]

57. Guo ZY, Brooks CL, Boczko EM. Proc Natl Acad Sci USA 1997;94:10161. [PubMed: 9294180]

58. Sato S, Religa TL, Daggett V, Fersht AR. Proc Natl Acad Sci USA 2004;101:6952. [PubMed: 15069202]

59. Jagielska A, Scheraga HA. manuscript to be submitted for publication

60. Liwo A, Khalili M, Scheraga HA. Proc Natl Acad Sci USA 2005;102:2362. [PubMed: 15677316]

61. Kaźmierkiewicz R, Liwo A, Scheraga HA. J Comput Chem 2002;23:715. [PubMed: 11948589]

62. Kaźmierkiewicz R, Liwo A, Scheraga HA. Biophys Chem 2003;100:261. [PubMed: 12646370] Biophys Chem 2003;106:9.Erratum

63. Faradjian AK, Elber R. J Chem Phys 2004;120:10880. [PubMed: 15268118]

64. van Kampen CW. Phys Rep 1985;124:69.

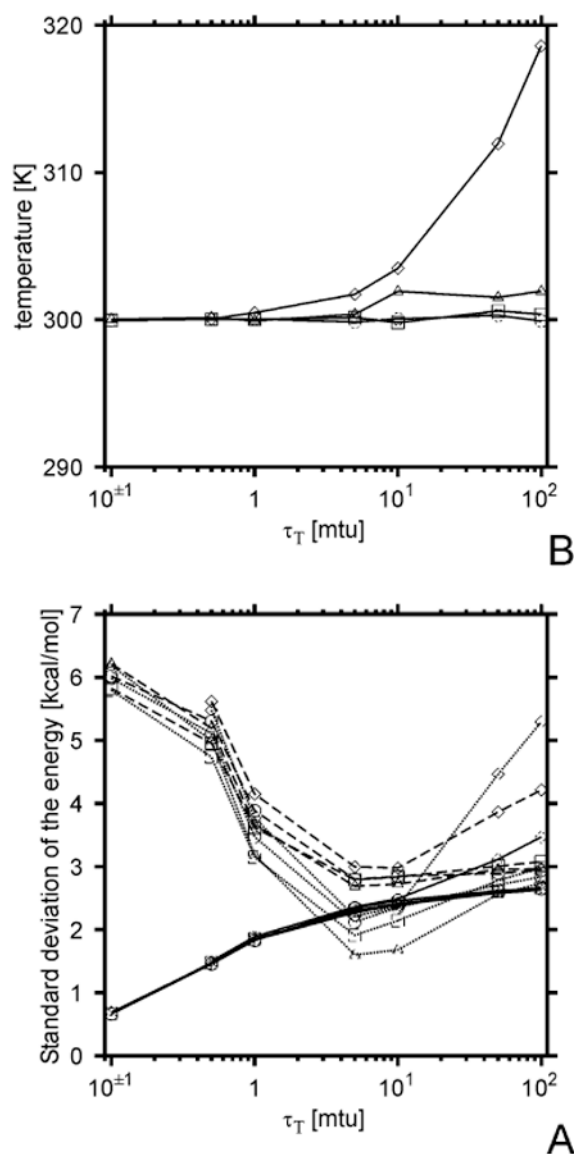65. Kubelka J, Hofrichter J, Eaton WA. Curr Opin Struct Biol 2004;14:76. [PubMed: 15102453]

Figure 1.
Plots of the standard deviation in the kinetic (σ_T, solid lines), the potential (σ_V, dashed lines), and the total (σ_E, dotted lines) energy (A), and average temperature (kinetic energy, B), as functions of the coupling parameter τ_T for different values of the time step δt. Simulations were carried out on the Ala$_{10}$ system, starting from the fully α-helical structure and 10 200-ps trajectories were run for each δt and each τ_T value. The symbols correspond to the values of the time step: circles, $\delta t = 0.01$ mtu (0.489 fs); squares, $\delta t = 0.02$ mtu (0.978 fs); triangles, $\delta t = 0.05$ mtu (2.445 fs); diamonds, $\delta t = 0.10$ mtu (4.89 fs).
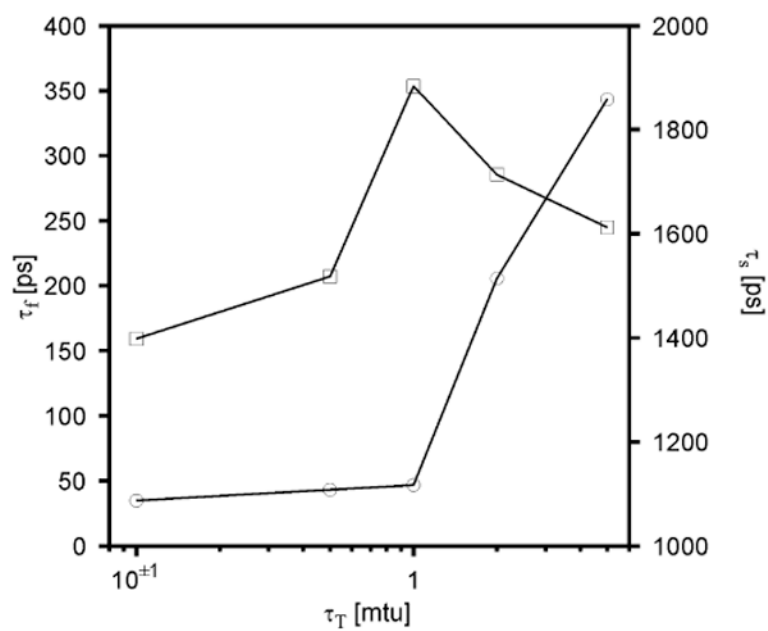
Figure 2.
Dependence of the folding (τ_f, circles) and stability (τ_s, squares) time on the coupling parameter τ_T for the Ala$_{10}$ system obtained in UNRES/MD runs at the temperature $T = 300$ K with the Berendsen thermostat. Ten 200-ps trajectories were run for each value of τ_T.
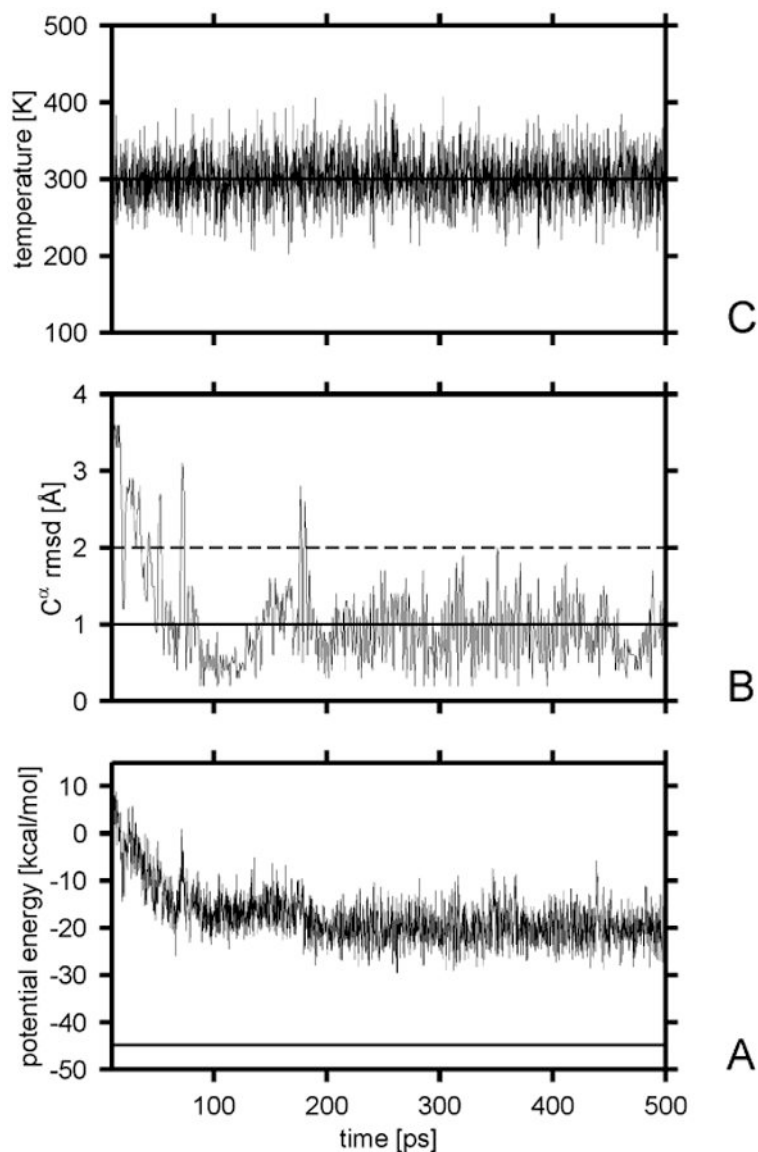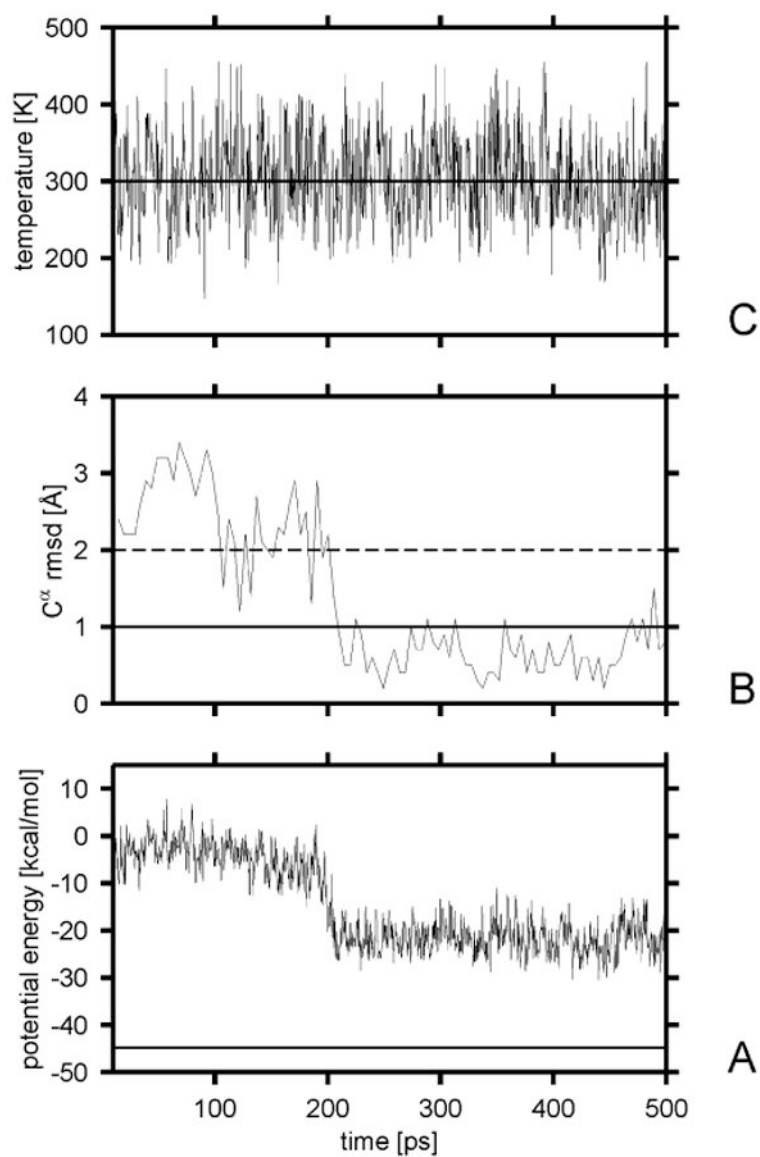
Figure 3.
Variation of the potential energy (A), the C^α rmsd from the α-helical structure (B), and the temperature (C) with time for a sample Berendsen trajectory of Ala_{10} with the UNRES force field obtained with $\tau_T = 1$ mtu. The solid horizontal line in part A is the lowest potential energy found by the CSA method. The solid horizontal line in panel B corresponds to the 1-Å cutoff of the rmsd over $C_2^\alpha \cdots C_9^\alpha$; after this value was reached, the chain was considered to have folded. The dashed horizontal line in panel B corresponds to the 2-Å rmsd cutoff above which a structure is considered to have left the native basin.
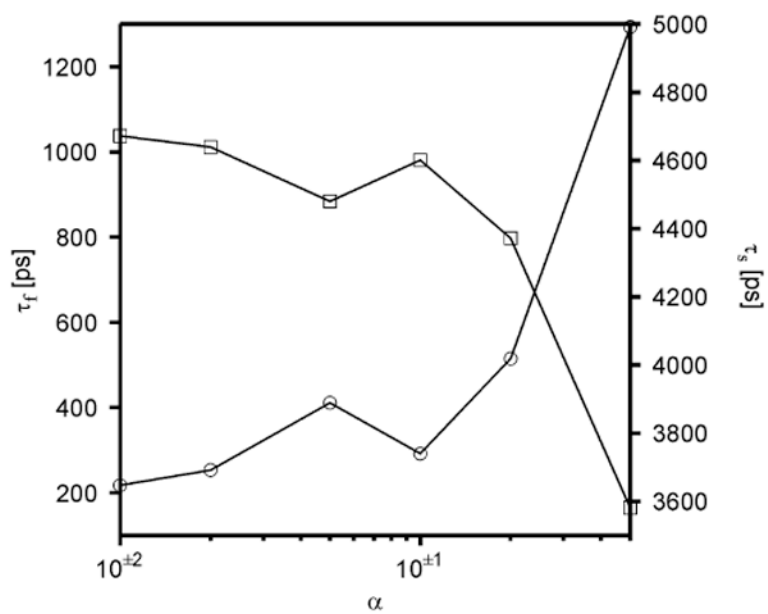
Figure 4.
Variation of the potential energy (A), the C^α rmsd from the α-helical structure (B), and the temperature (C) for a sample Langevin trajectory of Ala_{10} with the UNRES force field obtained with the scaling factor of the friction coefficients of $\alpha = 0.05$. The solid horizontal line in part A is the lowest potential energy found by the CSA method. The solid horizontal line in panel B corresponds to the 1-Å cutoff of the rmsd over $C_2^\alpha \cdots C_9^\alpha$; after this value was reached, the chain was considered to have folded. The dashed horizontal line in panel B corresponds to the 2-Å rmsd cutoff above which a structure is considered to have left the native basin.

Figure 5.
Dependence of the folding time (τ_f, circles) and stability time (τ_s, squares) of Ala$_{10}$ in UNRES Langevin simulations on the scaling factor of the friction coefficients (α of eq 6).
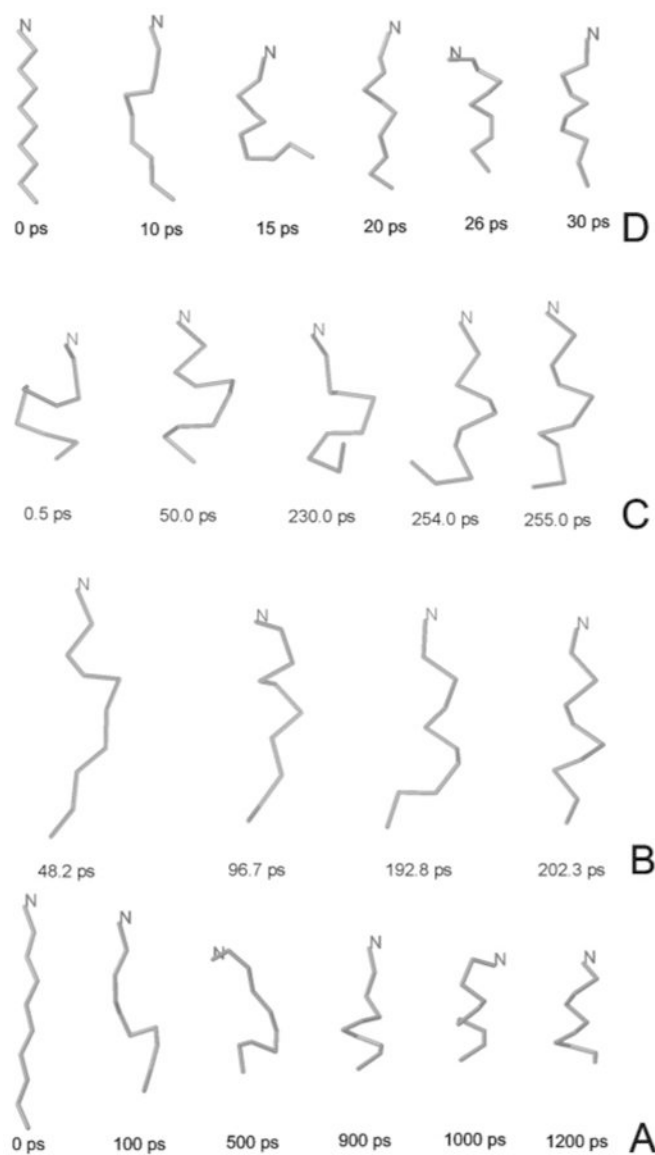
Figure 6.
Snapshots from the AMBER all-atom simulations with explicit water molecules (A) compared with Langevin UNRES with $\alpha = 0.05$ (B), and AMBER all-atom simulations with implicit water (C) compared with Berendsen UNRES (D). The trajectories of the Ala_{10} polypeptide are shown as α-carbon traces.
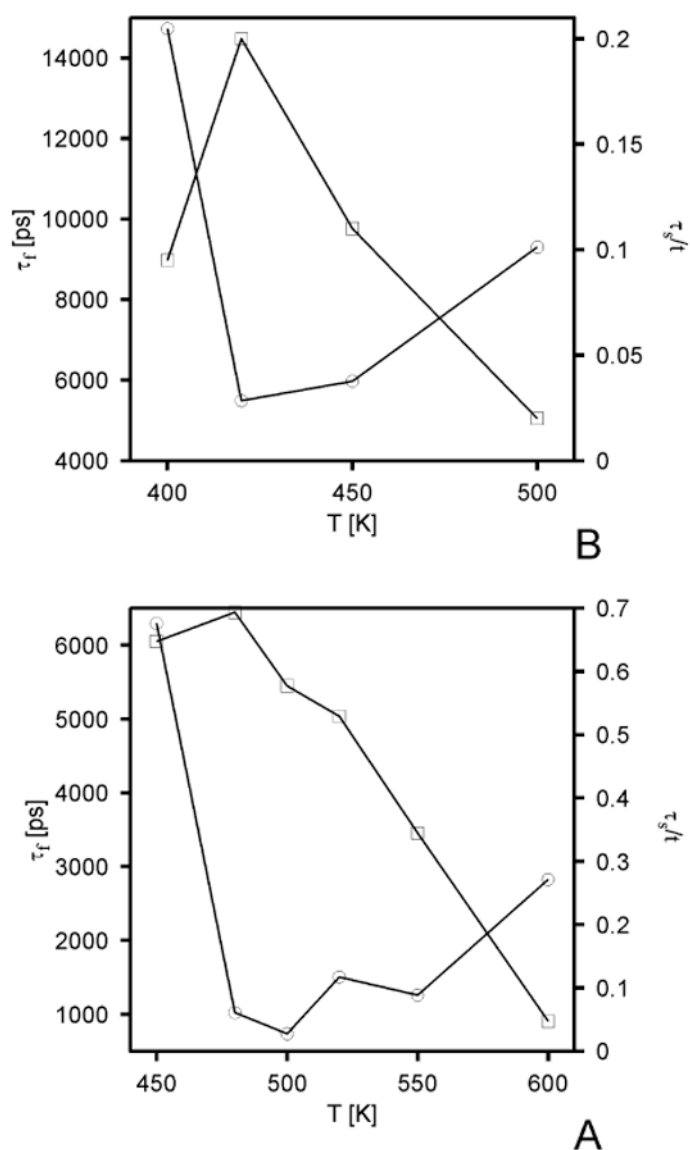
Figure 7.
Dependence of the folding (τ_f, circles) and stability (τ_s, squares) time for protein A on the temperature in UNRES simulations with the Berendsen thermostat (A) and all-atom AMBER simulations with implicit water modeled at the GBSA level (B). In the all-atom simulations, 5 20-ns trajectories were run, and 10 20-ns trajectories in the UNRES simulations with the Berendsen thermostat. All trajectories were started from fully extended structures.
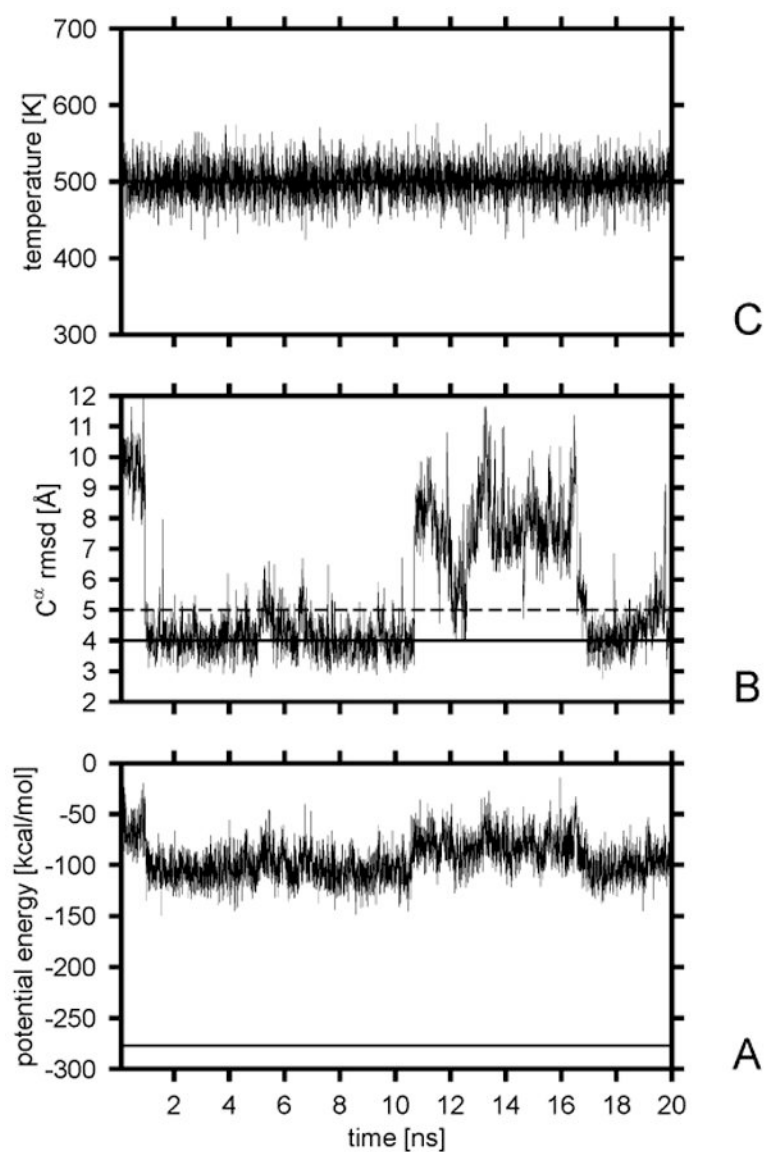
Figure 8.
Variation of the potential energy (A), the C^α rmsd from the native structure (B), and the temperature (C) with time for a sample trajectory of protein A with the UNRES force field and Berendsen thermostat at the folding temperature. The solid horizontal line in part A is the lowest potential energy found in our earlier study[14] by the CSA method. The solid horizontal line in panel B corresponds to the 4-Å C^α rmsd cutoff; after reaching this value, the chain was considered to have folded. The dashed horizontal line in panel B corresponds to the 5-Å rmsd cutoff above which a structure is considered to have left the native basin.
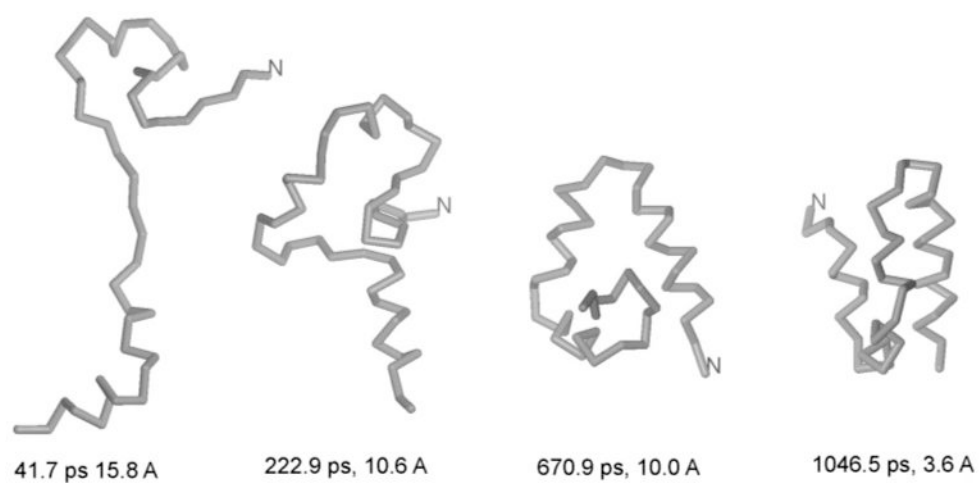
41.7 ps 15.8 A   222.9 ps, 10.6 A   670.9 ps, 10.0 A   1046.5 ps, 3.6 A

Figure 9.
Snapshots from a representative folding trajectory of protein A at 500 K obtained with the UNRES force field and Berendsen thermostat.
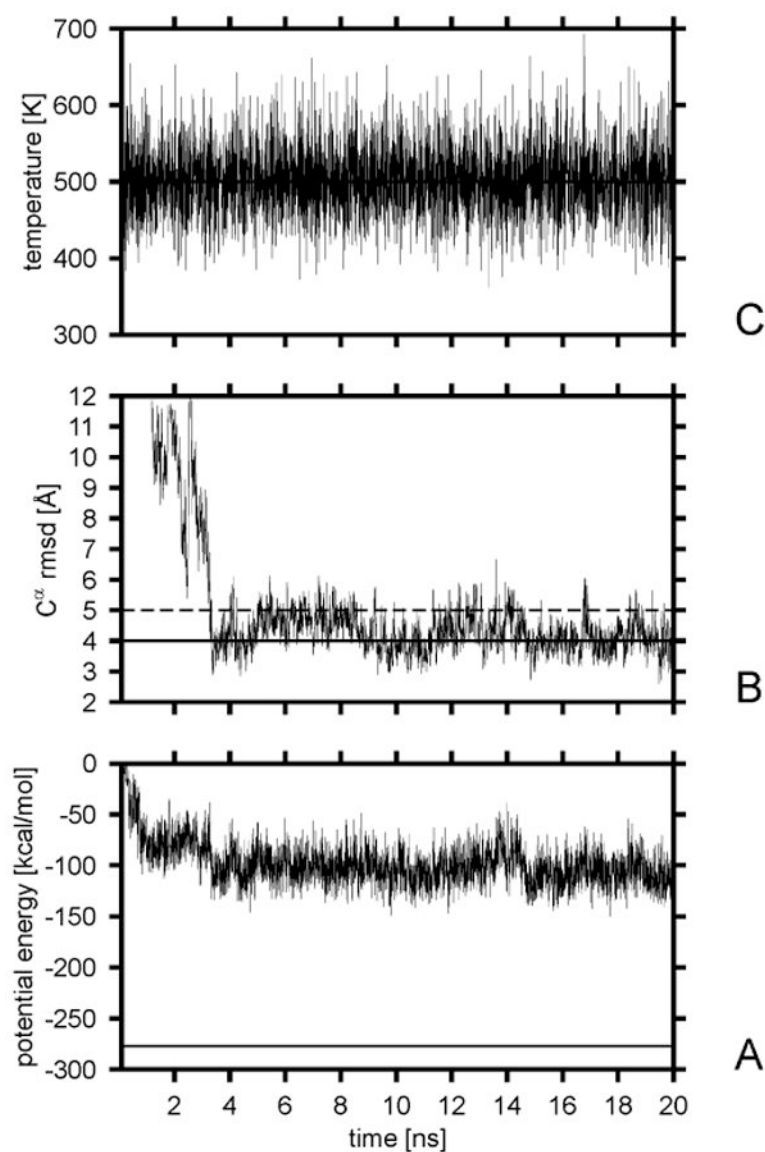
Figure 10.
Variation of the potential energy (A), the C^α rmsd from the native structure (B), and the temperature (C) with time for a sample Langevin trajectory of protein A with the UNRES force field at the folding temperature. The solid horizontal line in part A is the lowest potential energy found in our earlier study[14] by the CSA method. The solid horizontal line in panel B corresponds to the 4-Å C^α rmsd cutoff; after reaching this value, the chain was considered to have folded. The dashed horizontal line in panel B corresponds to the 5-Å rmsd cutoff above which a structure is considered to have left the native basin.
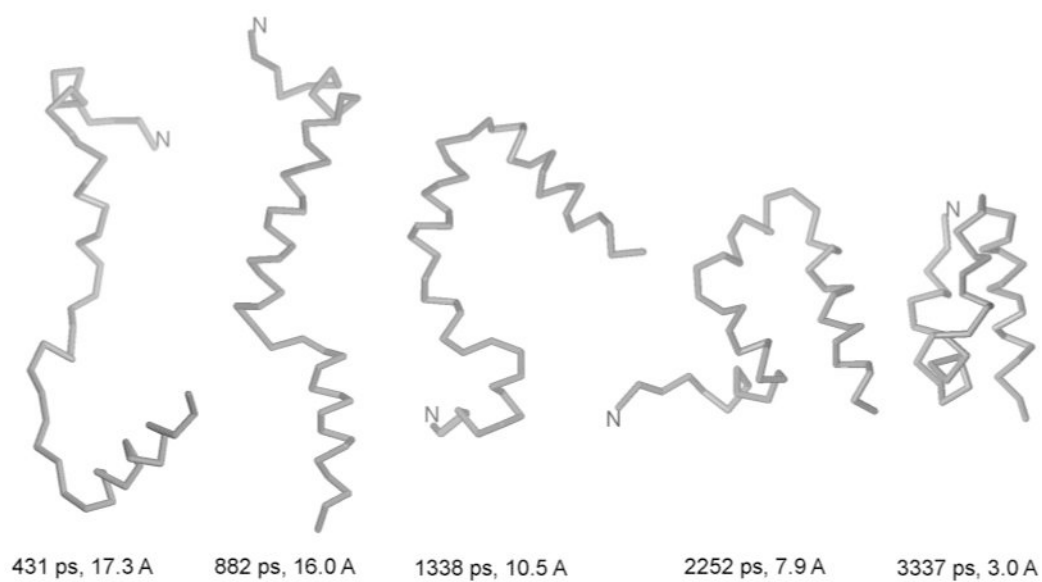
431 ps, 17.3 A   882 ps, 16.0 A   1338 ps, 10.5 A   2252 ps, 7.9 A   3337 ps, 3.0 A

Figure 11.
Snapshots from a representative Langevin folding trajectory of protein A at 500 K obtained with the UNRES force field.

TABLE 1

Folding (τ_f) and Relative Stability (τ_s/t)[a] Times for the Unblocked Ala$_{10}$ Polypeptide Run with the All-Atom Simulations with Explicit Solvent (7 trajectories), UNRES with Langevin Dynamics (10 trajectories), [b] the All-Atom Simulations with Implicit Solvent and the Berendsen Thermostat (10 trajectories; $\tau_T = 0.1$ ps), and UNRES Dynamics with the Berendsen Thermostat (10 trajectories; $\tau_T = 1$ mtu)

force field	τ_f (ps)			τ_s/t (ps)		
	min	max	average	max	min	average
AMBER explicit water	1107.0	2730.0	1748.7	0.14	0.02	0.16
UNRES/Langevin	19.6	1281.2	411.2	0.99	0.74	0.92
AMBER GBSA	8.0	670.0	200.2	0.67	0.35	0.47
UNRES/Berendsen	8.8	125.2	52.5	0.95	0.57	0.77

[a] t is the total length of a trajectory.

[b] The friction coefficients were scaled by the factor of 0.05.