Molecular Dynamics with the United-Residue Force Field: Ab Initio Folding Simulations of Multichain Proteins

Ana V. Rojas,[†,‡,§] Adam Liwo,[†] and Harold A. Scheraga*[,†]

Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301, Department of Physics and Astronomy, Louisiana State University, Baton Rouge, Louisiana 70803-4001, and Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana 70803-4001

Received: September 6, 2006; In Final Form: October 26, 2006

The implementation of molecular dynamics with the united-residue (UNRES) force field is extended to treat multichain proteins. Constant temperature was maintained in the simulations with Berendsen or Langevin thermostats. The method was tested on three α-helical proteins (1G6U and GCN4-p1, each with two chains, and 1C94, with four chains). Simulations were carried out for both the isolated single chains and the multichain complexes**.** The proteins were folded by starting from the extended conformation with random initial velocities and with the chains parallel to each other. No symmetry constraints or structure information were included for the single chains or the multichain complexes. In the case of single-chain simulations, a high percentage of the trajectories (100% for 1G6U, 90% for GCN4-p1, and 80% for 1C94) converged to nativelike structures (assumed as the experimental structure of a monomer in the multichain complex), showing that, for the proteins studied in this work with the UNRES force field, the interactions between chains are not critical for stabilization of the individual chains. In the case of multichain simulations, the native structures of the 1G6U and GCN4-p1 complexes, but not that of 1C94, are predicted successfully. The association of the subunits does not follow a unique mechanism; the monomers were observed to fold both before and simultaneously with their association.

1. Introduction

Predicting the native structure of a protein from knowledge of its amino acid sequence by an ab initio (physics-based) approach remains one the most difficult problems in contemporary computational biology. An even more challenging problem is the prediction of the folding pathway of a protein. The ab initio approach has the advantage that it provides thermodynamic and kinetic information about the different stages of the folding process as well as the final structure. To accomplish such predictions, it is necessary to simulate the folding process in real time, starting from a statistical coil (unfolded) conformation, until the native structure is reached. For such a simulation to be realistic, it should ideally include atomic details of both the system and the solvent.[1] However, with today's computational power, explicit-solvent all-atom molecular dynamics (MD) algorithms can simulate only events that range up to nanoseconds for typical proteins or microseconds for very small ones.[1−3] These time scales are at least 1 order of magnitude smaller than the folding times of proteins. To overcome this problem, all-atom simulations either implement alternative sampling methods, such as umbrella sampling,[4] or simulate the unfolding process and some aspects of its refolding;[1,2] simulations primarily treat single-chain proteins, but in some cases,[5−10] computations are carried out for oligomers. In general, simulations of oligomers either study the stability of a specific structure[6,10] or the kinetics of folding and/or assembly[5,7−9] of the subunits. Stability studies are usually carried out by all-atom MD,[6,7,10] but this technique is computationally too expensive to study the kinetics of the folding process. To reduce the computational cost, the main approach has made use of minimal models; a minimal model is one for which each amino acid is represented by a few interaction sites, reducing the dimensionality of the problem. Although the information that they can provide is not as detailed as that obtained by all-atom models, they can achieve longer simulation times. Some minimal models have further reduced the computational cost by using a Gō-type potential,[7−9] which creates a funnel-like landscape biased toward the native structure, thereby speeding up the folding process. It has also been possible to study the kinetics of oligomeric proteins without including any structural knowledge of the particular protein of interest. For example, Vieth et al.[5] used a lattice model with a statistical potential (i.e., biased toward structures in a library but not toward the particular structure being studied) and Monte Carlo (MC) dynamics to study the folding pathway of the GCN4 leucine zipper from randomly generated initial structures.

With a minimal model, we have recently[11−14] developed a molecular dynamics algorithm for the physics-based united-residue (UNRES) force field that was previously developed in our laboratory.[15−23] We will refer to this implementation of UNRES as UNRES/MD. UNRES was originally designed and parametrized to locate nativelike structures of proteins as the lowest in potential energy by unrestricted global optimization. The latest version of UNRES, referred to as the 4P force field,[23] was optimized on four training proteins: 1GAB (all-α), 1E0L (all-β), 1E0G ($\alpha + \beta$), and 1IGD ($\alpha + \beta$). It performed well in the CASP6 exercise;[24] the largest molecule that was folded with this force field contained 208 amino acid residues. The average

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

[†] Baker Laboratory of Chemistry and Chemical Biology, Cornell University.

[‡] Department of Physics and Astronomy, Louisiana State University.

[§] Center for Computation and Technology, Louisiana State University.

10.1021/jp065810x CCC: $37.00
Published on Web 12/15/2006

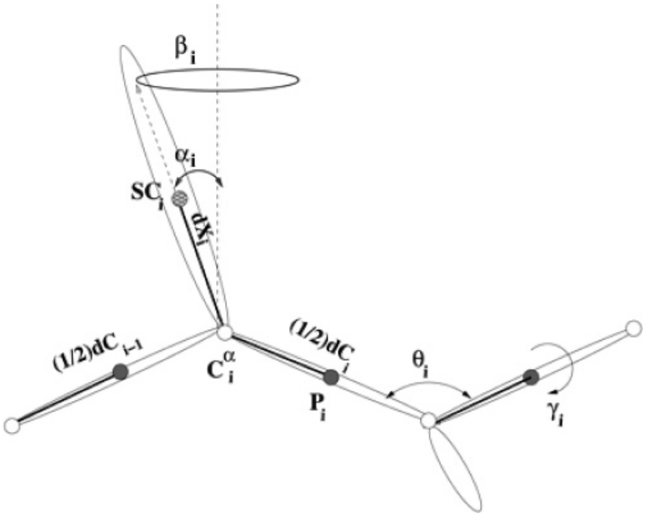

Figure 1. UNRES representation of a polypeptide chain. Filled circles represent the united peptide groups (p), and open circles represent the C$^{\alpha}$ atoms, which serve as geometric points. Ellipsoids with their centers of mass at SC positions represent UNRES side chains. The p's are located halfway between two consecutive C$^{\alpha}$ atoms, at positions ($^{1}/_{2}$)-**dC**. The conformation of the polypeptide chain can be described fully by either the coordinates of all the **dC** and **dX** vectors or by the virtual bond angles θ, the virtual bond dihedral angles γ, and the angles α and β defining the orientation of the side chain with respect to the backbone.

length of correctly predicted segments of α-helical proteins with this force field is 67 residues (Table 8 in ref 23).

Since the degrees of freedom corresponding to the fastest motions are averaged out[18] in UNRES, UNRES/MD was able to simulate events that fall into the microsecond time scale.[13] After the success of UNRES/MD with single-chain proteins, it seemed natural to generalize the method to treat multiple-chain proteins. A multichain version of UNRES and a global optimization search based on conformational space annealing (CSA) had previously been developed in our laboratory.[25] However, that implementation required the use of symmetry to achieve proper folding. Here, we present an extension of UNRES/MD, which can simulate the folding pathway of oligomeric proteins from an extended conformation without imposing symmetry constraints of any kind. As with the single-chain UNRES/MD calculations, Berendsen dynamics (BD) and Langevin dynamics (LD) are used to mimic energy exchange with the solvent and, consequently, to maintain constant temperature. LD provides a more realistic picture through explicit inclusion of the nonconservative friction and random forces, which account for collisions of the protein with the solvent molecules. Because the purpose of this work was to extend the UNRES/MD approach to multichain systems and not to develop an improved force field, we used systems that the 4P force field could treat to test the method. We ran simulations on the following three α-helical proteins of known native structure: 1G6U (two chains, 48 residues each), 2ZTA (two chains, 33 residues each), and 1C94 (four chains, 38 residues each). The complexity and size of these proteins is similar to that of the α-helical proteins tested in our previous work on single-chain UNRES/MD,[13] and the size of the smallest of them (2ZTA) is within the average size of structural segments of α-helical proteins that can be predicted successfully with the 4P force field.[23] These systems are, therefore, appropriate to test the UNRES/MD approach for multichain proteins, given the limitations of the present force field. We did not use β or $\alpha + \beta$ proteins because we found in our earlier work[13] that UNRES/MD generally produces non-native α-helical structures for such proteins, even though the native structures are global energy minima in the UNRES energy surface; this happens because the conformational entropy is neglected in force-field parametrization. We believe that this problem can be overcome by improving the force field and introducing entropic effects, an activity which is presently ongoing in our laboratory.

2. Methods

2.1. United-Residue Force Field. UNRES is a coarse-grained model[15−23] in which the backbone is represented as a sequence of α-carbon (C$^{\alpha}$) atoms linked by virtual bonds designated as dC, with united peptide groups (p's) in their centers. United side chains (SCs) are connected by virtual bonds designated as dX to the backbone at the C$^{\alpha}$ positions with the center of mass of the side chain at the end of dX (Figure 1). The geometry of the protein is then fully described by the virtual bond vectors dC's and dX's. Since the forces in UNRES are exerted on the peptide groups and side chains, hereafter we will use the term "interacting sites" to refer to both united peptide groups and side chains. The complete UNRES potential energy function for a single chain is given by

$$U_{\text{single chain}} = \sum_{j}\sum_{i<j} U_{\text{SC}_i\text{SC}_j} + \sum_{\text{ss}} U_{\text{Cys}_{iss}\text{Cys}_{jss}} + w_{\text{SCp}}\sum_{j}\sum_{i\neq j} U_{\text{SC}_i\text{p}_j} + w_{\text{el}}\sum_{j}\sum_{i<j-1} U_{\text{p}_i\text{p}_j} + w_{\text{tor}}\sum_{i} U_{\text{tor}}(\gamma_i) + w_{\text{tord}}\sum_{i} U_{\text{tord}}(\gamma_i, \gamma_{i+1}) + w_{\text{b}}\sum_{i} U_{\text{b}}(\theta_i) + w_{\text{rot}}\sum_{i} U_{\text{rot}}(\alpha_i, \beta_i) + \sum_{m=3}^{6} w_{\text{corr}}^{(m)} U_{\text{corr}}^{(m)} + w_{\text{vib}}\sum_{i} U_{\text{vib}}(d_i) \quad (1)$$

where the indices i and j run over the residues. The terms $U_{\text{SC}_i\text{SC}_j}$ (derived and parametrized in ref 16) correspond to the mean free energy of hydrophobic (hydrophilic) interactions between the side chains. These terms implicitly contain the contributions from the interactions of the side chain with the solvent. The terms $U_{\text{Cys}_{iss}\text{Cys}_{jss}}$ (derived and parametrized in ref 26) account for the energy of disulfide bonds, with ss running through all those pairs of half-cystines that are known a priori to form disulfide bonds.[26] The terms $U_{\text{SC}_i\text{p}_j}$ correspond to the excluded-volume potential of the side chain−peptide group interactions. The terms $U_{\text{p}_i\text{p}_j}$ (derived in ref 15 and parametrized in ref 21) represent the energy of average electrostatic interactions between backbone peptide groups. The terms U_{tor} and U_{tord} (derived and parametrized in ref 20) are the torsional and the double-torsional potentials, respectively, for the rotation about a given virtual bond or two consecutive virtual bonds. The terms U_{b} and U_{rot} (derived and parametrized in ref 17) are the virtual-angle-bending and side-chain-rotamer potentials, respectively. The terms $U_{\text{corr}}^{(m)}$ (derived in ref 18 and parametrized in ref 21) correspond to the correlations (of order m) between peptide-group electrostatic and backbone-local interactions. The terms $U_{\text{vib}}(d_i)$ (derived and parametrized in ref 11), d_i being the length of the ith virtual bond, are simple harmonic potentials defined by eq 2

$$U_{\text{vib}}(d_i) = \frac{1}{2}k(d_i - d_i^{\text{o}})^2 \quad (2)$$

where k is a force constant, currently set at 500 kcal/(mol Å^2) and d_i^{o} is the average length (corresponding to that used in the fixed-bond UNRES) of the ith virtual bond. The w's in eq 1 are the weights of the respective terms.

The UNRES force field has also been extended to multiple-chain proteins.[25] In the present work, the interchain interaction energies (and their form, parameters, and weights) were taken to be the same as those of the intrachain terms in the treatment of single chains. However, since the interacting sites between chains are not backbone-connected, not all the terms present in eq 1 contribute to the interchain energy. The interaction energy between two different chains (identified by superscripts *k* and *l*, respectively) can be expressed by

$$U_{\mathrm{interchain}}^{k,l} = \sum_i \sum_j U_{\mathrm{SC}_i{}^k\mathrm{SC}_j{}^l} + \sum_{ss^{k,l}} U_{\mathrm{Cys}_{iss}{}^k\mathrm{Cys}_{jss}{}^l} + w_{\mathrm{SCp}} \sum_i \sum_j U_{\mathrm{SC}_i{}^k\mathrm{p}_j{}^l} + w_{\mathrm{SCp}} \sum_i \sum_j U_{\mathrm{p}_i{}^k\mathrm{SC}_j{}^l} + w_{\mathrm{el}} \sum_i \sum_j U_{\mathrm{p}_i{}^k\mathrm{p}_j{}^l} + \sum_{m=3}^{6} w_{\mathrm{corr,nonadj}}^{(m)} U_{\mathrm{corr,nonadj}}^{(m)} \quad (3)$$

where $U_{\mathrm{corr,nonadj}}$ represents the correlation terms corresponding to interactions between nonadjacent residues. The different terms in eq 3 have the same form, and the weights have the same values, as those in eq 1. Detailed descriptions of each of the terms in eqs 1 and 3 can be found in refs 15−18, 20, 21, and 26. It should be noted that eq 3 is different from eq 1 of ref 25 because the latter represents the complete multiple-chain UNRES potential energy, whereas eq 3 accounts only for the interaction between two chains in the system. Hence, eq 3 is only part of the contribution to the complete multiple-chain potential energy. It should also be mentioned here that, for the force field used in this work (4P force field[23]), the weights of the fifth- and sixth-order correlation terms, $w_{\mathrm{corr}}^{(5)}$ and $w_{\mathrm{corr}}^{(6)}$ in eq 1 and $w_{\mathrm{corr,nonadj}}^{(5)}$ and $w_{\mathrm{corr,nonadj}}^{(6)}$ in eq 3, are zero,[23] but these terms have been included in the equations for completeness.

To mimic peptide concentrations, the system was confined within a soft sphere. This was done by adding another term, U_{conf}, to the potential energy, causing each interacting site (either a peptide group or a side chain) to feel an attractive force toward the center of the sphere whenever it is outside the boundary of the sphere. This potential, which is added to eqs 1 and 3, is defined by eq 4

$$U_{\mathrm{conf}} = \sum_k \sum_i u_{\mathrm{conf}_i{}^k} \quad (4)$$

where $u_{\mathrm{conf}_i{}^k}$, the confining potential acting on interacting site *i* in chain *k*, is given by

$$u_{\mathrm{conf}_i{}^k} = \begin{cases} 0 & \text{if } r_i^k \le R_0 \\ k_{\mathrm{c}}(r_i^k - R_0)^4 & \text{if } r_i^k > R_0 \end{cases} \quad (5)$$

where k_{c} is a force constant with unit value ($k_{\mathrm{c}} = 1$ kcal/(mol Å^4)), r_i^k is the distance from interacting site *i* to the center of the sphere (placed at the center of mass of the initial conformation), and R_0 is the radius of the sphere. The radius of the sphere determines the volume of the system (volume $= 4\pi(R_0)^3/3$). Therefore, the value of R_0 and the number of peptide chains in the solution determine the peptide concentration of the simulated solution (see section 3 for details of the concentrations used in the simulations); in all simulations, the number of chains was taken as the number of chains in the multichain complex.

Combining eqs 1, 3, and 4, we obtain the multiple-chain UNRES potential energy (eq 6)

$$U = \sum_k U_{\mathrm{single\ chain}}^k + \sum_k \sum_{l>k} U_{\mathrm{interchain}}^{k,l} + U_{\mathrm{conf}} \quad (6)$$

where the indices *k* and *l* run through the different chains.

2.2. Equations of Motion. To find the time evolution of a system, it is necessary to solve the equations of motion of the system. In general, for a system with generalized coordinates q_1, q_2, ..., q_{n} and generalized momenta $\dot{q}_1$, ..., $\dot{q}_n$, this is equivalent to solving the set of Lagrange's equations

$$\frac{\mathrm{d}}{\mathrm{d}t}[\nabla_{\dot{q}_i} L(q_1, q_2, ..., \dot{q}_1, \dot{q}_2, ...)] - \nabla_{q_i} L(q_1, q_2, ..., \dot{q}_1, \dot{q}_2, ...) = Q_i \quad (7)$$

where $i = 1, ..., n$, *L* is the Lagrangian of the system, and the Q_i's are the generalized dissipative (Rayleigh) forces acting on the system.

The Q_i's are nonconservative forces and, therefore, cannot be derived from the potential energy of the system. For our system, these nonconservative forces are the friction and stochastic forces; they represent collisions with the solvent molecules due to a net motion of the system and random impact of the fluctuating solvent molecules on the solute molecules, respectively, as well as the net effect of averaging out the internal secondary degrees of freedom of the protein molecule. Each Cartesian component of each generalized force will have the form

$$Q_i = -\gamma_i v_i(t) + f_i^{\mathrm{rand}} \quad (8)$$

with γ_i and $v_i(t)$ being the friction coefficient and velocity related to the *i*th coordinate and f_i^{rand} being a stochastic force with zero mean and intensity given by[27] eq 9

$$\langle f_i^{\mathrm{rand}}(t) f_i^{\mathrm{rand}}(t+\tau)\rangle = 2\gamma_i R T_0 \delta(\tau)\delta_{ij} \quad (9)$$

where *R* is the universal gas constant, T_0 is the temperature of the bath, $\delta(\tau)$ is the Dirac delta function (evaluated at an arbitrary time interval τ), and δ_{ij} is the Kronecker delta function. When the Q_i's are identified with the sum of the stochastic and friction forces, they account for the coupling of the protein chain(s) under study to the solvent, which in turn acts as a thermostat, thereby maintaining an average constant temperature of the system.

Following previous work,[11] we chose to describe each chain by a set of virtual bond vectors $\mathbf{dC}_i{}^{\mathrm{k}}$ and $\mathbf{dX}_i{}^{\mathrm{k}}$, with $\mathbf{dC}_i{}^{\mathrm{k}}$ being the vector pointing from $\mathrm{C}_i{}^k$ to $\mathrm{C}_{i+1}{}^k$, except for $\mathbf{dC}_0{}^k$ which points from the origin to the first C$^\alpha$ in the chain, and $\mathbf{dX}_i{}^{\mathrm{k}}$ being the vector pointing from $\mathrm{C}_i{}^k$ to $\mathrm{SC}_i{}^k$ (Figure 1). The superscript *k* indicates the chain to which reference is being made. The entries corresponding to glycine residues are omitted from the list of **dX**'s since they have zero length. A "dummy" C$^\alpha$ atom is introduced at the beginning (end) of the chain if the first (last) residue is not glycine and if the chain is unblocked.[15]

To simplify the notation, the $\mathbf{dC}_i{}^{\mathrm{k}}$ and $\mathbf{dX}_i{}^{\mathrm{k}}$ vectors will be grouped in a single vector $\mathbf{q}^k = (\mathbf{dC}_0{}^k, \mathbf{dC}_s{}^k, ..., \mathbf{dC}_e{}^k, \mathbf{dX}_1{}^k, \mathbf{dX}_2{}^k, ..., \mathbf{dX}_m{}^k)^{\mathrm{T}}$. The indices *s* and *e* correspond to the first and last real residue, i.e., $s = 1$ if the first residue is Gly and $s = 2$ otherwise. Likewise, if the last residue is a dummy one, then the index $e = n - 1$, with *n* being the number of residues in the chain, and $e = n$ otherwise. The index *m* is the number of non-glycine residues in the chain. It should be noted that, although we have omitted the superscripts, the values of *s*, *e*, *n*, and *m* might in principle be different for different chains within the complex.

The coordinates $\mathbf{x}_{\mathrm{p}_i{}^k}$ and $\mathbf{x}_{\mathrm{SC}_i{}^k}$ of the united peptide groups and side chains can be reconstructed from the $\mathbf{dC}_i{}^k$ and $\mathbf{dX}_i{}^k$ vectors through eqs 10 and 11

$$\mathbf{x}_{\mathrm{p}_i{}^k} = \mathbf{dC}_0{}^k + \sum_{j=s}^{j=i-1} \mathbf{dC}_j{}^{\mathrm{k}} + \frac{1}{2}\,\mathbf{dC}_i{}^{\mathrm{k}} \qquad (10)$$

$$\mathbf{x}_{\mathrm{SC}_i{}^{\mathrm{k}}} = \mathbf{dC}_0{}^k + \sum_{j=\mathrm{s}}^{j=\mathrm{i}-1} \mathbf{dC}_j{}^k + \mathbf{dX}_i{}^k \qquad (11)$$

Defining vectors $\mathbf{x}^k = (\mathbf{x}^k_{\mathrm{p}_s}, ..., \mathbf{x}^k_{\mathrm{p}_e}, \mathbf{x}^k_{\mathrm{SC}_1}, ..., \mathbf{x}^k_{\mathrm{SC}_m})$, eqs 10 and 11 can be expressed in matrix form, obtaining a single equation for each chain

$$\mathbf{x}^k = \mathbf{A}^k\mathbf{q}^k \qquad (12)$$

where $\mathbf{A}^k$ is the matrix that transforms from the generalized coordinates $\mathbf{q}^k$ of the kth chain to the Cartesian coordinates of the interacting sites $\mathbf{x}^k$ of the same chain. The same relation holds for the velocities $\mathbf{v}^k = (\mathbf{v}^k_{\mathrm{p}_s}, ..., \mathbf{v}^k_{\mathrm{p}_e}, \mathbf{v}^k_{\mathrm{SC}_1}, ..., \mathbf{v}^k_{\mathrm{SC}_m})$

$$\mathbf{v}^k = \mathbf{A}^k\dot{\mathbf{q}}^k \qquad (13)$$

Then, when solving Lagrange's equations, we obtain a relation, for each chain, of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\,[\nabla_{\dot{\mathbf{q}}^k} K^k\,(\mathbf{q}^k, \dot{\mathbf{q}}^k)] + \nabla_{\mathbf{q}^k} U\,(\mathbf{q}^1, \mathbf{q}^2, ..., \mathbf{q}^N) = \mathbf{f}^{\mathrm{fric}_k} + \mathbf{f}^{\mathrm{rand}_k} \qquad (14)$$

where k indicates the chain in question, K^k is its kinetic energy, $\mathbf{f}^{\mathrm{fric}_k}$ and $\mathbf{f}^{\mathrm{rand}_k}$ are the friction and random forces acting on that chain, and N is the total number of chains in the protein.

The different chains are coupled only through the UNRES potential energy U, which also includes the free energy of the solvent implicitly in the $U_{\mathrm{SC}_i\mathrm{SC}_j}$ terms. The kinetic energy of a specific chain does not contain any dependence on the coordinates from a different chain. This enabled us to easily generalize the single-chain equations derived in refs 11 and 12 to the multichain problem. We obtained the set of equations

$$\ddot{\mathbf{q}}^k = -[\mathbf{G}^k]^{-1}\,\nabla_{\mathbf{q}^\mathbf{k}} U\,(\mathbf{q}^1, \mathbf{q}^2, ..., \mathbf{q}^N) - [\mathbf{G}^k]^{-1}\,[(\mathbf{A}^k)^{\mathrm{T}}\Gamma^{\mathbf{k}}(\mathbf{A}^k)]\dot{\mathbf{q}}^k + [\mathbf{G}^k]^{-1}\,(\mathbf{A}^k)\mathrm{T}\mathbf{f}^{\mathrm{rand}_k} \qquad (15)$$

where Γ^k is a diagonal matrix containing the friction coefficients of the interacting sites (peptide groups and side chains) and $\mathbf{G}^k$ is the inertia matrix, defined by eq 16

$$\mathbf{G}^k = (\mathbf{A}^k)^{\mathrm{T}}\mathbf{M}^k\,(\mathbf{A}^k) + \mathbf{H}^k \qquad (16)$$

where $\mathbf{M}^k$ is a diagonal matrix containing the masses of the interacting sites and $\mathbf{H}^k$, also a diagonal matrix, is the part of the inertia matrix corresponding to the internal stretching of the virtual bonds. $\mathbf{M}^k$ and $\mathbf{H}^k$ are defined by eqs 28 and 29, respectively, of ref 11. Details of the derivation of eqs 15 and 16 can be found in refs 11 and 12.

The components of the vector of random forces are calculated from a normal distribution according to[28−30]

$$(\mathbf{f}^{\mathrm{rand}_k})_i = \sqrt{\frac{2\gamma_i RT}{\delta t}}\,\mathbf{N}(0,1) \qquad (17)$$

where $(\mathbf{f}^{\mathrm{rand}_k})_i$ is the random force acting on the ith site from chain k, γ_i is the friction coefficient associated with that site, R is the universal gas constant, T is the temperature of the bath, δt is the integration time step, and $\mathbf{N}(0,1)$ is a tridimensional normal distribution with zero mean and unit variance.

2.3. Simulations in the Microcanonical Ensemble. As in our previous work,[11] we first carried out MD calculations in the microcanonical ensemble. In this case, the stochastic and the friction forces are set to zero; therefore, the total energy of the system should be conserved. To check that the total energy condition was satisfied, we carried out simulations on two chains of an unblocked Ala_{10} polypeptide with the variable time step as described in section 3 of ref 11. The simulations showed that the fluctuations in the total energy are negligible when compared with those in the kinetic and potential energies. The total energy is conserved, although only to the extent that it is conserved in ref 11. The results of the microcanonical simulations are not shown here since that is an issue that has already been addressed in ref 11.

2.4. Simulations in the Canonical Ensemble. The microcanonical picture of a completely isolated system in which all the forces are known does not always correspond to typical experimental conditions. For this reason, the canonical ensemble, NVT, for which the temperature (i.e., the average kinetic energy) of the system remains constant, is a more desirable choice for carrying out MD simulations.

2.4.1. Langevin Dynamics (LD). As pointed out in section 2.2, the system can be kept at a constant temperature by inserting stochastic and friction terms in the equations of motion, yielding a Langevin equation, namely, eq 15. The trajectory of the system is obtained by numerical integration of eq 15, as described in eqs 12−17 of ref 12.

2.4.2. Berendsen Dynamics (BD). There are other methods to maintain a constant temperature heat bath with less computational effort. These methods can be classified in two large groups: extended Lagrangian methods[31,32] and rescaling of velocities.[27,33] The method that we chose for our MD simulations belongs to the second category and is known as the Berendsen thermostat.[27] The idea behind this method is that the system is forced to have the same kinetic energy as if it were subject to the forces in eq 8. To accomplish this, the velocities are rescaled by a factor

$$\lambda = \left[1 + \frac{\delta t}{\tau_{\mathrm{T}}}\left(\frac{T_0}{T(t)} - 1\right)\right]^{1/2} \qquad (18)$$

at every simulation step, where δt is the time step, T_0 is the reference temperature, τ_{T} is an adjustable parameter (known as the time constant of the thermostat), and $T(t)$, the instantaneous temperature of the system at time t, is given by eq 19

$$T(t) = \frac{2K(t)}{RD} \qquad (19)$$

where $K(t)$ is the kinetic energy of the system, R is the universal gas constant, and D is the number of degrees of freedom of the system. As a result, the system is globally coupled to a heat bath at temperature T_0. Although this method has not been proven to generate a true canonical ensemble, it has the advantage that the coupling can be made as weak as desired by manipulating the constant τ_{T}. It has been shown[27] that small values of τ_{T} (strong coupling) reduce the fluctuations in the kinetic energy K at the expense of increasing fluctuations in the total energy E. On the basis of our earlier work,[11] we set $\tau_{\mathrm{T}} = 48.9$ fs $= 1$ mtu (molecular time unit) and $\delta t = 0.05$ mtu. These values were tested by carrying out MD simulations with Berendsen dynamics on a system composed of two chains of an unblocked Ala_{10} polypeptide at a concentration of 1 mM. During the simulations, the fluctuations in the total (E), kinetic (K), and potential (U) energy were monitored. The simulations showed that the parameters used for the single chain were appropriate for the multichain complex as well.

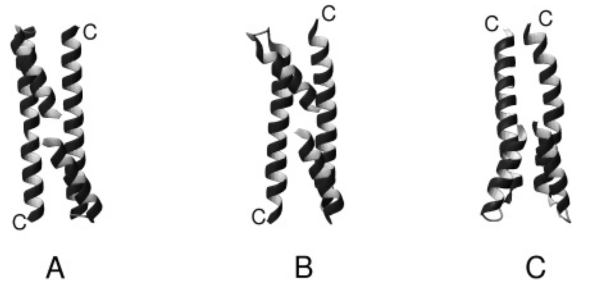

Figure 2. (A) Experimental structure of 1G6U. (B) The most nativelike structure (C$^\alpha$ rmsd = 1.79 Å) obtained with BD UNRES/MD. (C) An example of a misfolded structure. The C-terminus of each chain is marked.

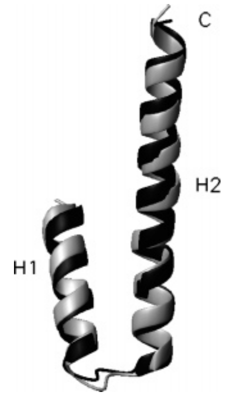

Figure 3. Superposition of one of the monomers in the 1G6U experimental dimer structure (black) on the most nativelike structure (gray) (C$^\alpha$ rmsd = 1.22 Å) obtained with the UNRES/MD simulations of the monomer using BD. The N-terminal helix H1 and the C-terminal helix H2 are indicated as well as the C-terminus.

The parameters and weights in UNRES have been determined by a hierarchical optimization method.[19,21,22,34] The idea behind this method is to reproduce a funnel-like energy landscape with energy decreasing as the number of nativelike elements in a structure increases.[19,34] Because the 4P force field was designed to find nativelike structures as global minima in the potential energy surface, the free-energy gaps between the nativelike structures and the lowest-energy non-native structure of the training protein were overemphasized in the optimization process.[23] Consequently, the optimal folding temperature for the MD simulations with the UNRES 4P force field turned out to be 800 K.[13] This value gave the best compromise between folding time and stability of the nativelike structures for several benchmark proteins.[13] This high temperature was not a problem while carrying out single-chain simulations because the internal forces acting on a polypeptide chain were tuned to this high temperature. However, in multichain simulations, the chains move with respect to each other, and the external motions are too strong to allow association. Therefore, we rescaled all energy term weights by a factor of $^3/_8$ to reduce the folding temperature to 300 K. This operation changes only the energy scale but not the structure of the energy landscape.

3. Results

To study different aspects of the UNRES/MD multiple-chain implementation, we carried out a number of tests. We compared Langevin (LD) and Berendsen dynamics (BD) by carrying out multiple-chain simulations with the same initial conditions with each method. To test whether the presence of other chains was a necessary condition to fold the monomers, we also carried out single-chain simulations with BD and LD and compared the structures obtained with those of the monomers in the crystal structures of the oligomers. Finally, since the method failed to predict the native structure of 1C94, additional simulations starting from the PDB structure were carried out for this protein. This was done to check whether the native structure was not found because of insufficient simulation time or because the force field was not good enough to properly represent the energy landscape of this protein.

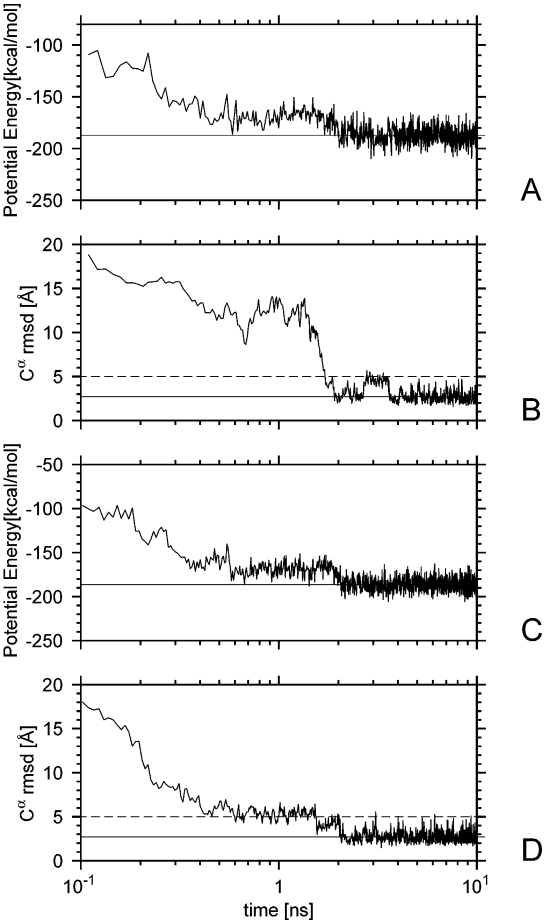

Figure 4. (A) Variation of the potential energy and (B) the C$^\alpha$ rmsd from the native structure of the monomer in the dimer during the folding of an isolated monomer of 1G6U obtained with Langevin dynamics. The solid horizontal line at −187.1 kcal/mol in panel A is the mean value of the energy after the monomer has reached the native basin. In panel B, the dashed horizontal line at 5 Å corresponds to the cutoff rmsd above which the monomer structure is considered to have left the native basin, and the solid horizontal line at 2.7 Å indicates the mean C$^\alpha$ rmsd of the monomer inside the native basin. Panels C and D contain the same information as panels A and B, respectively, for a trajectory obtained with Berendsen dynamics. The solid horizontal line at −187.1 kcal/mol in panel C is the mean value of the energy after the monomer has reached the native basin, and the solid horizontal line at 2.7 Å in panel D is the mean C$^\alpha$ rmsd inside the native basin of the monomer from the monomer in the native structure of the dimer.

All the runs (both single-chain and multichain), except those starting from the PDB structure, were started with the chains in an extended conformation. In all cases, the initial velocities of the peptide groups and side chains were randomly generated. In the multichain runs, the chains were placed parallel to each other, separated by a distance large enough (20 Å for GCN4-p1 and 1C94 and 40 Å for 1G6U) to allow them to rearrange independently. Since the chains rapidly adjust to an equilibrium ensemble, after starting from extended conformations, the simulations are practically independent of the starting condition. The initial velocities were selected from a Gaussian distribution

TABLE 1: Summary of Trajectories for 1G6U

algorithm	N_f[a]	$\langle\tau_f\rangle$[b] (ns)	$\langle\tau_f(H1)\rangle$[c] (ns)	$\langle\tau_f(H2)\rangle$[d] (ns)	ρ_{min}[e] (Å)	$\langle\tau_{res}\rangle$[f]	$\langle E\rangle_f$[g] (kcal/mol)	N_{mf}[h]	$\langle E\rangle_{mf}$[i] (kcal/mol)	CPU time[j] (h)
					Dimer					
Berendsen	9(20)	4.8(0.30)	0.14	0.16	1.79	49%	−402	1	−401	2.9
Langevin	2(19)	14.9(4.0)	0.35	0.20	2.38	36%	−403	2	−398	3.9
					Monomer					
Berendsen	10	0.92	0.18	0.21	1.22	86%	−186			
Langevin	10	2.6	0.25	0.24	1.28	69%	−188			

[a] Number of trajectories (out of 10) that folded to nativelike structures. In the dimer simulations, the number of monomers (out of 20, since there were 2 monomers on each of the 10 dimer simulations) that folded to a nativelike structure is indicated between parentheses. [b] Average folding time. The folding time was defined as the time at which the rmsd with respect to the crystal structure fell below the cutoff value (7 Å for the dimers and 5 Å for the monomers). In those runs for which the rmsd never went below the cutoff, the folding time was considered to be the simulation time (12 ns for Berendsen and 16 ns for Langevin). In the dimer simulations, the average folding time of the monomers is indicated between parentheses. [c] Average folding time for the N-terminal helix, H1. The folding time was defined as the time at which the rmsd with respect to the crystal structure fell below 1.5 Å. [d] Average folding time for the C-terminal helix, H2. The folding time was defined as the time at which the rmsd with respect to the crystal structure fell below 4 Å. [e] The lowest rmsd in all of the fluctuating trajectories. [f] Fraction of the time that the peptide spent in the native basin averaged over all of the folding trajectories. [g] Average potential energy over all structures in the native f basin. [h] Number of trajectories (out of 10) that yielded misfolded structures; [i]Average potential energy over all structures in the misfolded mf basin; [j] Average CPU time (in hours) per 1 ns of simulation on a single 3.06 GHz Intel Pentium IV Xeon processor.

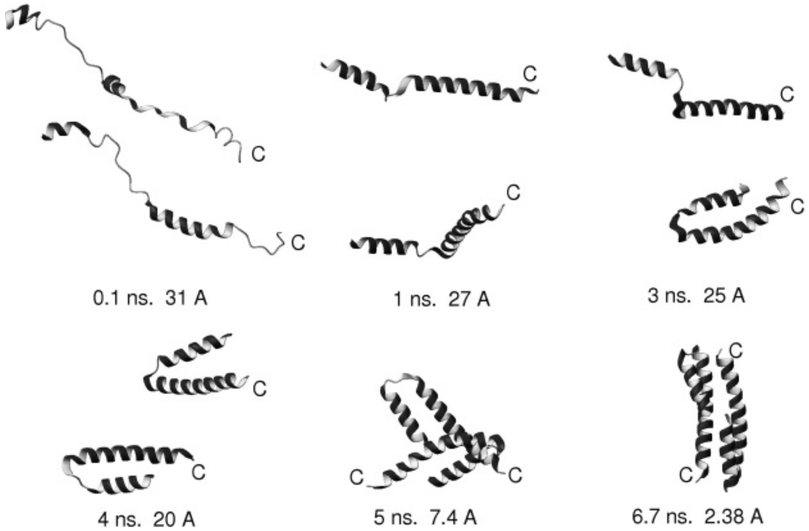

Figure 5. Example of a successful trajectory of 1G6U obtained with Langevin dynamics. The C-terminus of each chain is marked.

corresponding to the average kinetic energy at the simulation temperature, as in our earlier work,[11] and the temperature was held constant at 300 K during all of the simulations.

In the multichain runs, for those starting from the extended conformation, the radius of the confining sphere was initially set large enough to fit the extended chains. After the first 24 ns of simulation, the radius of the sphere was decreased slowly until the desired concentration (1 mM for the dimers and 10 mM for the tetramer) was reached. This concentration, although higher than those concentrations used in the experiments,[35−37] was chosen because it resulted in a volume large enough to fit the chains without altering their structures and small enough for the monomers to find each other and interact in a short period of time.

To classify the runs into success and failure, we monitored the C^α root-mean square devation (rmsd) between the computed structures and the crystal structure. If this value, hereafter referred to as ρ, fell below a cutoff value, ρ_{cut}, then the protein was considered to have folded. The folding time τ_f, defined as the time at which ρ fell below the cutoff ρ_{cut} for the first time, and the residence time τ_{res}, defined as the fraction of the total time that ρ was below ρ_{cut}, were also computed. For 1G6U, ρ_{cut} was 5 Å for the monomers and 7 Å for the dimers, for GCN4, ρ_{cut} was 3.4 Å for the monomers and 4.8 Å for the dimers, and for 1C94, ρ_{cut} was 4 Å for the monomers, 5.6 Å for the dimers, and 8 Å for the tetramers. If the monomers were folded by this criterion and were stable, and the arrangement of the chains was stable but not native, then the overall structure was classified as misfolded. If this criterion was not met, then the structure was classified as nonfolding.

3.1. Domain Swapped Dimer (PDB Code 1G6U). 1G6U is a synthetic α-helical homodimer with 48 residues per chain.[37] Each monomer consists of two α-helix segments, with the shortest (14 residues) helix packed against the longest (28 residues) helix. The monomers assemble forming a three-α-helix bundle with the long helices in the antiparallel position (Figure 2A). We will refer to the shortest helix as H1 and the longest helix as H2 (Figure 3). To provide a better description of the folding trajectories, we monitored the rmsd (Table 1) with respect to the native structure for the entire protein, for each of the monomers, and for each of the helices (H1 and H2). To determine the folding times of H1 and H2, we set their cutoff rmsd's at 1.5 and 4 Å, respectively.

3.1.1. Monomers. As can be seen in Table 1, all of the simulations of the monomers converged to nativelike structures, showing that dimerization is not necessary for the folding and stabilization of the individual chains. The most nativelike structure, 1.22 Å from native, was produced by BD. A superposition of this structure and the native structure is shown in Figure 3.

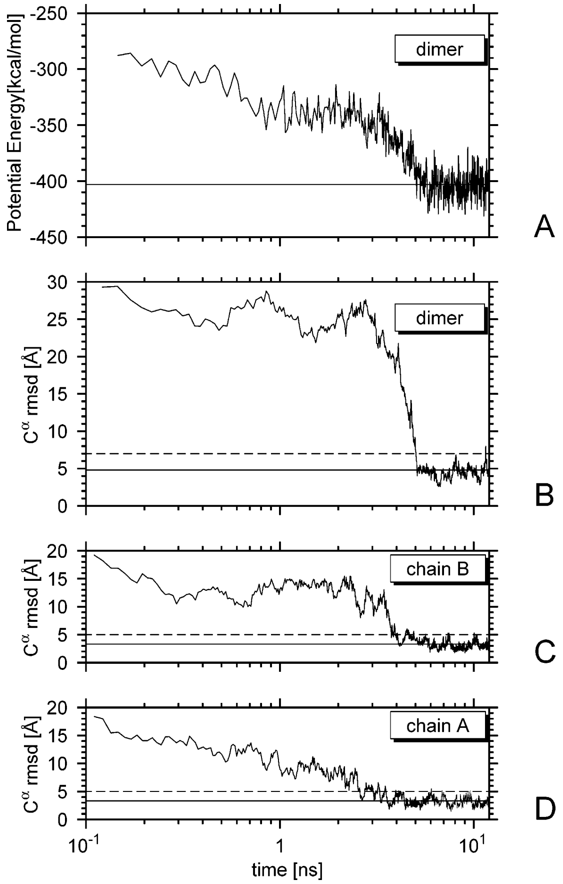

Figure 6. (A) Variation of the potential energy and (B) the C^{α} rmsd from the native structure for the dimer in a successful trajectory of 1G6U obtained with Langevin dynamics. For the same trajectory, panels C and D show the variation of the C^{α} rmsd from the native for each of the monomers. The solid horizontal line at −403 kcal/mol in panel A is the mean value of the energy after the dimer has reached the native basin, and the solid line at 4.8 Å in panel B is the mean C^{α} rmsd inside the native basin of the dimer. The dashed horizontal line in panels B, C, and D corresponds to the cutoff rmsd (7 Å for the dimer and 5 Å for the monomers) above which a structure is considered to have left the native basin. The solid horizontal line at 3.3 Å in panels C and D is the mean C^{α} rmsd inside the native basin of the monomer.

Figure 4 shows potential energy and ρ values for an LD trajectory (panels A and B, respectively) and a BD trajectory (panels C and D, respectively) for an isolated monomer of 1G6U. As can be seen from Figure 4, the native basin was very stable, and with both methods, once the peptide adopted nativelike structures, the fluctuations in the potential energy and ρ became smaller, and the peptide remained in the native basin.

3.1.2. Dimers. In the simulation of dimers, the initial separation distance between chains was 40 Å, the initial arrangement was parallel, and the simulation time was approximately 12 ns for BD and 16 ns for LD. The final concentration of 1 mM was achieved within the first nanosecond. The results are summarized in Table 1. Both algorithms, BD and LD, folded the protein. In general the folding times with BD were shorter than those with LD, as observed in our earlier work on single-chain proteins.[12] BD also produced the most nativelike structure, which is shown in Figure 2B.

From the simulations, it became evident that the energy landscape generated by the 4P UNRES force field has two basins with low free energy. One of these basins corresponds to the native structure, and the other one to a structure that differs from the native in that the long helices are parallel to each other instead of antiparallel (Figure 2C). We will refer to the latter structure as a misfolded one. Both structures were very stable, and once the protein fell into one of these basins, it would not escape within the simulation time (12 ns for BD and 16 ns for LD). The difference in average potential energy between the native and the misfolded basin is very small (Table 1). Thus, it is natural to expect that, for some trajectories, the forces will drive the system to the native basin and, for some others, to the misfolded basin. Indeed, this is what was observed in these simulations. Presumably, improvement of the 4P UNRES force field will stabilize the native basin to a greater extent compared to the non-native basin.

Snapshots of a successful trajectory obtained with LD are shown in Figure 5. For the same trajectory, the values of ρ and the potential energy as a function of time are shown in Figure 6. The snapshots show that helix formation takes less than 1 ns, and for this particular example, the packing of the helices on both monomers takes about 3 ns. Also for this example, the monomers fold independently, but they are close enough so that, after the subunits have folded, they can overcome the friction forces to turn around (since the initial orientation of the helices is parallel, but in the native structure the orientation is antiparallel) and assemble in less than 2 ns. The folding of the

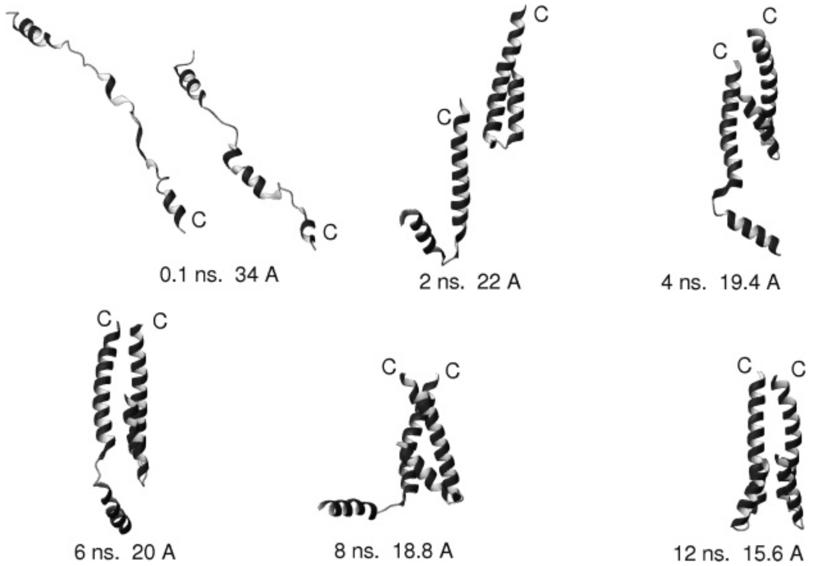

Figure 7. Example of a trajectory of 1G6U, obtained with Langevin dynamics, leading to the misfolded structure. The C-terminus of each chain is marked.

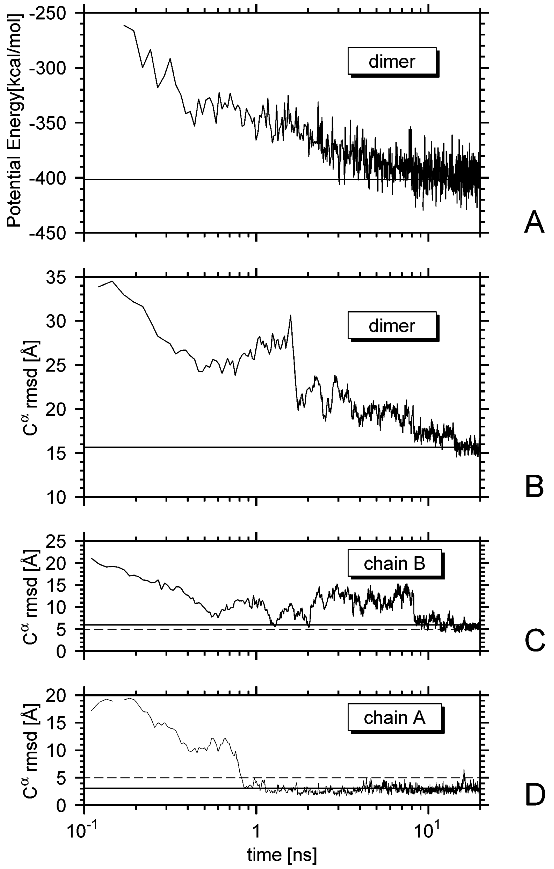

Figure 8. (A) Variation of the potential energy and (B) the C$^\alpha$ rmsd from the native structure for the dimer for a misfolding trajectory of 1G6U obtained with Langevin dynamics. The misfolded structure differs from the native in that the long helices are parallel to each other instead of antiparallel. For the same trajectory, panels C and D show the variation of the C$^\alpha$ rmsd from the native for each of the monomers. In panels A and B, the solid horizontal line (at −401 kcal/mol in panel A and 15.7 Å in panel B) is the mean value of the energy and the C$^\alpha$ rmsd from the native, respectively, after the protein has fallen into the misfolded basin. The dashed horizontal line in panels C and D corresponds to the 5 Å cutoff rmsd, above which the monomers are considered to have left the native basin; i.e., the monomers folded but the overall structure was misfolded. The solid horizontal line in panels C and D (at 5.9 Å in panel C and 3.1 Å in panel D) is the mean C$^\alpha$ rmsd inside the native basin of the monomer.

dimer is completed in a total time of 5 ns. The two LD trajectories that converged to the native basin (Table 1) showed the folding mechanism illustrated in Figure 5.

Figure 7 shows snapshots of an LD trajectory leading to a misfolded structure. The values of ρ and the potential energy for this trajectory are shown in Figure 8. For this particular trajectory, chain A folds first (cf. panels C and D), and chain B folds while it binds to form the dimer (cf. panels B and C). The formation of the dimer in Figure 8 corresponds to the stabilization of ρ around 15.6 Å in panel B. For the other LD trajectory that converges to the misfolded basin, the assembly mechanism was similar to that described in Figure 5, in the sense that the monomers folded completely before they assembled. Thus, folding of the monomers followed by their assembly does not always lead to the native basin.

Those LD trajectories that did not converge to the native or misfolded basin reached a state (called nonfolded) in which either one or both monomers were folded, but they had not yet assembled within the 16 ns simulation time. Their structures were similar to either the 3- or the 4-ns snapshot in Figure 5.

With BD, all of the simulations converged to either the native or the misfolded basin (Table 1). Among those runs that converged to the native basin, two different pathways were observed, one on which the subunits fold before their assembly ("lock-and-key" mechanism) and another one on which the subunits fold simultaneously with their assembly ("induced-fit" mechanism). Although only a few runs followed the latter assembly mechanism (3 out of 9 folding trajectories), this pathway seems to be 2.5 times faster on average than the assembly of already folded subunits, which is not surprising since, after the monomers are folded, they might collide several times until they find the right orientation, which will in general slow down the process. Figures 9 and 10 illustrate these two folding pathways. For the trajectory shown in Figure 9 (fast folding pathway), folding and association of the chains occurs simultaneously, with the dimer folding in less than 0.4 ns, while for the trajectory shown in Figure 10 (slow folding pathway), although the chains collide several times (snapshots at 0.23, 0.33, 5.83, and 6.44 ns), only the last collision results in the formation of the dimer. There is a long period between the snapshots at 0.56 and 5.83 ns (this period is not shown in the snapshots) during which the chains remain folded but they do not collide at all. Figure 11 contains the values of ρ and the potential energy as a function of time, corresponding to the trajectory shown in the snapshots in Figure 10. In Figure 11, two pronounced drops in energy can be seen (panel A). The first one corresponds to the folding of the monomers (ρ below 5 Å in panels C and D), and the second one corresponds to the assembly of the dimer (ρ below 7 Å in panel B).

The folding mechanism of the only BD trajectory that converged to the misfolded basin was similar to the one described in Figure 9 (fast folding pathway), except that the orientation of the chains was parallel instead of antiparallel as in the native structure.

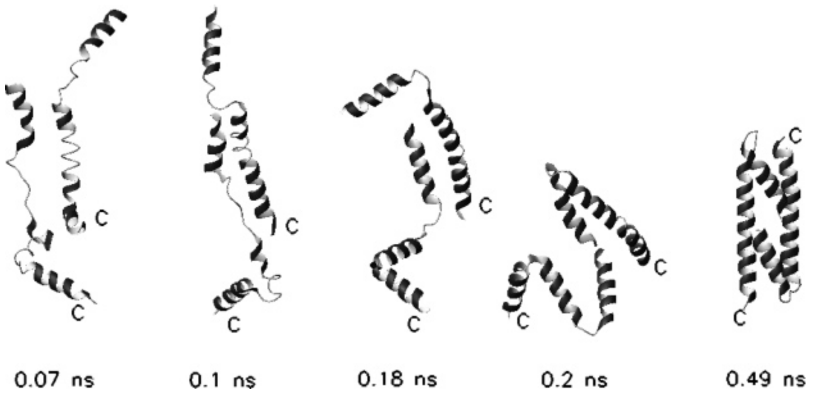

Figure 9. Example of a fast folding trajectory of 1G6U obtained with Berendsen dynamics. The C-terminus of each chain is marked.

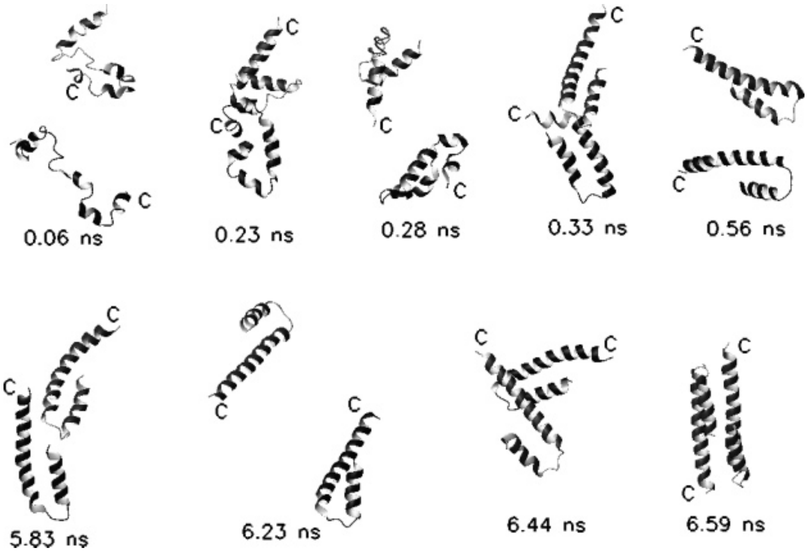

Figure 10. Example of a slow folding trajectory of 1G6U obtained with Berendsen dynamics. The C-terminus of each chain is marked.

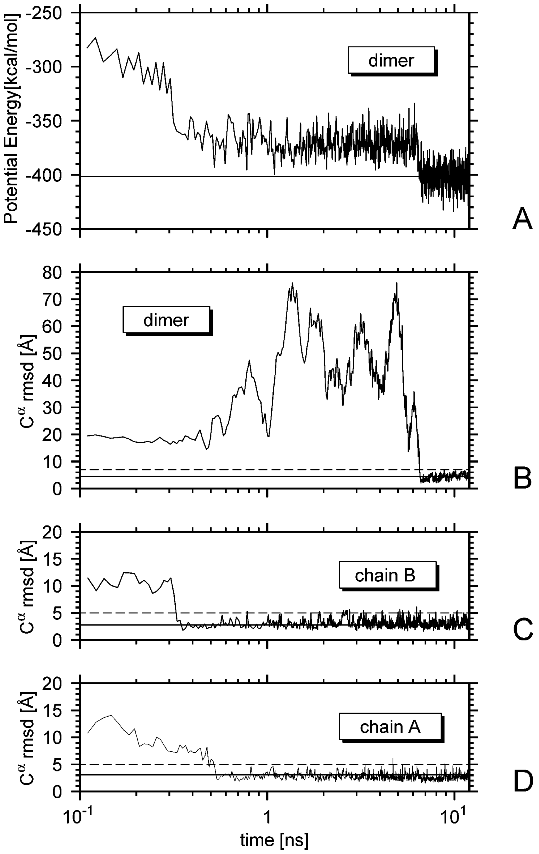

Figure 11. (A) Variation of the potential energy and (B) the C^α rmsd from the native structure for the dimer in a successful trajectory of 1G6U obtained with Berendsen dynamics. For the same trajectory, panels C and D show the variation of the C^α rmsd from the native structure for each of the monomers. The solid horizontal line at −401 kcal/mol in panel A is the mean value of the energy after the protein has reached the native basin. The dashed horizontal line in panels B, C, and D corresponds to the cutoff rmsd (7 Å for the dimer and 5 Å for the monomers) above which a structure is considered to have left the native basin. The solid horizontal line at 4.5 Å in panel B is the mean C^α rmsd inside the native basin of the dimer. The solid horizontal line in panels C (2.8 Å) and D (3.1 Å) is the mean C^α rmsd inside the native basin of the monomer.

When comparing the folding of the isolated monomers of 1G6U in the single- and multichain simulations, we found that, with LD, the average folding time of the monomers in the single-chain simulations was shorter than that in the multichain simulations (Table 1), which suggests that the interactions between chains might slow down the folding of the individual chains. To further elucidate whether this delay occurs in the formation of helices H1 and H2 or in their packing, we compare the folding times of H1 and H2 in the single-chain simulations with their folding times in the multichain simulations. We found almost no difference in the average folding time of H2, and in the case of H1, the formation of the helix seems to be slightly faster for the single-chain simulations (Table 1). This suggests that, for 1G6U with LD, the interactions between the chains can hinder the packing of helices H1 and H2 and can also slow down the formation of the shortest helix (H1).

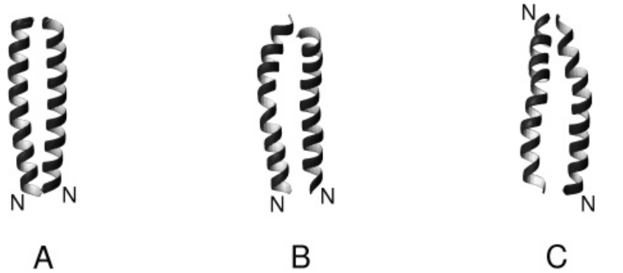

Figure 12. (A) Experimental structure of GCN4-p1. (B) The most nativelike structure (C^α rmsd = 1.19 Å) obtained with LD UNRES/MD. (C) An example of a misfolded structure. The N-terminus of each chain is indicated.

With BD, on average, the monomers folded 3 times faster in the multichain simulations than in the single-chain simulations (0.30 ns compared to 0.92 ns) (Table 1). Further analysis of the folding times of helices H1 and H2 showed that H1 and H2 fold at approximately the same rate for single-chain and multichain simulations (Table 1). This indicates that interactions between chains enhance the packing of H1 and H2 but have no substantial effect on the formation of the helical structures.

The fact that the packing of H1 and H2 is favored by multichain interactions with BD and hindered with LD might be explained as follows: With BD, in which the friction forces are absent, the chains can move very fast, and if a collision that does not favor the packing of H1 and H2 has taken place, then the chains can quickly rearrange to find a better orientation while, with LD, the reorientation of the chains is much slower due to the friction forces from the solvent. With both methods,

TABLE 2: Summary of Trajectories for GCN4-p1

algorithm	N_f[a]	$\langle\tau_f\rangle$[b] (ns)	ρ_{min}[c] (Å)	$\langle\tau_{res}\rangle$[d]	$\langle E\rangle_f$[e] (kcal/mol)	N_{mf}[f]	$\langle E\rangle_{mf}$[g] (kcal/mol)	CPU time[h] (hs)
				Dimer				
Berendsen	4(17)	6.6(3.2)	1.22	29%	−214	6	−217	1.5
Langevin	3(16)	9.1(3.4)	1.19	81%	−218	1	−225	1.9
				Monomer				
Berendsen	9	1.5	0.59	69%	−104			
Langevin	9	2.7	0.70	74%	−97			

[a] Number of trajectories (out of 10) that folded to nativelike structures. In the dimer simulations, the number of monomers (out of 20, since there were 2 monomers on each of the 10 dimer simulations) that folded to a nativelike structure is indicated between parentheses. [b] Average folding time. The folding time was defined as the time at which the rmsd with respect to the crystal structure fell below the cutoff value (4.8 Å for the dimers and 3.4 Å for the monomers). In those runs for which the rmsd never went below the cutoff, the folding time was considered to be the simulation time (12 ns). For both BD and LD, in the dimer simulations, the average folding time of the monomers is indicated between parentheses. [c] The lowest rmsd in all of the fluctuating trajectories. [d] Fraction of the time that the peptide spent in the native basin averaged over all of the folding trajectories. [e] Average potential energy over all structures in the native f basin. [f] Number of trajectories (out of 10) that yielded misfolded structures. [g] Average potential energy over all structures in the misfolded mf basin. [h] Average CPU time (in hours) per 1 ns of simulation on a single 3.06 GHz Intel Pentium IV Xeon processor.

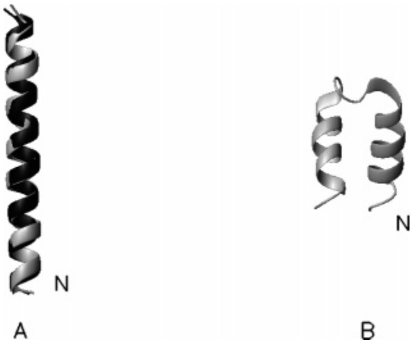

Figure 13. (A) Superposition of one of the monomers from the experimental structure of GCN4-p1 (black) on the most nativelike structure (gray) (C^α rmsd = 0.59 Å) obtained with BD UNRES/MD. (B) A structure that was often found during the folding pathway of GCN4-p1 (with both BD and LD) and was the final structure of those trajectories that did not find the native basin. The N-terminus is indicated.

collisions will sometimes favor the packing of H1 and H2 and other times hamper it, the only difference is that, with BD, the chains can collide more frequently, and overall (when averaged over several trajectories) the presence of another chain will favor single-chain folding.

3.2. GNC4 Leucine Zipper (PDB Code 2ZTA). The GCN4 leucine zipper (GCN4-p1), derived from the yeast transcriptional activator GCN4, is an α-helical homodimer consisting of two parallel chains with 33 residues per chain[35] (Figure 12A). Since the helices in GCN4-p1 wrap around each other, its motif is known as a coiled coil. The coiled coil motif is found in many proteins, and for this reason, GCN4-p1 and its mutants have been the subject of numerous studies.[5,35,38] In particular, simulations of the folding pathway of GCN4-p1 have been carried out by Vieth et al.,[5] as mentioned in the Introduction.

3.2.1. Monomers. With both the BD and LD methods, 9 out of 10 monomer trajectories converged to nativelike structures, as can be seen from Table 2. Moreover, these nativelike structures were quite stable, indicating that dimerization is not necessary for the folding and stabilization of the individual chains. A superposition of the most nativelike structure, obtained with BD, and the experimental structure is shown in Figure 13A. Those BD and LD trajectories that did not find the native basin by the end of the simulation showed structures with ρ values around 11 Å, in which the helix was bent, packing against itself, as shown in Figure 13B.

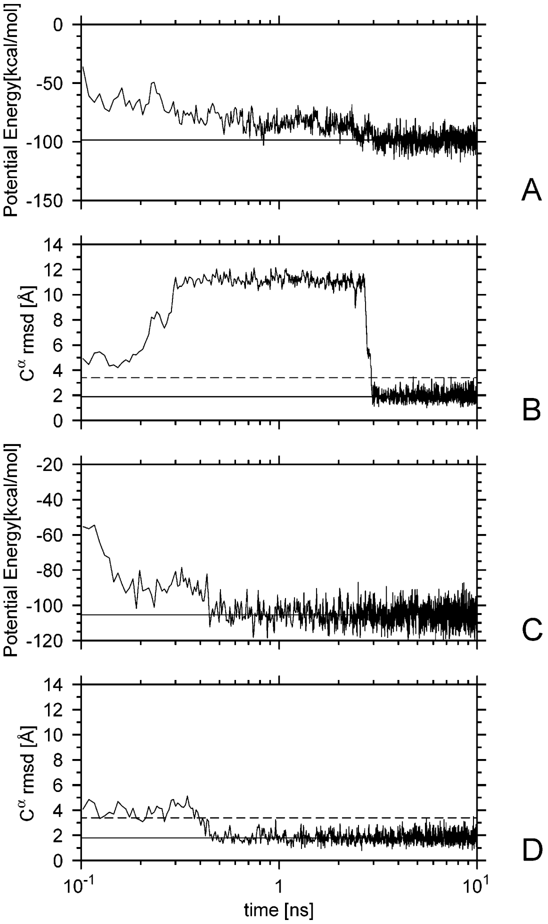

Figure 14. (A) Variation of the potential energy and (B) the C^α rmsd from the native structure of the monomer in the dimer during the folding of an isolated monomer of GCN4-p1 obtained with Langevin dynamics. The solid horizontal line at −98 kcal/mol in panel A is the mean value of the energy after the monomer has reached the native basin. In panel B, the dashed horizontal line at 3.4 Å corresponds to the cutoff rmsd above which the monomer structure is considered to have left the native basin, and the solid horizontal line at 1.9 Å is the mean C^α rmsd of the monomer inside the native basin. Panels C and D contain the same information as panels A and B, respectively, for a trajectory obtained with Berendsen dynamics. The solid horizontal line at −105 kcal/mol in panel C is the mean value of the energy after the monomer has reached the native basin, and the solid horizontal line at 1.8 Å in panel D is the mean C^α rmsd inside the native basin of the monomer from the monomer in the native structure of the dimer.

The structure shown in Figure 13B was also found along the pathway of some of the trajectories that converged to nativelike structures. Potential energy and ρ values as a function of time, for an LD trajectory showing such a behavior, are shown in

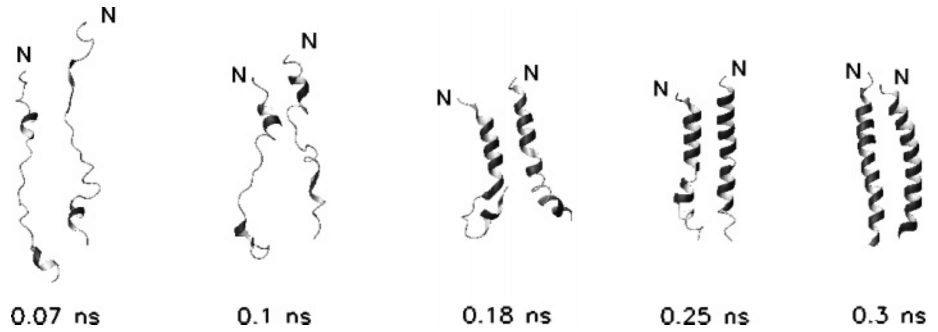

Figure 15. Example of a fast folding trajectory of GCN4-p1 obtained with Langevin dynamics. The N-terminus of each chain is marked.

Figure 14 (potential energy in panel A and ρ values in panel B). During the first 3 ns of simulation of this trajectory, the peptide adopts structures similar to that shown in Figure 13B, which corresponds to the plateau in ρ values around 11 Å in panel B. At the third nanosecond of simulation, the monomer finds the native basin (ρ falls below the 3.4 Å cutoff in panel B), and the energy drops considerably (panel A), showing that the structure in Figure 13B is only a local minimum and does not compete with the native structure.

Not all the trajectories that converged to the native basin exhibited the folding pathway described in the previous paragraph. In other simulations, a fast folding pathway was observed, with the monomer rapidly finding the native basin without spending time in any intermediate structure. An example of such behavior can be seen in the BD trajectory shown in panels C and D of Figure 14 (potential energy in panel C and ρ values in panel D). This behavior was the most commonly observed among all the runs (both BD and LD).

In general, with either BD or LD, the native basin was very stable, which can be inferred from the behavior of ρ in panels B and D of Figure 14; once ρ crossed the 3.4 Å rmsd cutoff (equivalent to finding the native basin), it remained within this cutoff most of the time.

3.2.2. Dimers. The initial separation distance between chains was 26 Å, and the initial arrangement was parallel. Both methods, BD and LD, generated trajectories leading to nativelike structures within 12 ns of simulation. The results are summarized in Table 2. The equilibrium concentration of 1 mM was reached during the first 24 ps of simulation. Again, as for 1G6U, two families of stable structures (corresponding to basins with low free energy) were found; one of them was nativelike, and the other one differed from the native structure in that the orientation of the helices was antiparallel instead of parallel. The most nativelike structure generated by UNRES/MD as well as an example of a misfolded structure are shown in Figures 12B and 12C, respectively.

When running in the LD mode, two different pathways were observed, one on which folding and assembly of subunits were coupled, "induced-fit" mechanism, and another one in which the subunits folded before they assemble, "lock-and-key" mechanism. Of the three LD trajectories that converged to the native basin, two of them folded by the induced fit mechanism and the remaining one by the lock-and-key mechanism. Snapshots from one of the runs that folded by the induced fit mechanism are shown in Figure 15, and the potential energy and ρ values for the same trajectory are shown in Figure 16. In Figure 15, dimerization starts with the association of the small helical segments at the N-termini and propagates toward the C-termini simultaneously with formation of the helices. The two trajectories folding by this mechanism folded in less than 0.3 ns, which was 10 times faster than the trajectory folding by the lock-and-key mechanism. Snapshots from the trajectory folding by the lock-and-key mechanism are shown in Figure 17, and the corresponding potential energy and ρ values as a function of time are shown in Figure 18. In Figure 17, the folding of the helices is almost completed at the 0.20 ns snapshot, but the chains fail to bind and move apart. It takes almost 5 ns more for the chains to find the right orientation and form the dimer. This folding mechanism will in general lead to a larger folding time since, once the individual chains adopt their native structure, moving through the solvent to find the proper packing is difficult, while if the subunits are already attached (in the

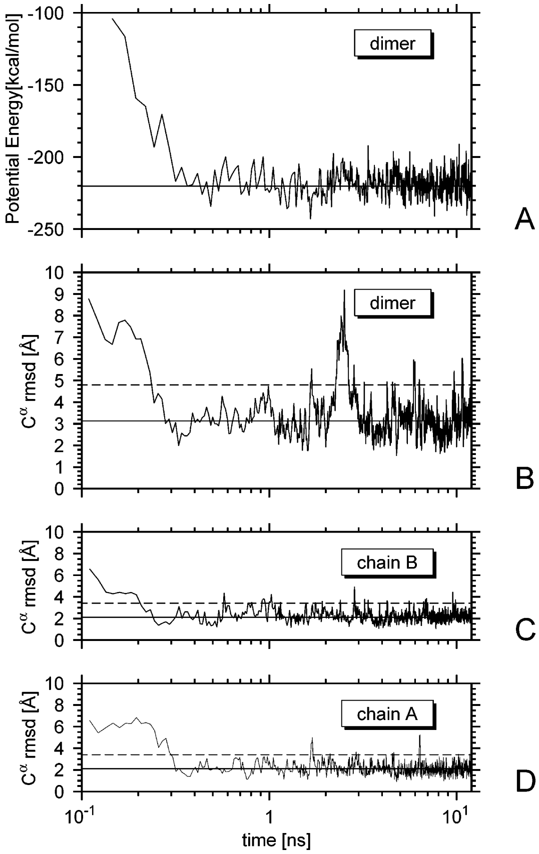

Figure 16. (A) Variation of the potential energy and (B) the C$^\alpha$ rmsd from the native structure for the dimer in a fast folding trajectory of GCN4-p1 obtained with Langevin dynamics. For the same trajectory, panels C and D show the variation of the C$^\alpha$ rmsd from the native structure for each of the monomers. In panel A, the solid horizontal line at −220 kcal/mol is the mean value of the energy after the dimer has reached the native basin. The dashed horizontal line in panels B, C, and D corresponds to the cutoff rmsd (4.8 Å for the dimer and 3.4 Å for the monomers) above which a structure is considered to have left the native basin. The solid horizontal line at 3.1 Å in panel B is the mean C$^\alpha$ rmsd inside the native basin of the dimer. The solid horizontal line in panels C and D (at 2.1 Å in panel C and 2.2 Å in panel D) is the mean C$^\alpha$ rmsd inside the native basin of the monomer.

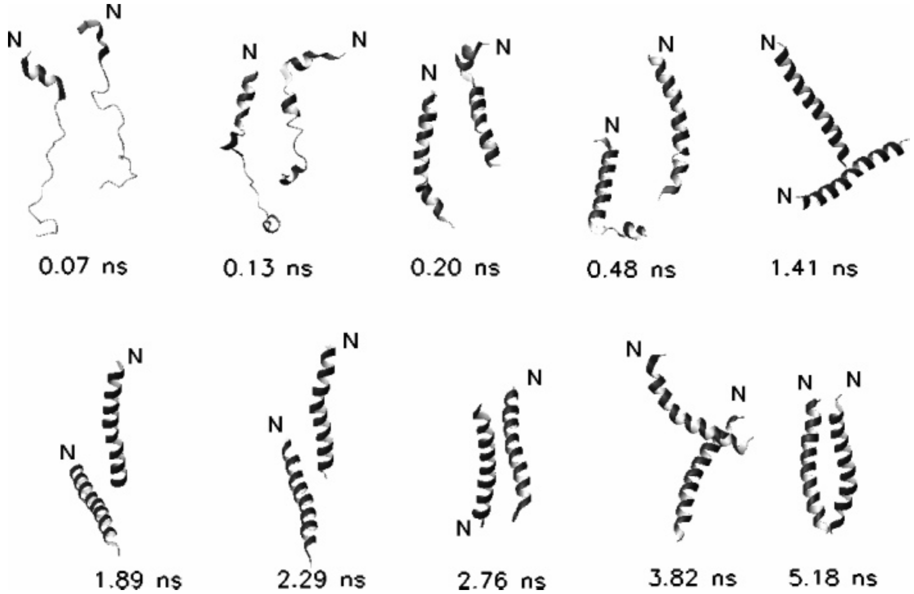

Figure 17. Example of a slow folding trajectory of GCN4-p1 obtained with Langevin dynamics. The N-terminus of each chain is marked.

right place) the rate of folding is limited only by the folding of the individual chains.

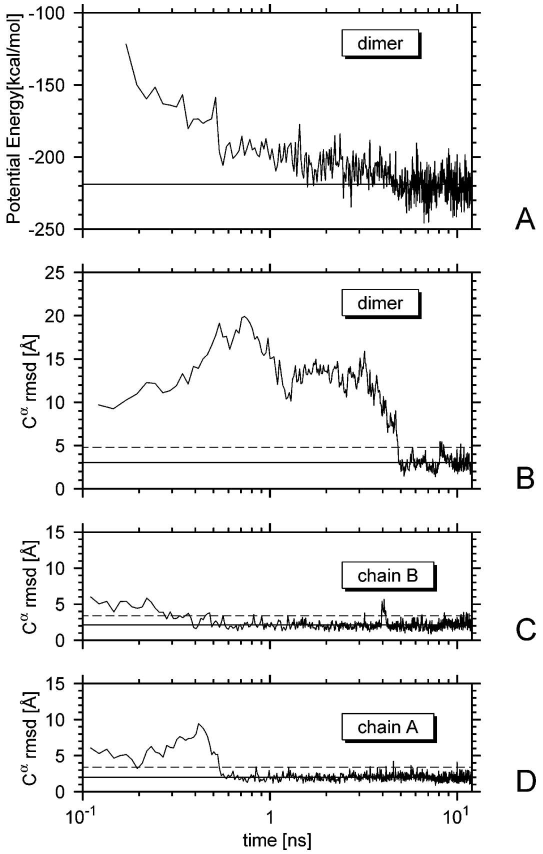

Figure 18. (A) Variation of the potential energy and (B) the C^α rmsd from the native structure for the dimer in the slow folding trajectory of GCN4-p1 obtained with Langevin dynamics. For the same trajectory, panels C and D show the variation of the C^α rmsd from the native structure for each of the monomers. In panel A, the solid horizontal line at −219 kcal/mol is the mean value of the energy after the dimer has reached the native basin. The dashed horizontal line in panels B, C, and D corresponds to the cutoff rmsd (4.8 Å for the dimer and 3.4 Å for the monomers) above which a structure is considered to have left the native basin. The solid horizontal line at 3.0 Å in panel B is the mean C^α rmsd inside the native basin of the dimer. The solid horizontal line in panels C and D (at 2.0 Å in panel C and 2.1 Å in panel D) is the mean C^α rmsd inside the native basin of the monomer.

When running in the BD mode, for some of the trajectories, the protein jumped from one basin to the other one. The potential energy and ρ values for a representative trajectory presenting this behavior are shown in Figure 19. It can be seen that the dimer (panel B) folds and misfolds without affecting the structure of the monomers (panels C and D), which is consistent with our results from single-chain simulations indicating that the monomers are stable by themselves.

As observed for 1G6U, the average potential energies of the native and misfolded basins were very similar (Table 2), the slightly lower values for the misfolded structures being within the expected error in the potential function.

When comparing the folding times for the monomers in the multichain simulations with those in the single-chain simulations, we notice that, with both BD and LD, the isolated monomers fold, on average, slightly faster. A closer look at those monomers that, in multichain simulations, have the largest folding times, or did not fold at all, shows that the folding was delayed because the monomers are trapped in structures similar to that shown in Figure 13B. In all simulations, the dimers were formed, but one or both chains have this bent structure. As already mentioned, this structure was also found along the pathway of some of the trajectories in the simulations of isolated monomers, but the fact that the isolated monomers were able to find the native structure faster indicates that multichain interactions might stabilize the structure shown in Figure 13B.

Those trajectories that did not converge to the native or misfolded basin reached a state (called nonfolded) in which a dimer was formed, but one or both chains had the non-native-like structure shown in Figure 13B.

It should be emphasized that UNRES/MD reflects the energy landscape produced by the UNRES 4P force field. The presence of non-native stable structures is a feature of the force field, not the method. Improvement of the 4P UNRES force field is expected to stabilize the native over the non-native basin to a greater extent.

3.3. Retro-GNC4 Leucine Zipper (PDB Code 1C94). 1C94 is a synthetic α-helical homotetramer of 38 residues per chain. The sequence of 1C94 corresponds to the reversed sequence of the leucine zipper portion of GCN4, viz., GCN4-p1 (section 3.2). GCN4-p1 consists of 33 residues, and 1C94 consists of the same 33 residues but in reversed order from N- to C-terminus; in addition 1C94 is extended at the N-terminus with

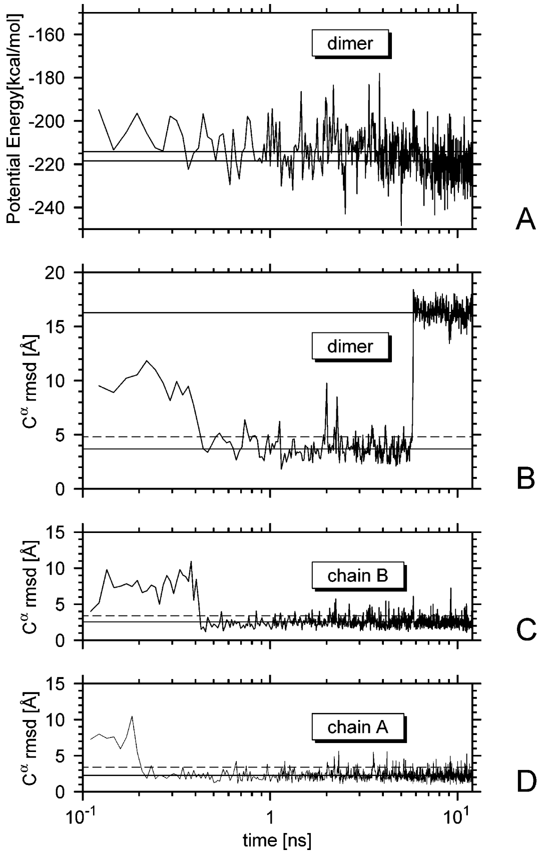

Figure 19. (A) Variation of the potential energy and (B) the C^α rmsd from the native structure for the dimer in a folding trajectory of GCN4-p1 obtained with Berendsen dynamics. For the same trajectory, panels C and D show the variation of the C^α rmsd from the native structure for each of the monomers. The dimer remains in the native basin for almost 5 ns after which it jumps to the misfolded basin. The solid horizontal lines at −214 and −218 kcal/mol in panel A correspond to the mean values of the potential energy inside the native basin and the misfolded basin, respectively. The solid horizontal lines at 3.7 and 16.3 Å in panel B correspond to the mean C^α rmsd inside the native basin and the misfolded basin, respectively. The dashed horizontal line in panels B, C, and D corresponds to the cutoff rmsd (4.8 Å for the dimer and 3.4 Å for the monomers) above which a structure is considered to have left the native basin. The solid horizontal line in panels C (at 2.3 Å) and D (at 2.6 Å) corresponds to the mean C^α rmsd inside the native basin of the monomer.

the tripeptide sequence Cys-Gly-Gly and at the C-terminus with Gln-Leu.[36] Thus, 1C94 is referred to as the retro-GNC4 leucine zipper. The crystal structure, consisting of four α-helices oriented parallel to each other (Figure 20A), was modeled[36] as a dimer of dimers since mass spectroscopic analysis indicated that the chains were covalently linked in pairs by disulfide bonds.[36]

3.3.1. Monomers. As can be seen from Table 3, 9 out of 10 monomer Langevin trajectories and all 10 Berendsen trajectories converged to nativelike structures. The remaining trajectories that did not find the native basin by the end of the simulation showed structures with ρ values around 13 Å where the helix is broken, packing against itself. An example of such a structure is shown in Figure 21B. With an older version of the UNRES force field (α_0 force field[39]), Saunders and Scheraga[25] identified a structure of the type shown in Figure 21B as the lowest UNRES energy structure. With the force field used in this work (4P force field),[23] however, these types of structures have a higher energy than the nativelike structures, as can be seen by comparing the two Langevin trajectories shown in Figure 22. Panels A and B show the energy and ρ values, respectively, for the LD trajectory with final structures similar to that shown in Figure 21B, and panels C and D show the same information for the LD trajectory converging to the native basin. The mean value of the potential energy in the native basin is indicated with the solid line at −152 kcal/mol in panel C, which is 12 kcal/mol lower than the same quantity in panel A, showing that the UNRES 4P potential energy is lower in the native basin.

Figure 23 shows potential energy (panel A) and ρ values (panel B) for a sample trajectory obtained with BD. As can be seen in this example, all Berendsen trajectories showed higher-energy values (panel A) and higher fluctuations in the ρ values (panel B) compared to LD runs (panels C and D in Figure 22). This could be explained by the fact that, for BD, the absence of friction forces allows for larger conformational changes. No simulations were carried out for 1C94 dimers.

3.3.2. Tetramers. Berendsen and Langevin simulations were carried out starting with the four chains in the extended conformation, with each pair of chains cross-linked by disulfide bonds. The chains were in the same plane, parallel to each other and with a 20 Å distance between consecutive chains. On the basis of the experimental data,[36] the Cys residue at the first N-terminal position was assumed to form a disulfide bond with the corresponding Cys residue in another chain; however, this residue was never included in the rmsd calculations since it is not resolved in the experimental structure. The simulation time was 35 ns for LD runs and 28 ns for BD runs. The equilibrium concentration of 10 mM was reached during the first 50 ps of simulation.

None of the trajectories obtained with UNRES/MD yielded nativelike structures. On the other hand, both methods found stable structures consisting of two parallel dimers bound together in an antiparallel orientation (instead of parallel as in the native structure), examples of which are shown in Figures 20B and 20C. In the structure shown in Figure 20B, the dimers have nativelike structures, but the area of contact between the dimers is very small. However, the structure shown in Figure 20C has better packing, but the dimers have non-native-like structures, and the disulfide-linked monomers are not parallel to each other but slightly twisted to align in an antiparallel orientation with the monomers from the other dimer. These two structures have approximately the same potential energy (approximately −507 kcal/mol); we will refer to either of them as misfolded structures. Figure 24 shows the potential energy (panel A) and ρ values for the tetramer (panel B) and for the dimers (panels C and D) as a function of time for the trajectory leading to the structure in Figure 20B. It can be seen that, by the end of the simulation, the ρ values for the tetramer stabilize around 22 Å (indicated by a solid line in panel B) while, for the dimers, it remains below or close to the 5.6 Å cutoff (indicated by the dashed lines in panels C and D). The potential energy also stabilizes by the end of the simulation, with values around −510 kcal/mol (indicated by a solid line in panel A).

To determine whether the native structure of the tetramer could not be found because of imperfections in the UNRES 4P force field or simply because the simulation times were too short, we carried out a set of 8 ns simulations with the crystal structure as the initial conformation using Langevin dynamics. As can be seen in Table 3, 3 out of 10 simulations remained in the native basin. Potential energy and ρ values corresponding to one of the trajectories that did not remain in the native basin are shown in Figure 25. It is important to notice that although the tetramer leaves the native basin (ρ values crossing the dashed line at the 8 Å cutoff in panel B) there is no substantial change in the potential energy (panel A). We calculated the average

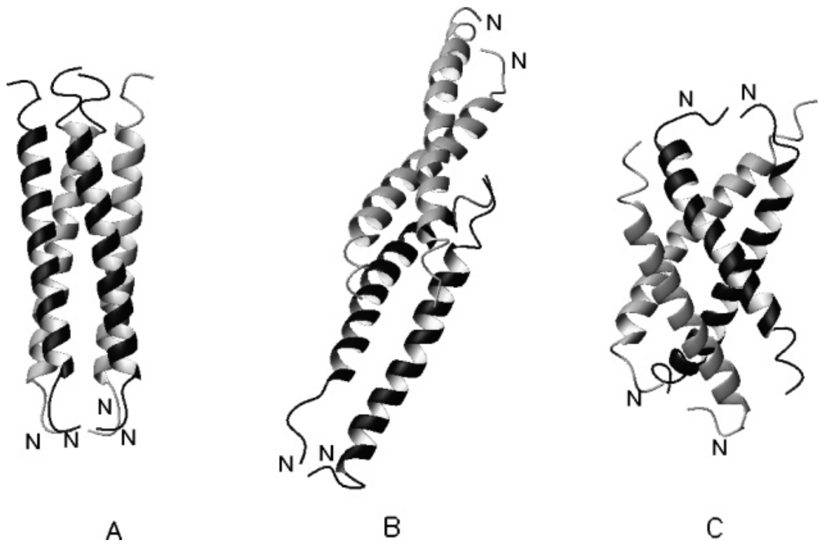

Figure 20. (A) Experimental structure of 1C94 and (B and C) examples of misfolded structures obtained with LD and BD UNRES/MD. The N-terminus of each chain is indicated.

TABLE 3: Summary of Trajectories for 1C94

	tetramer											
	from extended conformation				from crystal structure			monomer				
algorithm	N_f[a]	$\langle\tau_f\rangle$[b] (ns)	N_{mf}[c]	$\langle E\rangle_{mf}$[d] (kcal/mol)	N_n[e]	$\langle E\rangle_n$[f] (kcal/mol)	CPU time[g] (hs)	N_f[a]	$\langle\tau_f\rangle$[b] (ns)	ρ_{min}[h] (Å)	$\langle\tau_{res}\rangle$	$\langle E\rangle_f$[i] (kcal/mol)
Berendsen	0(18)	2.2	4	−504			6.9	10	1.4	1.36	81%	−107
Langevin	0(18)	2.6	3	−508	3	508	8.1	9	2.0	1.28	83%	−152

[a] Number of trajectories (out of 10) that folded to nativelike structures, starting from the extended conformation. In the multichain simulations, the number of monomers (out of 40, since there were 4 monomers on each of the 10 simulations of tetramers) that folded to a nativelike structure is indicated between parentheses. [b] Average folding time of the monomers. The folding time was defined as the time at which the rmsd with respect to the crystal structure fell below 4 Å. In those runs for which the rmsd never went below the cutoff, the folding time was considered to be the simulation time (12 ns for the isolated monomer simulations, 35 ns for the tetramers simulations with LD, and 26 ns for the tetramer simulations with BD). The average folding times for the multichain complex are not calculated since none of the simulations led to nativelike tetramers. [c] Number of trajectories (out of 10) that yielded misfolded structures. [d] Average potential energy over all the structures in the misfolded basin. [e] Number of trajectories, out of 10 simulations started with the crystal structure as the initial conformation that, after 8 ns of simulation, still had nativelike structures (rmsd with respect to crystal structure below 8 Å). [f] Average potential energy over all those trajectories that, starting with the crystal structure, remained in the native basin after 8 ns of simulation. [g] Average CPU time (in hours) per 1 ns of simulation on a single 3.06 GHz Intel Pentium IV Xeon processor. [h] The lowest rmsd in all of the fluctuating trajectories. [i] Average potential energy over all structures in the native f basin.

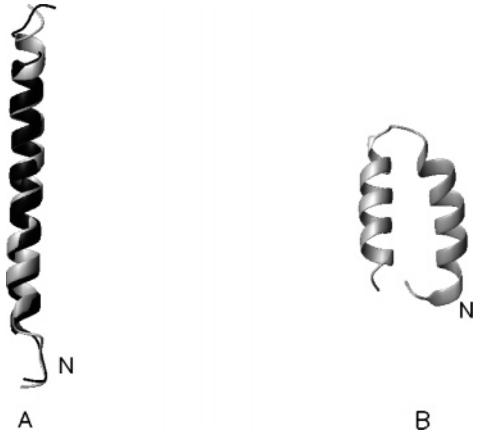

Figure 21. (A) Superposition of one of the monomers in the experimental structure of 1C94 (black) on the most nativelike structure (gray) (C$^\alpha$ rmsd = 1.28 Å) obtained with the BD UNRES/MD. (B) A structure that was often found during the folding pathway of 1C94 (either with BD or LD) and was the final structure of the monomer LD trajectory that did not find the native basin. The N-terminus is indicated.

potential energy among those structures that remained in the native basin and compared it with the average energy among the misfolded structures. The values obtained were almost equal (Table 3), indicating that the protein might choose either conformation with the same probability. However, when starting from the extended conformation, none of the simulations led to nativelike structures. Therefore, the energy landscape generated by the UNRES 4P potential makes the antiparallel conformation more easily accessible than the parallel (native) conformation; i.e., the free energy of the misfolded basin has a lower value compared to that of the native basin.

When comparing the folding times of the monomers in the single- and multichain simulations (Table 3), we did not find any appreciable difference, indicating that, for this protein, multichain interactions do not play an important role in the folding of the monomers.

We conclude that the failure to fold the protein to the native tetramer with the UNRES 4P force field should be attributed to the imperfections in the potential rather than to insufficient simulation time because, first, for the two preceding proteins (1G6U and GCN4-p1), we observed the formation of both the native and the non-native dimers and, second, in our previous implementation of UNRES to search for the native structures of multichain proteins with CSA,[25,40] the native structure of retro-GNC4 could be predicted by global optimization only when native symmetry constraints were imposed. Improvement

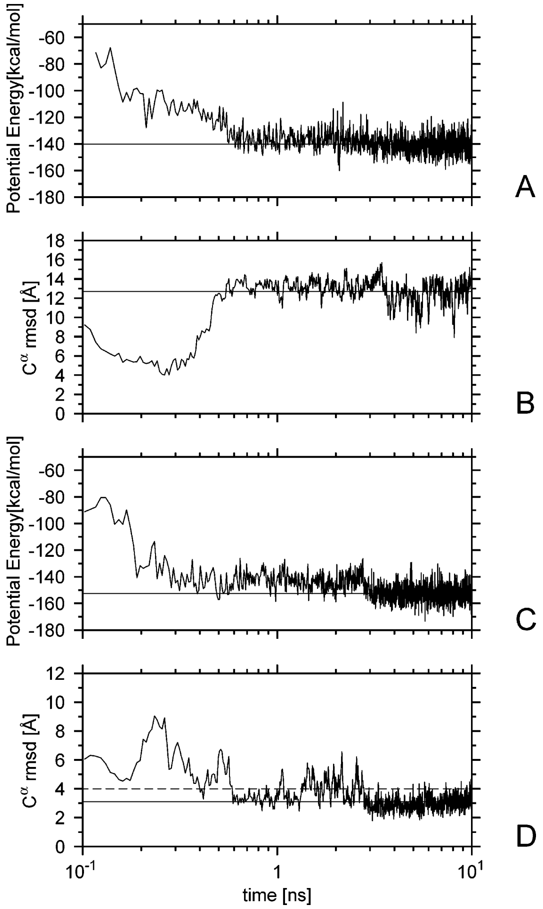

Figure 22. (A) Variation of the potential energy and (B) the C$^\alpha$ rmsd from the native structure of the monomer in the tetramer as a function of time for an LD trajectory of an isolated monomer of 1C94 converging to a non-native-like structure (which is shown in Figure 21B). In panel A, the solid horizontal line at −140 kcal/mol is the mean value of the energy after the monomer has adopted the non-native stable structure. In panel B, the solid horizontal line at 12.8 Å is the mean C$^\alpha$ rmsd after the peptide has adopted the non-native structure. Panels C and D contain the same information as panels A and B, respectively, for an LD trajectory converging to the native basin. The solid horizontal line at −152 kcal/mol in panel C is the mean value of the energy after the peptide has reached the native basin. In panel D, the dashed horizontal line at 4 Å corresponds to the cutoff rmsd above which the structure is considered to have left the native basin, and the solid horizontal line at 3.1 Å is the mean C$^\alpha$ rmsd inside the native basin. The solid horizontal line at −152 kcal/mol in panel C is the mean value of the energy after the peptide has reached the native basin.

of the 4P UNRES force field is expected to stabilize the native basin to a greater extent compared to the non-native basin.

4. Conclusions

The UNRES/MD implementation described in ref 13 was extended to treat multichain proteins. The method was tested on three α-helical proteins, two dimers and one tetramer.

To simulate a constant temperature bath, two alternative methods were implemented, the Berendsen thermostat (BD) and a method based on the Langevin equation (LD). The latter method includes friction and stochastic forces explicitly as opposed to the former for which these forces are included implicitly. When comparing the time required for each method to find the global minimum of the energy, BD proved to be much faster than LD, as observed in our earlier studies on single-chain proteins.[12] However, it should be noted that, despite its predicting efficiency, BD might not reproduce the true folding pathway. LD, which reproduces a true canonical ensemble, should be used instead when studying the kinetics of the folding process, as in ref 14.

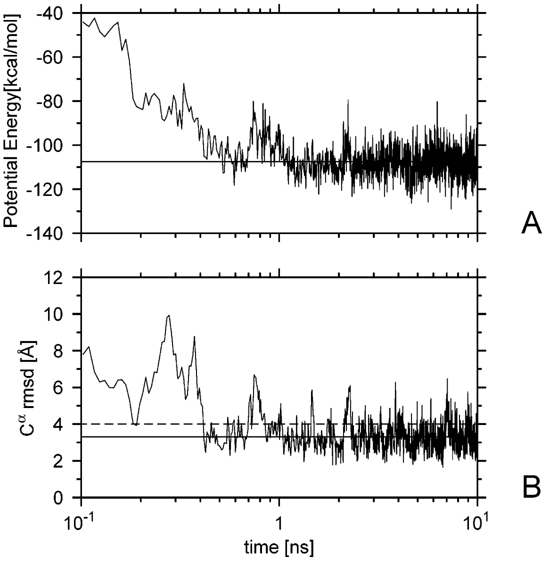

Figure 23. (A) Variation of the potential energy and (B) the C$^\alpha$ rmsd from the native structure during the folding of an isolated monomer of 1C94 obtained with Berendsen dynamics. In panel A, the solid horizontal line at −105 kcal/mol is the mean value of the energy after the protein has reached the native basin. The dashed horizontal line, at 4 Å, in panel B corresponds to the cutoff rmsd above which the structure is considered to have left the native basin, and the solid horizontal line at 3.3 Å in the same panel is the mean C$^\alpha$ rmsd inside the native basin.

Simulations of single chains and multichain complexes were carried out with BD and LD. Single-chain simulations indicate that, for each of the three α-helical proteins tested in this work, the structure adopted by the monomer in the multichain complex is also the lowest UNRES 4P energy structure of the isolated monomer. In general, the folding times of the monomers in the single-chain simulations were shorter than those in the multichain simulations, which indicates that, with the UNRES 4P force field, the short-range interactions, responsible for the folding of the single-chain α-helices, are impaired by the interactions between different chains. However, the folding of 1G6U with BD (section 3.1) was the exception. In these simulations, the monomers folded faster when they were allowed to interact with another monomer; i.e., the correct packing of the two helices on each monomer is favored by the interactions with another monomer. Although the wrong orientation of the monomers with respect to each other can sometimes hinder the packing of the helices, with BD, in which the friction forces are absent, the chains can rearrange quickly to find a more favorable orientation that will aid the packing of each monomer. This behavior is probably an artifact of BD and might not represent the folding mechanism of 1G6U.

It is important to note that, although some of the trajectories led to non-native-like structures, these structures were indeed free-energy minima within the context of UNRES 4P. In the case of the two dimers, the non-native structure was competing with the native one. This competition was reflected in our simulations, especially in the case of GCN4-p1 for which the dimer switched from one structure to the other. In the case of 1C94, the results were poor since none of the trajectories yielded the native structure. The reason for this failure might be found in the defects of the UNRES parameters. Improvement of these parameters is ongoing research in our laboratory.

It must be emphasized that the goal of this work was to test the implementation of UNRES/MD on multichain proteins and

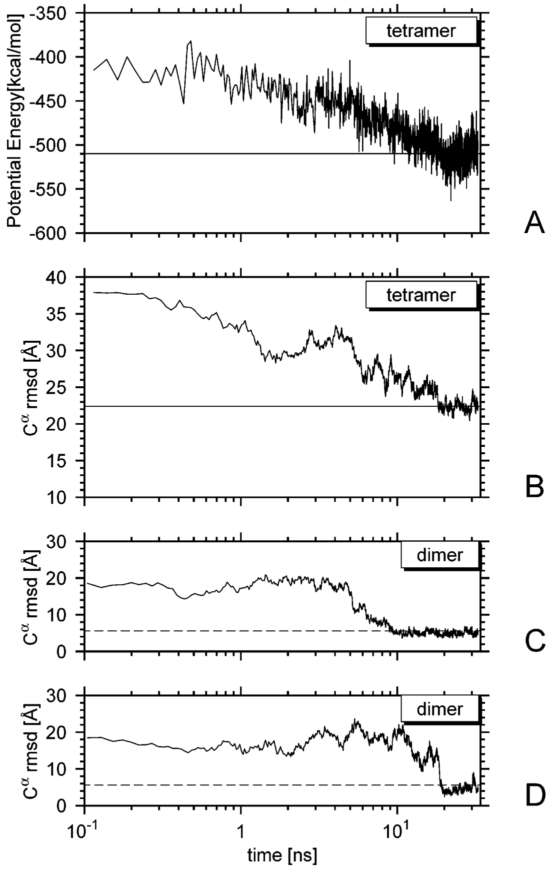

Figure 24. (A) Variation of the potential energy and (B) the C^α rmsd from the native structure for a misfolding trajectory of 1C94, starting with extended chains, obtained with Langevin dynamics. For the same trajectory, panels C and D show the variation of the C^α rmsd from the native for each of the dimers. In panels A and B, the solid horizontal line is the mean value of the energy (at −510 kcal/mol) and C^α rmsd from the native structure (at 22.4 Å), respectively, after the tetramer has found the misfolded basin. The dashed horizontal line in panels C and D corresponds to the 5.6 Å cutoff rmsd, above which the dimers are considered to have left the native basin; i.e., the dimers folded but the overall structure was misfolded.

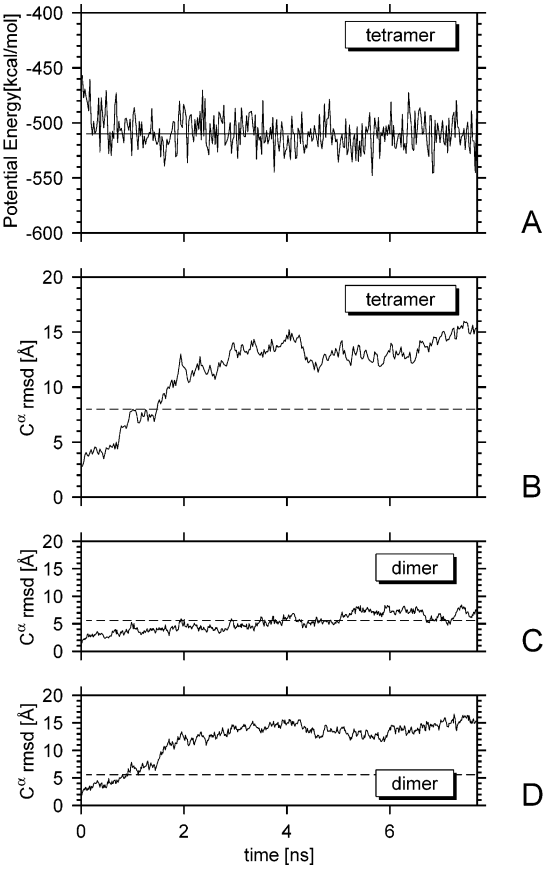

Figure 25. (A) Variation of the potential energy and (B) the C^α rmsd from the native structure for a trajectory of 1C94 that did not remain in the native basin, obtained with Langevin dynamics, with the crystal structure as the initial conformation. The solid horizontal line at −510 kcal/mol in panel A is the mean value of the energy during the simulation. The dashed horizontal line in panel B corresponds to the 8 Å cutoff rmsd, above which the tetramer is considered to have left the native basin. For the same trajectory, panels C and D show the variation of the C^α rmsd from the native for each of the dimers. The dashed horizontal line in panels C and D corresponds to the 5.6 Å cutoff rmsd, above which the dimers are considered to have left the native basin.

not to improve the 4P force field, and therefore, we chose relatively simple systems which the force field could treat to test the approach, as pointed out in the Introduction. The UNRES 4P force field was trained using four proteins with different topologies and tested on 66 proteins with chain lengths from 28 to 144 amino acid residues. The average size of correctly predicted segments of α-helical proteins was approximately 67 residues.[23] The parametrization procedure and the limitations of the UNRES 4P force field are described extensively in ref 23. The reason for such limitation must be found in the old parametrization procedure,[19,21−23] which neglected conformational entropy, an issue that has been addressed in the new procedure that is currently being developed in our laboratory, and preliminary results are reported in a separate paper.[41] The multichain UNRES/MD method, with a force field developed by this new procedure, was recently used to predict the structure of a homodimer in the CASP7 experiment. The predicted monomeric structures were complemented with MD simulations of dimers, considerably improving the quality of the predictions. Our CASP7 results will be reported in a separate paper.

Finally, in contrast to earlier calculations of multichain complexes,[25,40] with CSA[42,43] as a global optimization algorithm, in which symmetry constraints had to be imposed to simulate the experimental structure, no such constraints were imposed here. Apparently, in the time scale achieved in MD with UNRES, the search of the conformational space of a dimer is more efficient than that with CSA.

Acknowledgment. We thank Dr. M. Khalili for comments on this paper. This research was supported by grants from the National Institutes of Health (NIH; GM-14312), from the NIH Fogarty International Center (TW-7193), from the National Science Foundation (MCB05-41633), and from the Polish Ministry of Education and Science (3 T09A 032 6). It was carried out by using the resources of our 880-processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University, the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, the National Center for Supercomputing Applications System at the University of Illinois at Urbana−Champaign, and the resources of the Center for Computation and Technology at Louisiana State University, which is supported by funding from the Louisiana legislature's Information Technology Initiative.

References and Notes

(1) Day, R.; Daggett, V. *Adv. Protein Chem.* **2003**, *66*, 373.
(2) Fersht, A. R.; Daggett, V. *Cell* **2002**, *108*, 573.

(3) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76.

(4) Shea, J. E.; Brooks, C. L., III. *Annu. Rev. Phys. Chem.* **2001**, *52*, 499.

(5) Vieth, M.; Kolinski, A.; Brooks, C. L.; Skolnick, J. *J. Mol. Biol.* **1994**, *237*, 361.

(6) Ma, B.; Nussinov, R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14126.

(7) Levy, Y.; Caflisch, A.; Onuchic, J. N.; Wolynes, P. G. *J. Mol. Biol.* **2004**, *340*, 67.

(8) Yang, S.; Cho, S. S.; Levy, Y.; Cheung, M. S.; Levine, H.; Wolynes, P. G.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 13786.

(9) Levy, Y.; Papoian, G. A.; Onuchic, J. N.; Wolynes, P. G. *Isr. J. Chem.* **2004**, *44*, 281.

(10) Yang, S.; Levine, H.; Onuchic, J. N.; Cox, D. L. *FASEB J.* **2005**, *19*, 1778.

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

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

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

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

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

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

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

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

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

(20) Ołdziej, S.; Kozlowska, U.; Liwo, A.; Scheraga. H. A. *J. Phys. Chem. A* **2003**, *107*, 8035.

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

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

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

(24) 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.; Kazmierkiewicz, 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.

(25) Saunders, J. A.; Scheraga, H. A. *Biopolymers* **2003**, *68*, 300.

(26) Czaplewski, C.; Ołdziej, S.; Liwo, A.; Scheraga, H. A. *Protein Eng., Des. Sel.* **2004**, *17*, 29.

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

(28) de Gennes, P.-G. *Scaling Concepts in Polymer Physics*; Cornell University Press: Ithaca, NY, 1979; Chapter VI.

(29) Veitshans, T.; Klimov, D.; Thirumalai, D. *Folding Des.* **1996**, *2*, 1.

(30) Cieplak, M.; Hoang, T. X.; Robbins, M. O. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 104.

(31) Nosé, S. *Mol. Phys.* **1984**, *52*, 255.

(32) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.

(33) Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384.

(34) Liwo, A.; Arłukowicz, P.; Ołdziej, S.; Czaplewski, C.; Makowski, M.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 16918.

(35) O'Shea, E. K.; Klemm, J. D.; Kim, P. S.; Alber, T. *Science* **1991**, *254*, 539.

(36) Mittl, P. R. E.; Deillon, C.; Sargent, D.; Liu, N.; Klauser, S.; Thomas, R. M.; Gutte, B.; Grütter, M. G. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2562.

(37) Ogihara, N. L.; Ghirlanda, G.; Bryson, J. W.; Gingery, M.; DeGrado, W. F.; Eisenberg, D. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 1404.

(38) Harbury, P. B.; Zhang, T.; Kim, P. S.; Alber, T. *Science* **1993**, *262*, 1401.

(39) Lee, J.; Ripoll, D. R.; Czaplewski, C.; Pillardy, J.; Wedemeyer, W. J.; Scheraga, H. A. *J. Phys. Chem. B* **2001**, *105*, 7291.

(40) Saunders, J. A.; Scheraga, H. A. *Biopolymers* **2003**, *68*, 318.

(41) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Ołdziej, S.; Wachucik, K.; Scheraga, H. A. *J. Phys. Chem. B* **2006**, in press.

(42) Lee, J.; Scheraga, H. A.; Rackovsky, S. *J. Comput. Chem.* **1997**, *18*, 1222.

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