THE JOURNAL OF
PHYSICAL CHEMISTRY B

Article

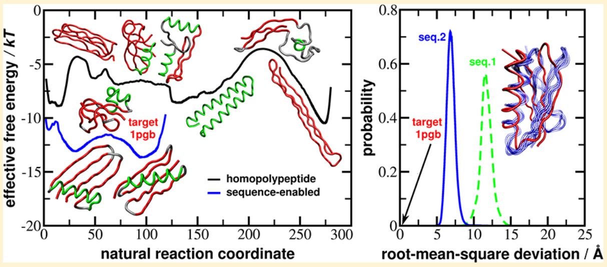
pubs.acs.org/JPCB

Tracking Polypeptide Folds on the Free Energy Surface: Effects of the Chain Length and Sequence

Andrey V. Brukhno,* Piero Ricchiuto, and Stefan Auer

Centre for Molecular Nanoscience, School of Chemistry, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, United Kingdom

Supporting Information



ABSTRACT: Characterization of the folding transition in polypeptides and assessing the thermodynamic stability of their structured folds are of primary importance for approaching the problem of protein folding. We use molecular dynamics simulations for a coarse grained polypeptide model in order to (1) obtain the equilibrium conformation diagram of homopolypeptides in a broad range of the chain lengths, $N = 10, ..., 100$, and temperatures, T (in a multicanonical ensemble), and (2) determine free energy profiles (FEPs) projected onto an optimal, so-called "natural", reaction coordinate that preserves the height of barriers and the diffusion coefficients on the underlying free energy hyper-surface. We then address the following fundamental questions. (i) How well does a kinetically determined free energy landscape of a single chain represent the polypeptide equilibrium (ensemble) behavior? In particular, under which conditions might the correspondence be lost, and what are the possible implications for the folding processes? (ii) How does the free energy landscape depend on the chain length (homopolypeptides) and the monomer interaction sequence (heteropolypeptides)? Our data reveal that at low T values equilibrium structures adopted by relatively short homopolypeptides ($N < 60$) are dominated by α-helical folds which correspond to the primary and secondary minima of the FEP. In contrast, longer homopolypeptides ($N > 70$), upon quasi-equilibrium cooling, fold preferentially in β-bundles with small helical portions, while the FEPs exhibit no distinct global minima. Moreover, subject to the choice of the initial configuration, at sufficiently low T, essentially metastable structures can be found and prevail far from the true thermodynamic equilibrium. We also show that, by sequence-enabling the polypeptide model, it is possible to restrict the chain to a very specific part of the configuration space, which results in substantial simplification and smoothing of the free energy landscape as compared to the case of the corresponding homopolypeptide.

■ INTRODUCTION

Folding of a protein into its functional native structure is often described by using the concept of a free energy "landscape".[1,2] Although this concept is very useful for analysis of any thermodynamic system undergoing a phase or structural transition, the free energy landscapes are generally inaccessible experimentally and, thus, can be only evaluated with the use of purely computational methods. Moreover, the true free energy landscape of a protein is essentially a high-dimensional hypersurface determined in the space of its vast configurational degrees of freedom. Thus, accurate quantitative characterization of the transition states and the associated free energy barriers is rather challenging, as it implies unambiguous determination of the (relative) free energy levels of both native and all non-native protein structures. In numerical simulation, being nowadays the primary tool for assessment of free energy differences, three major problems arise: (i) the common ergodicity issue concerned with efficient sampling of numerous conformations of a protein (polymer, in general) separated by large free energy barriers, (ii) the closely related problem of the restricted accessible time scale in molecular dynamics (MD) simulations due to the still severe computational hardware limitations,[3] and (iii) the theoretical, methodological issue of adequate representation of the underlying free energy hypersurface by projecting it on a reduced parameter set, most often with only one or two, maximally three, order parameters (as it is troublesome to visualize and envisage more).[4,5]

On the one hand, the most tractable protein models for which evaluating free energies is relatively straightforward are confined to lattice[2,6] and Gō models.[7,8] Although these models are instrumental and valuable insights have been gained in these studies, their applicability is limited by either the very restricted operational configuration space or the interaction potentials biased toward a prechosen native protein structure. On the other hand, using full-atom models has allowed for examining free energy landscapes for small peptides,[4,9,10] and recent advances in hardware and simulation methodology enabled evaluation of free energy barriers for small fast-folding proteins.[11−13]

Special Issue: Macromolecular Systems Understood through Multiscale and Enhanced Sampling Techniques

Received: January 31, 2012
Revised: May 23, 2012
Published: May 24, 2012

The MD simulations listed above were performed at constant temperature, which incurs the risk of undersampling the configuration space and missing some of the metastable configurations. Therefore, the relation of so-determined "kinetic" free energy landscapes to the true equilibrium free energy surface is not clear. In addition, the results in the given examples are mostly specific to particular peptides and proteins, whereas less is known about the general effects on the underlying free energies due to, e.g., the chain length and the amino acid sequence. In this regard, two fundamental questions arise. (i) Does the "kinetic" free energy landscape accurately represent the equilibrium free energy surface, and if it is not always the case, under which conditions does the description break down? (ii) How does the shape of the free energy landscape depend on the chain length and the amino acid sequence?

In this report, we address these questions with the aid of discontinuous molecular dynamics (DMD) simulations for a coarse grained off-lattice model of polypeptide that was originally developed by Hoang and co-workers.[14−16] The distinct feature of the model we use (see Methodology) is that the chain backbone is represented by partly overlapping hard spheres centered on C_α atoms of a polypeptide. The effective stepwise interactions are parametrized so as to include a well-tuned directional hydrogen bonding scheme (with well depth ε), pairwise attractive hydrophobic forces (well depth ε_{hp}), and local bending stiffness (energy penalty ε_s). The model is, therefore, regarded as a semiflexible "tube".

Despite its rather simplified parametrization, only spanning the solvent-mediated interactions and the chain bending regidity, i.e., the set (ε, ε_{hp}, ε_s) with the corresponding distances, the homopolypeptide tube model is known to possess a rich conformation spectrum.[14] In particular, the *ground state* diagram of the tube model in the (ε_{hp}, ε_s) plane embraces most of the structural elements found in native proteins, either stable or metastable. On the basis of this observation, the tube-like topology has been considered the essential feature of all polypeptides that "pre-sculpts" the native secondary and tertiary protein structure by constraining the polypeptide conformations to a rather specific "menu" of possible folds.[14] Furthermore, within the simplest model allowing for heterogeneous interations within a polypeptide chain, with monomers (residues) of only two kinds, strongly hydrophobic (H) and "polar" (P), it has also been shown[16] that prechosen conformation folds can be *energetically* stabilized by selecting specific sequences of H and P monomers. We note though that only relatively short polypeptide chains (up to 50 residues) have been considered in those seminal studies and the amino acid sequence was specified by using only two levels of the hydrophobicity, while the hydrogen bonding and the backbone stiffness interactions were homogeneous along the chain.

In a recent communication,[17] we reported the equilibrium conformation state diagram of a single homopolypeptide chain within the 2D space of the chain length, N, and temperature, T, i.e., in the (N, T) plane. The state diagram was obtained in multicanonical molecular dynamics simulations in a broad range of N and T, which ensured the true equilibrium statistics. The diagram revealed a chain length induced α-helix to β-sheet transition with increase of the degree of polymerization, N. In the current study, we extend our simulations to determine one-dimensional (projected) free energy profiles (FEPs) with respect to a "natural" (optimal) reaction coordinate that preserves the free energy barriers and the reaction coordinate dependent diffusion coefficients on the free energy hypersurface.[4] In what follows, the Methodology section describes the model and simulation details. In the Results and Discussion section, we first shortly present the previously obtained equilibrium conformation state diagram for homopolypeptides. Then, we discuss the corresponding FEPs, which elucidate the effects of the homopolypeptide chain length and conformation metastability at low T values. Finally, we present the FEPs for heteropolypeptides while aiming at secondary and tertiary structures of two target proteins, where the amino acid sequence is effectively modeled by a three-letter code for heterogeneous interactions within the chain.

■ METHODOLOGY

Model. The model is based on stepwise (discrete) pairwise potentials of square-well type, which are graphically represented in Figure 1 and have been described in detail elsewhere.[18,19] Below, we recapitulate the major features of the model.

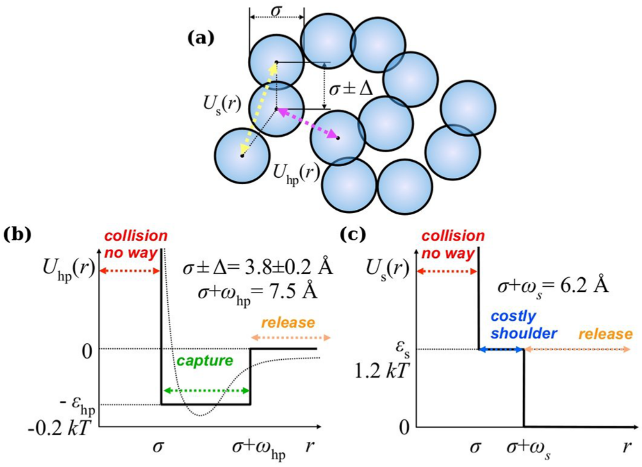


Figure 1. (a) A sketch of the polypeptide model, with the hydrophobic and steric interactions schematically indicated by arrows. (b and c) Graphical representation of the square-well model potentials used in DMD, where r is the separation between two C_α sites, σ is the monomer hard-core diameter, 2Δ is the bond length allowing an overlap of Δ between adjacent monomers, ω is the effective potential width, and ε is the corresponding interaction strength.

A polypeptide is modeled as a nonbranched semiflexible "tube" where only the positions of C_α-atoms in the amino acid sequence are accounted for. In the original description by Hoang et al.,[14] the finite thickness of the polypeptide was imposed by requiring the radius of the circle drawn through any three C_α atoms to be larger than 2.5 Å. In contrast, here we gain computational efficiency by representing amino acid residues as large spheres centered on their C_α-atoms and capable of partially overlapping with the neighbors; see Figure 1. The so coarse grained monomeric units have the diameter σ = 3.8 Å and are connected into a chain with bonds of length constrained to 3.8 ± 0.2 Å, which allow for slight overlaps between bonded monomers. Overlapping is, however, prohibited for nonadjacent monomers, thereby accounting for excluded volume (steric) constraints.

In order to mimic local constraints imposed on the chain conformation by the presence of side chains (absent within the model), the bond angle between three consecutive C_α atoms is restricted between 82 and 148°, and an energetic penalty, $\varepsilon_s > 0$, is applied for bond angles smaller than 107.15°. Overall, the peptide thickness and bending regidity are tuned so as to mimic clash-free packing of side groups.[19] The anglular penalty and restrictions are conveniently turned into the corresponding criteria for distances, $r_{i,i+2}$. Likewise, the hydrophobic effects are captured by use of a pairwise square-well potential with depth $-\varepsilon_{\mathrm{hp}} < 0$, acting between monomers with indexes i, $j > i + 1$ when their separation becomes smaller than $\sigma + w_{\mathrm{hp}} = 7.5$ Å.

As was mentioned above, a specific trait of this model is the unique geometric rules for hydrogen bonding, designed to reproduce on average the probability distributions for distances between H-bonded C_α-atoms in native protein structures.[19] Three major HB types are recognized: "α-helix" and parallel and antiparallel "β-sheets". The terminal monomers are incapable of hydrogen bonding, whereas any inner monomer can form two hydrogen bonds at most (one as donor and one as acceptor). Although all the nonterminal monomers are initially equivalent in terms of the interaction potency, which implies the initial isotropy of the interactions within the chain, different H-bond types can emerge depending on the momentary local environment around a particular pair of monomers approaching each other. Since the cooperativity in positions of three or four C_α sites is often required for establishing an individual H-bond,[20] eventually monomers start differentiating in their H-bonding propensities, which follow, of course, the arising structural patterns. Thus, the evolution of H-bonds within the chain proceeds on the "first-come, first-served" basis. The H-bonding potential is also of a square-well type, with its depth $-\varepsilon < 0$ at full strength. We note that a single nonlocal H-bond between residues i and $j > i + 3$ is somewhat weaker, -0.7ε, than H-bonds formed locally ($j = i + 3$). When two nonlocal H-bonds are formed cooperatively between consecutive monomer pairs (i, $i + 1$) and (j, $j + 1$), an additional energy of -0.3ε per H-bond is gained, thereby bringing the cooperative H-bonds to the full strength, $-\varepsilon$. The corresponding set of distance criteria for α-helix and β-sheet H-bonds are summarized in ref 19. As previously,[17−19,21] hydrophobic contacts are 20 times weaker than H-bonds, $-\varepsilon_{\mathrm{hp}} = -0.05\varepsilon$, whereas the energetic penalty for bending, $\varepsilon_s = 0.3\varepsilon$, is assigned to each monomer in the vertex of a local bond angle smaller than 107.15°.

Throughout the text, dimensionless energetic units are used so that only relative energy values are physically meaningful. The reduced temperature $T' = k_{\mathrm{B}}T$ (K)$/\varepsilon$, where $\varepsilon = (1.6, ..., 2.4) \times 10^{-20}$ (J), is the energy stored in a polypeptide hydrogen bond corresponding to the experimentally measured folding temperatures, T_f, between 276 and 363 K.[21] That is, the conversion to the absolute temperature scale can be done by selecting the appropriate value of ε so as to match the transition (folding) temperature, T_f (K) $= T'\varepsilon/k_{\mathrm{B}}$, with the experimental value for a particular polypeptide.

Simulation Aspects. We use the discontinuous molecular dynamics (DMD) method developed specifically for models with stepwise energetic landscapes.[19,22,23] DMD propagates the system on a collision-by-collision basis and, hence, does not require force (re)calculation or tracking of intermediate configurations between subsequent events of abrupt changes on the potential hypersurface. The velocities are then only updated during these momentary "collisions". These include actual absolutely elastic collisions due to infinite potential walls (geometric constraints) or insufficient kinetic energy to overcome a finite potential barrier, and also potential−kinetic energy conversions while traversing the stepwise changes in the model potentials. The Andersen thermostat is implemented for maintaining constant temperature.[24]

Two types of simulations have been carried out: (i) in a multicanonical ensemble, in order to obtain converged equilibrium statistics over a range of temperatures simulated in one run, and (ii) extensive NVT runs at selected T' values, in order to produce sufficiently long trajectories for collecting reliable statistics on rough free energy landscapes at low temperatures, which may not necessarily correspond to the equilibrium but rather reflect the dynamics of polypeptide folding within local traps in the conformation space (depending on the initial configuration).

Multicanonical Simulations. In order to facilitate (quasi-)equilibrium configuration sampling and allow the polypeptide to avoid being trapped in metastable conformations at low temperatures, we employed the multicanonical expanded ensemble (MEE) method.[25] That is, the configuration trajectory was generated by the DMD means, whereas additionally the temperature was periodically varied with the aid of Monte Carlo (MC) moves, as is briefly described next. Simulations were carried out in two sets of NVT (sub)-ensembles within two temperature ranges, $T' \in [0.2, 0.3]$ or $[0.3, 0.5]$. In each case, a discrete grid, $\{T'_m; m = 0, ..., M; M = 20\}$, was chosen such that $T'_0 < T'_M$ and the spacing between T' values followed the inverse proportionality, $1/T'_m = 1/T'_0 + m\Delta'_T$, with step $\Delta'_T = (1/T'_M - 1/T'_0)/M$, which implied tightening of the T' grid for lower temperatures. This choice ensured reasonably high acceptance rate of T' variations and also eased possible ergodicity issues associated with finite relaxation times upon successful jumps in T'. Each simulation started with a randomized configuration in an NVT subensemble corresponding to T'_M, the highest in a T'-set, which guaranteed the shortest possible initial relaxation period. The standard MEE algorithm, based on hopping between adjacent NVT substates (in addition to normal configuration sampling), was reinforced by using a self-consistent "on-the-fly" iteration for free energy accumulation similar in spirit to the density of states (DOS) algorithm of Wang and Landau.[26] Thus, whenever a jump in temperature was being attempted, $m \to m \pm 1$, the penalty (or balancing) function, $\eta(T')$, in the resulting substate was incremented by a certain "drop-in" amount, initially 0.05 in reduced energy units, and decremented upon sufficient flattening of the running histogram of substate visits. Such an MC move was accepted with probability $p(m \to m \pm 1) = \min\{1, \exp[\mp\Delta'_T E_m + \Delta\eta_{m,m\pm1}]\}$, where $\Delta\eta_{m,m\pm1}$ accounted only for the currently accumulated increments in the η values. This way, the balancing of visits to all NVT substates within the MEE was made efficient and robust. Note that for our purposes we do not need a high accuracy in the (Helmhotz) free energy estimates, $\Delta A_{m,m\pm1}/(k_{\mathrm{B}}T) \approx \Delta\eta_{m,m\pm1}$, but rather we seek uniformity of sampling and statistics collected over the entire temperature range being simulated.

Since the time between subsequent collisions varies broadly, depending on the compactness of the chain conformation, it is crucial to ensure that T-sampling is performed uniformly in *time*, as opposed to attempting T-moves after a certain number of collisions. Therefore, during a simulation, the time passed since the last T-step was monitored and compared to the chosen time lag between MC moves within the MEE. When such the accumulated time up to the next collision exceeded the time lag, the system was only allowed to propagate to the point at which the next T-attempt had to be made, rather than to the next collision. In the case of a rejected T-attempt, i.e., when the temperature remained constant, the system was propagated further and the next collision was dealt with in a normal manner. Otherwise, upon an accepted T-jump, the Andersen thermostat was applied to the entire system and the search for the next collision had to be redone.

A simulation was arranged in blocks, comprising 10^4 collisions each, with at least 2×10^6 blocks per single MEE-DMD run. The MC time lag was chosen such that, on average, one T-step was attempted every 10 blocks, i.e., a trajectory of $\sim 10^5$ collisions was generated at constant temperature. Hence, upon completion of 10^6 blocks, the total number of T-attempts reached $\sim 10^5$, and assuming uniform sampling over the MEE, the average number of collision events in each subensemble was proportional to $10^{10}/(M + 1)$. For longer chains ($N > 50$), we doubled the simulation length in terms of the total block number.

Free Energy Calculation along Folding/Refolding Pathways. Quantitative analysis of polypeptide/protein dynamics in terms of the free energy landscapes is notoriously difficult. A poorly chosen reaction coordinate may hide the complexity of the free energy landscape and associated dynamics.[27−29] The optimal reaction coordinate and associated free energy landscape is constructed as follows. First, the trajectory is used to build a network, the equilibrium kinetic network (EKN), which describes the system kinetics at equilibrium. This is obtained by clustering the trajectory in the principal component space defined by the distance between selected atom pairs, and counting the number of transitions between clusters.[7] Once such a network has been determined, its free energy profile (FEP) is built along the second eigenvector reaction coordinate.[30,31] The FEP is plotted as a function of a "natural coordinate" which is constructed so that the diffusion coefficient is constant along the profile.[32] It is assumed that the constructed EKN accurately describes the dynamics of the system. More rigorous methods to determine the optimal reaction coordinate without constructing the EKN have been suggested;[29,30,33] however, they are more computationally expensive.

In the current study, in order to obtain free energy profiles while tracking the propagation of the polypeptide chain along the optimal, so-called "natural", reaction-coordinate, we perform an NVT simulation during which the chain configurations from the generated DMD trajectory are regularly stored. Upon completion of the simulation, we use Krivov's analysis tools so as to obtain the FEP. All such simulations consisted of at least 2×10^6 DMD blocks, providing comprehensive statistics in most cases.

■ RESULTS AND DISCUSSION

In the following, we aim to explore (1) how the *equilibrium* folding behavior of homopolypeptides relates to *kinetically determined* free energy landscapes at low temperatures and (2) if and how the polypeptide conformation free energy profiles are affected by the choice of (a) its initial configuration (homopolypetides) and (b) its monomer interaction sequence (heteropolypetides).

Homopolypeptides: Equilibrium Conformation State Diagram. By using multicanonical molecular dynamics simulations, we obtained the *equilibrium conformation* state diagram for homopolypeptides in the (N, T) plane, spanning a broad range of temperatures and chain lengths, Figure 2a.[17] In our model, the values of $\varepsilon_{\rm hp}$ and $\varepsilon_{\rm s}$ correspond to the middle of the ground state diagram[14] of a 25-residue homopolypeptide in the ($\varepsilon_{\rm hp}$, $\varepsilon_{\rm s}$) plane, within the α-helix domain but close to the various transition lines. This choice ensures that the model exhibits the entire spectrum of the possible polypeptide structures and realistically differentiates between the preferential folds as the chain length varies. This is evident from the contour-plot of the degree of helicity, $\Phi_\alpha = \langle n_\alpha \rangle/N$ (i.e., the average fraction of H-bonds in helical structures), overlaid with the state diagram in Figure 2a.

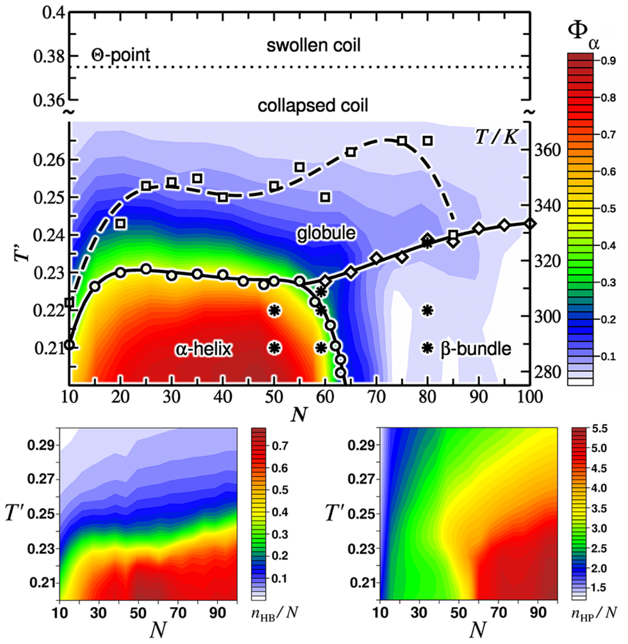


Figure 2. (a) The equilibrium conformation state diagram in (N, T) plane overlaid with the contour-plot for helicity, Φ_α. Squares correspond to the maxima of the partial heat capacity due to hydrophobic contacts, $C_{V,{\rm h}}(T) = \partial^2 U_{\rm hp}(T)/\partial T^2 = 0$; horizontally spread circles and diamonds are the data for the maxima of the total $C_V(T)$; vertically arranged circles are the inflection points of $\Phi_\alpha(N)$. The eye-guiding lines are the fourth degree polynomial fits to the data points. Stars correspnd to the (N, T) pairs for which additional separate NVT simulations have been performed and the free energy profiles obtained (see next section). (b, c) The (N, T) contour-plots for the average numbers of H-bonds (HB) and hydrohobic contacts (HP) per monomer, $\langle n_{\rm HB} \rangle/N$ and $\langle n_{\rm HP} \rangle/N$, respectively.

The details of obtaining the various transition lines in the state diagram have been reported earlier,[17] whereas here we recapitulate that the main method was to locate the inflection points on the surface of either the total potential energy, $U(N, T)$, its different contributions, or the helicity, $\Phi_\alpha(N)$; see Figure 2a. To exemplify, panels b and c in Figure 2 show contour-plots for the average numbers of H-bonds and hydrophobic contacts per C_α site, $\langle n_{\rm HB} \rangle/N$ and $\langle n_{\rm HP} \rangle/N$ (the respective energy contributions being proportional to those). Clearly, the main folding transition line (circles and diamonds in Figure 2a) closely follows the yellow-green border in Figure 2b, $\langle n_{\rm HB} \rangle/N \approx 0.4$, revealing that formation of H-bonds is the major driving force of the final folding process for all chain lengths. Obvious from Figure 2 is also a remarkable correlation between the abrupt drop in helicity, $\Phi_\alpha(N)$, and the steep raise in hydrophobic contacts, $\langle n_{\rm HP} \rangle/N$, as N passes through the midregion, $N = 50, ..., 70$, whereas the relative contribution of H-bonding notably reduces for $N > 70$.

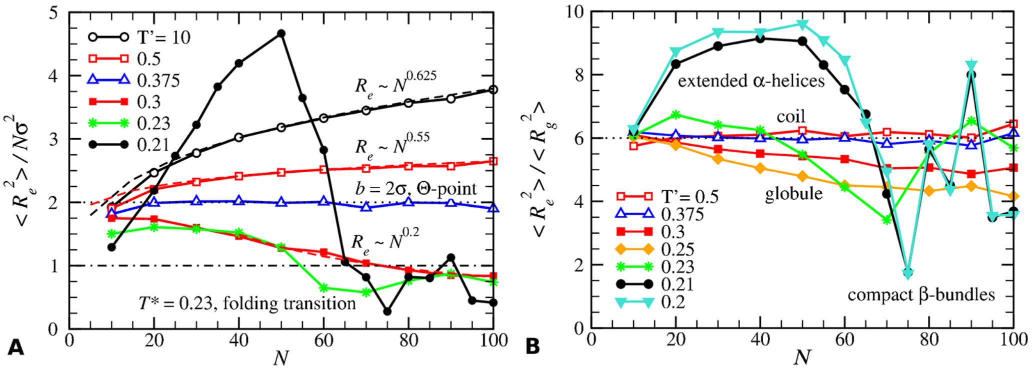


Figure 3. (A) Mean-square end-to-end distance $\langle R_e^2(N)\rangle$ scaled by $N\sigma^2$, as a function of the chain length, N (C_α atoms), and reduced temperature, $T' = T/\varepsilon$. $T' = 10$ - open circles, 0.375 - open triangles-up, 0.5 - open squares, 0.3 - filled squares, 0.23 - stars, and 0.21 - filled circles. Dashed lines are Flory's power fits for $\langle R_e^2(N)\rangle \sim N^\nu$. Dotted and dot-dashed lines represent the regimes where Kuhn's segment length is $b = 2\sigma$ and σ, respectively. (B). The ratio $\langle R_e^2\rangle/\langle R_g^2\rangle$ for the tube model, with additional lines for $T' = 0.25$ - filled diamonds and 0.2 - filled triangles-down.

The state diagram highlights the thermal regimes of stability for various conformational patterns in polypeptides, including swollen, random, and collapsed coils, globular structures, α helices, and compact β bundles. On the basis of the state diagram, the two major findings in regard to the *equilibrium folding dynamics* of homopolypeptides are the following: (i) Upon lowering the temperature, the main folding transition is preceded by a continuous, relatively smooth, hydrophobicity driven process of the coil compaction, resulting in a partially structured "molten globule" state (squares fitted by the dashed line in Figure 2a) where both α and β motifs are often present and interchange in a ratio dependent on the chain length. (ii) Below the folding temperature, a generic homopolypeptide undergoes a *chain-length-induced α−β* transition (circles and stars fitted by the vertical solid line in Figure 2a).

These observations are well reflected in the behavior of the mean-square end-to-end distance, $\langle R_e^2(N)\rangle$, and the chain elongation measure, $\langle R_e^2\rangle/\langle R_g^2\rangle$, presented in Figure 3, where $\langle R_g^2\rangle$ is the mean-square radius of gyration. Judging by the power fits for the data in Figure 3A, $\langle R_e^2(N)\rangle \sim N^\nu$, it is clear that the "molten globule" state occurring below $T' = 0.3$ is significantly more compact than the "classical" globule for which the predicted Flory's power is $\nu = 1/3$. Furthermore, the data for $T < T^*$ indicate that shorter chains, $N < 60$, have a strong tendency to self-arrange in extended α-helical structures, whereas the longer ones, $N > 70$, mostly fold into compact cross-linked β-bundles whose packing also depends on the chain length (note broad variations in $\langle R_e^2\rangle/\langle R_g^2\rangle$, Figure 3B).

The observed conformation equilibria and transitions can be rationalized as follows. Starting from coil-like conformations at high temperatures, $T' > 0.3$, upon *gradual* cooling down, polypeptide chains (not too short, $N > 20$), first, collapse into the "molten globule" state, thereby maximizing the hydrophobic contacts within the chain. With the temperature further reduced through the folding transition point, T^*, progressive formation of H-bonds within the chain results in stabilization of well-structured folds, α and β domains. This process yields, however, different secondary and tertiary structures within short and long chains. That is, sufficiently short polypeptides, $N < 60$, easily escape from the compact globular conformations and tend to stabilize in helical folds (most often extended helix), by saturating the H-bonds for the expense of the hydrophobic interactions (note the obvious maxima in $\langle n_{\mathrm{HP}}(T)\rangle_{N=\mathrm{const}}$ at $T' \approx 0.24$ in Figure 2c). In contrast, long chains, $N > 60$, are "precured" in cross β-sheet structures, or β-bundles, originating from the compact globular conformations that become more and more cluttered as N grows. Looking ahead, we emphasize the importance of the molten globule state as a precursor of the β-dominated folds in longer polypeptides. Namely, the process of gradual polypeptide compaction into molten globule and its subsequent restructuring into β-bundles upon *quasi-equilibrium* cooling follows essentially the same *condensation−ordering* route as that reported earlier for the formation of fibrillar structures in a concentrated solution of small peptides.[19]

Homopolypeptides: Conformation Free Energy—Diversity vs Stability. Multicanonical simulations which we used for obtaining the equilibrium state diagram of homopolypeptides are known to be very efficient in exploring the configuration space by spanning a range of temperatures and bypassing the free energy barriers existing at lower T values when the system returns from the higher T end. The obtained equilibrium statistics describes well the likelihoods of various conformations observed within an *ensemble* of the same noninteracting chains (i.e., an infinitely dilute solution of those). However, additional analysis of the *single* chain conformation behavior and the corresponding free energy landscapes at constant temperature is needed for assessing the metastability issues[34] which might be overlooked otherwise. This is important because well below the transition point sampling of the conformation space by a single chain can be significantly hindered and, thus, different from the equilibrium conformation sampling by an ensemble of same chains. Below, we address the metastability issues arising within the chosen polypeptide model.

In order to examine the (meta-)stablility of the α and β dominated structures in long polypeptides, we performed extensive simulations for N = 50, 60, and 80 and calculated the free energy profiles (FEPs) at a set of temperatures near and below the respective folding points, $T \le T^*$ (stars in Figure 2a); see Figure 4. For each N, we start the simulations near T^* with either collapsed random coil or molten globule configurations and, if not stated otherwise, for lower T values the initial configuration was taken from the most populated FEP minimum obtained in the previous run (at higher T). The objective was to assess the FEP variation for different chain lengths as the temperature dropped and also probe the stability of conformations far from equilibrium (see below).

The FEPs along the "natural" reaction coordinate, $F(\xi)/k_BT$, for N = 50 are shown in Figure 4A. We see that the FEP at T' =

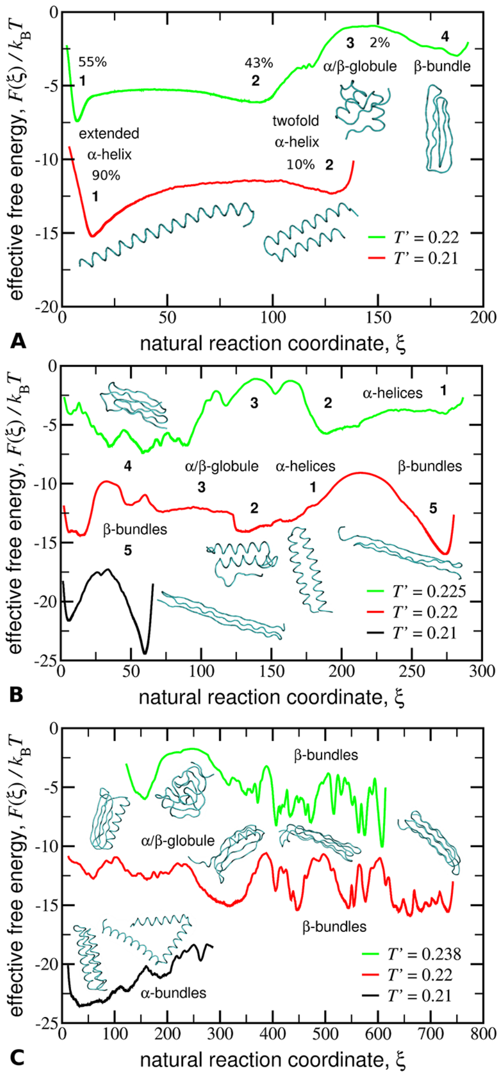


Figure 4. The free energy profiles along the "natural" reaction coordinate, $F(\xi)/k_BT$, for $N = 50$ (A), 60 (B), and 80 (C); see legends for the T' values used. The trace snapshots of representative conformations in main basins are illustrated, with a self-explanatory notation. Panel A: The data for $T' = 0.22$ were obtained starting from a collapsed random coil, and then, the stable extended helix was used as a seeding at $T' = 0.21$. Panel B: The green line, $T' = 0.225$, was obtained in a simulation initiated with a molten globule configuration, and the other two lines, $T' = 0.22$ and 0.21 (red and black), correspond to quasi-equilibrium initial configurations (taken from previous simulations). Panel C: The two simulations for $T' \geq 0.22$ (green and red lines) started from molten globule configurations, whereas the black line, $T' = 0.21$, resulted from a stretched C_α zigzag.

0.22 (≈4% below T^*) exhibits three minima and two transition states. Minima 1 (global) and 2 (local), corresponding to the extended and 2-fold helical structures, are separated by a rather low and flat free energy barrier (≈2 k_BT) due to relatively frequent bending of the helix followed by its subsequent folding onto itself. Obviously, the inevitable loss of a few H-bonds within the kink in the 2-fold helix is compensated by the gain of hydrophobic interactions within the contact area between the two helical barrels. In contrast, the free energy barrier of $\alpha-\beta$ transformation is rather high (≈5 k_BT), being associated not only with the loss of H-bonds but also the necessity of passing through an intermediate transition step—a globule containing both α and β motifs, which represents an entropic bottleneck between α and β domains. This explains the rather low population of both the globule and β-bundle states (the system spent ≈2% in basins 3 and 4 altogether). The data for lower T' illustrate how the FEP changes upon cooling down of the chain in *quasi-equilibrium*, i.e., taking the most populated conformation (the extended α-helix in this case) as a starting point. The free energy barriers progressively raise and, as a result, the metastable compact β-bundles, initially appearing ≈3 k_BT higher in free energy than α-helical folds, at $T' = 0.21$ become totally inaccessible for the chain with $N = 50$. In this case, the equilibrium fold is stabilized energetically, i.e., corresponding to the ground state conformation.

The effect of the chain length becomes evident from the comparison of the FEPs for homopolypeptides with $N = 50$, 60, and 80 obtained at $T' \geq 0.22$, Figure 4. The FEP for $N = 60$ at $T' = 0.225$ is similar to that observed with $N = 50$ at $T' = 0.22$, albeit reversing in terms of the dominance of α and β motifs. In accord with the equilibrium conformation diagram, Figure 2a, for $N = 60$ the extended α-helix structure is only a local minimum, whereas for $N = 80$ it is virtually absent from the FEP. Generally, the amount of α-helical elements within the combined $\alpha-\beta$ structures found in quasi-equilirium simulations progressively decreases with N.

Strikingly, the FEPs obtained for $N = 60$ and 80 suggest *thermodynamic coexistence of a variety of structural folds well below* T^*, without a clear preference of any particular structure (except for the general trend of greater stability of β-sheets over helices). As one can see in panels B and C of Figure 4, down to $T' = 0.22$, the FEPs for $N = 60$ and 80 are broad and exhibit a complex shape with many local minima, mostly of similar depth, which are separated by the barriers of a few k_BT units. The larger stepwise barriers indicate grouping of the minima into rather spread basins corresponding to domains of distinct folding habit, with different proportions of α and β patterns. Clearly, the pathways connecting these basins pass through entropic bottlenecks, implying substantial unfolding of the polypeptide in the course of its restructuring, as we already discussed above for $N = 50$. These refolding processes are progressively hindered at lower temperatures whereby the gamut of accessible conformations, once the system is quenched from one of the minima, becomes severely restricted.

As is illustrated by the free energy data in the cases of chains with $N = 60$ and 80 simulated at $T' = 0.21$, upon lowering the temperature, the choice of the final fold is ultimately limited by the selection of the local free energy minimum that is nearest to, or easiest to access from, the starting configuration. For instance, in a simulation with $N = 60$ at $T' = 0.21$ started from an elongated 3-fold β-bundle (see basins 5 at $T' = 0.22$ and 0.21 in Figure 4B), the chain configuration is only varied due to sliding of the β-strands with respect to each other and exchange of the H-bonds between the three sheets. One can argue, though, that in this case the initital configuration was taken from the global minimum separated from the rest of the conformation space by the highest free energy barrier (as it appears to be the case). Therefore, in order to probe the accessibity and stability of α-helices, i.e., structures far from the thermodynamic equilibrium for the chain with $N = 80$ (simulated at $T' = 0.21$), we opted to start with a completely stretched C_α zigzag configuration. This choice led to rapid emergence of purely α-helical folds which failed to transform into the thermodynamically stable β-bundle dominated state. The resulting "kinetic" FEP (black line in Figure 4C) illustrates the stability of the so-obtained, and otherwise inaccessible with $N = 80$, helical structures.

In the Supporting Information, we provide additional FEP data (Figure S1, jp300990k_si_001.pdf) for $N = 60$ obtained with even longer DMD trajectories (simulations with doubled length, 4×10^6 blocks) starting with either α- or β-dominated conformations, where one can clearly see the inaccessibility of certain parts in the conformation space in both simulations. Such a restricted conformation behavior can only be associated with *high "kinetic barriers"*, i.e., long-time dynamics of substantial cooperative restructuring within the chain configuration while minimizing the inevitable energetic losses during these rare events. In the case of even longer chains (e.g., $N = 80$, FEPs are not presented due to the complexity and roughness of the profiles), the issue of "kinetically isolated" α- and β-folds is, of course, even more pronounced. It is worth reminding here that this *"kinetic separation"* of the folded conformation domains is completely absent in our earlier quasi-equilibrium simulations in the multicanonical ensemble.

The effect of the initial configuration on the homopolypeptide folding preferences has been investigated further by performing a number of simulations for $N = 50$, 60, 70, 80, and 100 at $T' = 0.2$ starting from various configurations that either originated from earlier simulations at T^* or were generated from scratch. To exemplify the observed common trends, in Figure 5, the probability distributions for the length of helically folded fractions within the chain, $P(l_{\text{helix}})$, are given for $N = 50$, 60, and 100.

Thus, we examined the configuration dynamics in two simulations for $N = 50$ at $T' = 0.2$ starting with two distinct configurations: either a stretched C_α zigzag configuration or a twisted 3-fold β-bundle, resembling the collagen structure (from basin 4 in Figure 4A). The corresponding $P(l_{\text{helix}})$ distributions are given in Figure 5A, where we also included a selection of snapshots to illustrate typical conformations adopted by the fiftymer in the course of the two simulations for the β and α rich states. We emphasize that in both cases the two primary motifs survived over the entire simulation length (10^6 blocks), although in the second case the collagen-like arrangement eventually turned into a less specific β-bundle and from time to time a small α-helical fragment emerged. This is in contrast to our earlier quasi-equilibrium simulations where β-folds were absent from the free energy profile already at $T' = 0.21$, red line in Figure 4A. Notice how the rather unlikely and thermodynamically unstable (for $N = 50$) β-bundle conformations prevail upon rapid and deep cooling, when there is virtually no chance for the chain to refold.

In the other cases studied ($N = 60$, 70, 80, and 100, $T' = 0.2$), the two initial configurations were either a stretched C_α zigzag or a somewhat collapsed random coil, both generated from scratch. Remarkably, a completely stretched chain of any length readily adopts helical structures which do not unfold at later stages, often forming 2/3-fold α-bundles or the extended helix conformation, with longer chains being more prone to making helix bundles. To the contrary, in simulations started with a collapsed random coil, only short helices form initially, resulting from sufficiently long uncluttered portions within the polypeptide. Subject to the initial random conditions (configuration and velocities), at later stages, chains of $N = 60$, ..., 70 might end up eventually as either completely helical or β-bundle, whereafter the conformation merely fluctuates without any further global transformations (as we saw earlier for the 3-fold β-bundle, cf. red and black lines in Figure 4B). For $N > 70$, the clutter in the initial collapsed coil configuration completely prevents the chain from forming notable helix content, whence the resulting structure invariably being the β-bundle, as is exemplified for $N = 100$. These folding trends are well illustrated by animations for the case of $N = 80$ supplied in the Supporting Information [Animations S1 (jp300990k_si_002.zip), S2 (jp300990k_si_003.zip), and S3 (jp300990k_si_004.zip)].

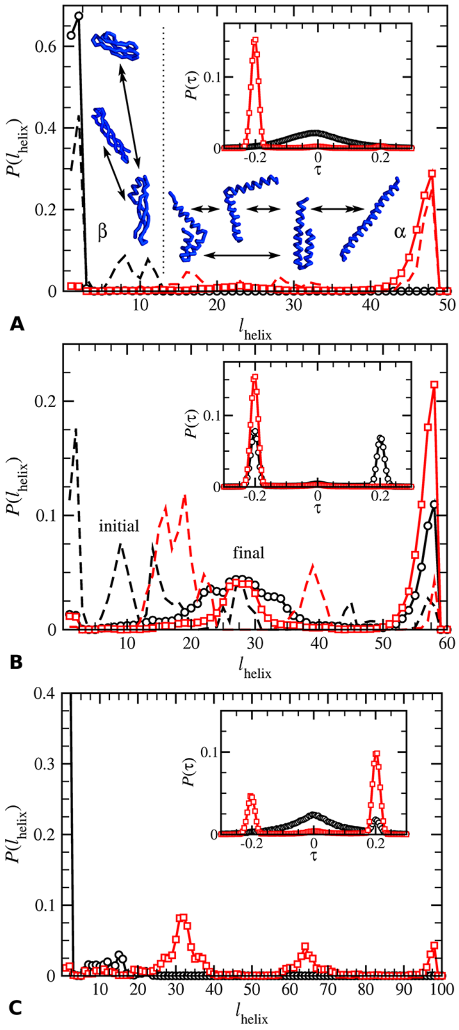


Figure 5. The probability distributions for the number of monomers (length) of helical fragments within a collapsed (folded) polypeptide chain, $P(l_{\text{helix}})$, obtained at $T' = 0.2$. The order of the graphs top to bottom: $N = 50$, 60, and 100. The dashed lines given for $N = 50$ and 60 are the partial data obtained during the equilibration (initial) stage of 10^4 DMD blocks, and the solid lines with symbols are the total (final) data. Insets: the corresponding distributions for the torsion along the chain, $P(\tau)$, where the sign of τ indicates the handedness (chirality) of the helical "torque": negative for left-handed turns and positive for right-handed turns.

We again note that, for long chains too, the thermodynamically unstable structures, i.e., α-helices, while being only rarely visited in quasi-equilibrium simulations, prevail at a sufficiently low temperature, provided the conditions are suitable for their emergence (initially uncluttered extended configurations). Judging by the $P(l_{\text{helix}})$, we conclude that most often the number of monomers in stable helical fragments, l_{helix}, is proportional to N/k, where $k = 1, 2, 3, ...$. With $N < 65$, i.e., in the helix-rich part of the equilibrium state diagram, Figure 2a, the most stable (both globally and locally) conformation is the fully stretched helix ($k = 1$), while the longer the chain the larger k values are populated, which minimizes the contribution of loose tails and helps packing by maximizing hydrophobic contacts, $n_{\rm HP}$, cf. $P(l_{\rm helix})$.

In the insets of Figure 5, we also show the distributions for the average torsion along the chain, $P(\tau)$,[35] which serves here as an indicator of the helix chirality: negative and positive values correspond to the left- and right-handed helical turns, respectively. We see that both left- and right-handed helical fragments are found in long chains, provided a few such fragments are present. Of course, the tube model does not have a preference for the handedness of emerging helices. It is instructive, though, to see that the chirality is well-conserved when the fully stretched helix emerges, implying that at sufficiently low temperatures it is rather unlikely for a helically shaped polypeptide to spontaneously unfold (and refold), even though the helix might be thermodynamically metastable.

To resume, we have seen in this section that even within the rather simplified "tube" model the free energy landscapes for sufficiently long homopolypeptides possess many basins with minima of comparable depth. The major basins are separated by high free energy barriers (>5 $k_{\rm B}T$) due to the bottlenecks associated with substantial rearrangements within the chain—either α–β transformations or restructuring within β-bundles. These features originate, of course, from the switch-like collective hydrogen bonding and hydrophobic interactions, which at low T make the selection of the actual folded structure a kinetic process based essentially on the "first come, first served" principle.

We conclude also that the low T conformations populated by homopolypeptide most in multicanonical simulations, with the chain being periodically "heated" and "quenched", are dominated by a rather restricted subset of all the possible relatively stable folded states. Moreover, our simulation results for long chains, $N > 50$, unveil a common mechanism of damping the energetically favorable helical structures on the background of the readily accessible β-bundles when starting from the collapsed coil or, especially, molten globule states (as is seen in our multicanonical simulations). That is, the cluttered molten globule, which results from the collapse and further compaction of a random coil, is effectively much closer in the configuration space to the cross-linked β-bundle structures than to α-helical folds. In other words, the helices form a rather small and isolated island in the configuration space that is only accessible starting with sufficiently extended and uncluttered configurations which are rarely visited by the molten globule.

Heteropolypeptides: Smoothing of the Free Energy Landscape. In what follows, we attempt to (approximately) reproduce two tertiary structures found in small proteins, namely, the 3-fold helical bundle of the 48-residue 1lp1 (we use $N = 50$) and 58-residue 1pgb domain (we use $N = 60$) in which four-strong β-sheet is combined with a helix diagonally attached to it; see Figure 6. Following the ideas of Hoang et al.,[16] we aim to examine the expected smoothing of the free energy landscape and appearance of a funnel leading the chain into a specific fold determined by the (amino acid) interaction sequence.

Our three-letter code distinguishes C_α sites with a propensity to participate in formation of either α-helices, β-strands, or unstructured coil-like fragments. Thus, we describe the interaction sequence by a set of numbers $\{n_i, m_j, ...\}$, where indices $i, j \in \{\alpha, \beta, \text{"c"}\}$ (the latter standing for "coil-like").

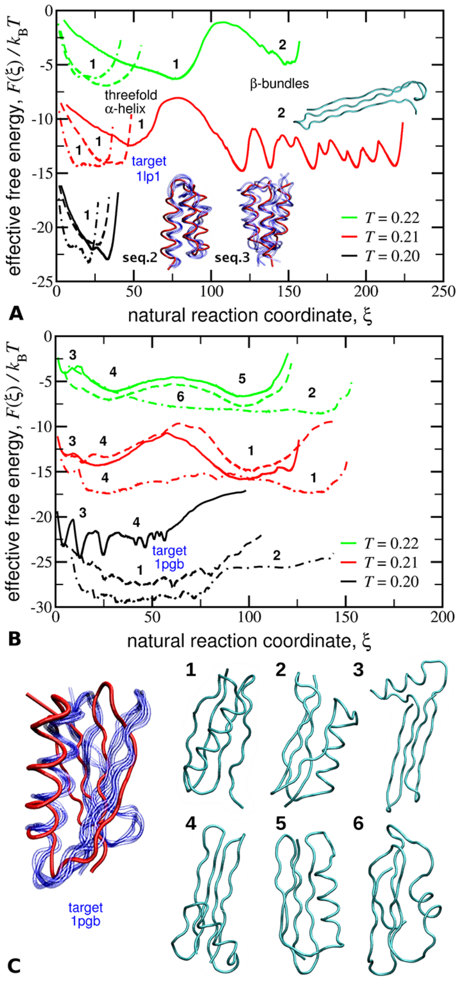


Figure 6. (A, B) The free energy profiles along the "natural" reaction coordinate, $F(\xi)/k_{\rm B}T$, for $N = 50$ (A) and 60 (B) and three different C_α sequences in each case: solid, dashed, and dot-dashed lines are for sequences 1, 2, and 3, respectively. See legends for the T values used. All simulations were started from a collapsed random coil. In panel A, the target protein structure, 1lp1, is shown (red) overlaid with several approximating configurations from simulations (blue) for sequences 2 and 3 (basin 1, rmsd ≈ 5.9 Å at $T' = 0.21$). In panel C, the trace snapshots of representative conformations from various basins for $N = 60$ are given, where the basin numbers correspond to those in panel B; the target protein structure, 1pgb, is overlaid with several approximating configurations for sequence 3 (basin 1, rmsd ≈ 7.2 Å at $T' = 0.21$).

These count the number of monomers of a particular type occurring in the sequence when counting starts from one of the chain ends. The simulated sequences for $N = 50$ and 60 and the relevant interaction parameters corresponding to each monomer type, ε_α, ε_β, and ε_c, are given in Table 1 (the values are in units of the energetic depth of one hydrogen bond, and the cross-interactions between monomers of different types are assinged a mean value so that, for instance, $\varepsilon_{\alpha\beta} = 0.5(\varepsilon_\alpha + \varepsilon_\beta)$). As one can notice, a slight complication arises due to the fact that different sequences can be introduced for different interaction types, although we kept such variations within one sequence to the minimum. Note also that we do not make a difference between α and β monomers in terms of H-bond strength, while for "coil" type $\varepsilon_c(\text{HB}) = 0$.

Targeting the 1lp1 Protein Domain. Knowing from our previous simulations that homopolypeptides up to ≈60

Table 1. The C_α Interaction Sequences Used in Summations

N=50, sequence 1	**target: 1lp1**			
Interaction	Sequence	ε_α	ε_β	ε_c
hydrophobic (HP)	{ 13α, 5c, 13α, 3c, 14α, 2c }	-0.05	–	-0.005
H-bonds (HB)	same as HP	-1.0	–	0
stiffness (ST)	same as HP	0.3	–	0.1
N=50, sequence 2				
HP	{ 12(3c, 1α), 2c }	-0.05	–	-0.005
HB	{ 14α, 4c, 13α, 2c, 15α, 2c }	-1.0	–	0
ST	same as HB	0.3	–	0.1
N=50, sequence 3	same as sequence 2 except HP:	-0.10	–	-0.005
N=60, sequence 1	**target: 1pgb**			
HP	{ 9β, 2c, 9β, 3c, 14α, 3c, 9β, 2c, 9β }	-0.04	-0.05	-0.005
HB	same as HP	-1.0	-1.0	0
ST	same as HP	0.3	0.9	0.1
N=60, sequence 2				
HP	{ 9β, 2c, 9β, 2c, 16α, 2c, 9β, 2c, 9β }	-0.025	-0.05	-0.005
HB	same as HP	-1.0	-1.0	0
ST	same as HP	0.3	0.9	0.1
N=60, sequence 3				
HP	{ 2[5(1β, 1c), 1c], 4(1α, 3c), 2[1c, 5(1c, 1β)] }	-0.1	-0.05	-0.005
HB	same as sequence 2	-1.0	-1.0	0
ST	same as sequence 2	0.3	0.9	0.1

monomers only rarely sample β-structures, in the case of the 1lp1 structure modeled with $N = 50$, it seemed straightforward to start simply with the original ε values for α-monomers, i.e., $\varepsilon_\alpha(\text{HP}) = \varepsilon_{\text{hp}} = -0.05$ and $\varepsilon_\alpha(\text{ST}) = \varepsilon_s = 0.3$, while significantly reducing these parameters for c-monomers. Thus, in sequence 1, we only introduced two short coil-like fragments, the sole purpose of which was to produce kinks in the presumed overall helical structure. Somewhat unexpectedly, though, these two kinks were sufficient to also produce a rather competitive 3-fold structure of cross-linked β-strands, which is depicted in Figure 6A near the corresponding local minima in the free energy profiles (basins 2). As is seen in Figure 6A, at $T' = 0.22$, this β-bundle appears virtually as stable as the target structure and possibly becomes thermodynamically stable at $T' = 0.21$ (with a reservation of insufficiently long simulation).

Nonetheless, comparing the FEPs and configurations obtained in this case with those of the homopolypeptide, Figure 4A, we see that, even with this *ad hoc* sequence, sampling of the configuration space has been essentially restricted to two free energy basins where the extended helix (mostly populated by the homopolypeptide) and compact β-bundles are avoided. Furthermore, at $T' = 0.2$, the sequence-enabled chain predominantly samples conformations similar to the target structure (basin 1), although the corresponding basin is still rather broad owing to large scale fluctuations in the chain configuration. The trend of constraining the polypeptide folding toward a selected structure becomes even more pronounced for the other two sequences (seq. 2 and 3 in Table 1) in which we shortened the *c*-fragments and introduced an alternating "every fourth" hydrophobicity pattern, taking into account the typical number of C_α sites per turn in the α-helix. Finally, in seq. 3, $\varepsilon_\alpha(\text{HP})$ was doubled so as to better stabilize the 3-fold α bundle structure. The corresponding FEPs reveal that, although the use of the helix-tailored HP pattern resulted in removal of the unwanted tendency to form β-strands, increasing $|\varepsilon_\alpha(\text{HP})|$ did not improve notably the stability of the target structure.

Targeting the 1pgb Protein Domain. With $N = 60$, the target structure (1pgb) is more complex, containing an α-helical fragment in the middle of the chain which is hydrophobically attached to an underlying β-sheet made of the β-stands self-arranging at the chain ends. In this case, we had to modulate the hydrophobicity and stiffness along the chain so as to prevent the mid region of the chain from participating in the β-sheet formation, and yet stabilize the putative α-helix on top of the β structure. To achieve this, we, first, increase the stiffness of β fragments (to prevent formation of helices) and, second, reduce the strength of hydrophobic contacts, making the α-monomers slightly less hydrophobic than the β-monomers; see Table 1. As with $N = 50$ above, short fragments of virtually inert c-monomers are introduced where the turns within the sought structure are expected. Again, initially we use the same sequence code for all the interactions (seq. 1). This already results in the FEPs with only two major minima; see solid lines in Figure 6B where basins 3/4 and 5/1 are identified by the respective conformations in Figure 6C; cf. Figure 4B. Note that two distinct conformation states appear (basins 3/4 and 5/1), which differ by the relative orientation of the chain tails within the β-sheet (parallel in basins 3/4 vs antiparallel in basins 5/1). Hence, in basins 3/4, the α helix has both its ends on the same butt of the β-sheet, which is not the case in the target 1pgb fold, whereas in basins 5/1 the ends of the helix occur on different β-sheet butts and, thus, the helix is correctly placed relative to the β-sheet. At $T' = 0.2$, however, basins 5/1 which correspond to the conformation state generally resembling the target fold (albeit containing a notably shorter helix) are not populated anymore. Instead, basins 3/4 split into a few local minima due to variations in the H-bonding patterns caused by shifts and drifts of the strands within the β-structure.

When analyzing the configuration evolution in basins 5/1 obtained with sequence 1, we noticed that the helix appeared on average to be too short because a few α-monomers at the helix ends were often pulled by the β-sheet and then happened to be stuck. Therefore, in sequence 2, we opted to further reduce $|\varepsilon_\alpha(\text{HP})|$ while retaining the original value of $\varepsilon_\beta(\text{HP})$; see Table 1. As is revealed by the resulting FEPs (dashed lines in Figure 6B), the stability of the target structure generally increased and it became the only conformation state populated at $T' = 0.2$. Next, in attempt to improve on the helix placement and attachment to the β-sheet, we introduced alternating hydrophobicity patterns, (1β,1c) and (1α,3c), within β and α fragments, respectively, and at the same time increased $|\varepsilon_\alpha(\text{HP})|$ in order to compensate for the decrease in the number of possible α–β hydrophobic contacts, cf. sequences 2 and 3 for $N = 60$ in Table 1. From the FEPs for sequence 3 (dot-dashed lines in Figure 6B), it is evident that this final change in the interaction sequence(s) resulted in a smaller free energy barrier separating the two competing structures, basins 1 and 4, whereas the target structure remained the only populated state at $T' = 0.2$.

A common measure of the convergence of the simulated folds toward target protein structures is the root-mean-square deviation (rmsd) of the atom positions within the model from their respective target locations. In our case, we calculated rmsd probability distributions for C_α sites, which are presented in Figure 7 for both $N = 50$ and 60 at the intermediate temperature, $T' = 0.21$. We see that in both cases there is an improvement of the target representation upon tailoring the sequences. For $N = 50$, we notice a steady convergence to the 1lp1 fold and, thus, the best fit in terms of rmsd is obtained with the third sequence. In contrast, for $N = 60$, seqence 2 deviates less from the target structure than sequence 3, but both of these sequences (2 and 3) are a major improvement in comparison with sequence 1.

To summarize, despite the use of a crude polypeptide model, both target protein folds, 1lp1 and 1pgb, were successfully

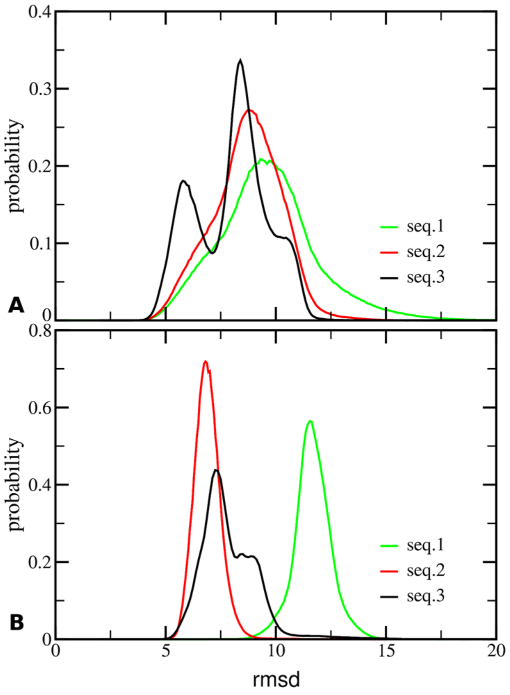


Figure 7. Probability distributions for root-mean-square deviation (rmsd) of C_α sites from the target structures obtained for three interaction sequences for $N = 50$ (A) and $N = 60$ (B), $T' = 0.21$.

reproduced and *thermodynamically stabilized* in a range of temperatures below the folding transition point. In these simulations, we have established a protocol for informed tailoring and fine-tuning of the polypeptide tube model against the real protein structures.

CONCLUSIONS

In this report, we used a generic coarse grained polypeptide model to study the folding behavior of homo- and heteropolypeptides, for which purpose we calculated the chain free energy profiles along the optimal, so-called "natural", reaction coordinate. In particular, in the case of homopolypeptides, we examined the low temperature folding trends predicted by the equilibrium conformation state diagram in the (N, T) plane. The FEPs obtained at a number of N and T values within and outside the region of the chain length induced $\alpha-\beta$ transition revealed that for a sufficiently long homopolypeptide there is no unique thermodynamically stable structure. The free energy landscapes for $N \geq 60$ appear to have multiple, similarly weighted and rugged basins, with minima of comparable depth, whereas the barriers between those represent narrow bottlenecks corresponding to hindered conformational rearrangements which are, at low temperature, costly energetically and unlikely entropically.

The most striking observation with homopolypeptides is that there is a discrepancy in the homopolypeptide folding trends upon gradual cooling down, as observed in quasi-equilibrium simulations, and the folding processes initiated directly at low temperatures and started from different initial configurations. That is, the accessibility of helical folds is severely reduced for long chains when the starting configuration originates from the collapsed coil or, especially, molten globule domains (dominant during the quasi-equilibrium cooling), whereas extended configurations (sampled very rarely by the molten globule) result mostly in purely helical structures.

We also saw that the quasi-equilibrium compaction and folding of polypeptides is driven by the same *condensation–ordering* mechanism as that governing the aggregation and fibrillation of small peptides in a solution.[19] This finding highlights the common trends in the emergence of β-dominated aggregated structures due to both the intra- and intermolecular forces in any kind of peptide chains.

With heteropolypeptides modeled by introducing monomer sequences based on a three-letter interaction code, we managed to reasonably well reproduce two small protein structures: 1lp1 with $N = 50$ and 1pgb with $N = 60$. Apart from that, we have clearly shown in both cases that upon enabling and tailoring the heteropolypeptide interaction sequence the free energy landscape becomes drastically simpler and smoother as compared to that of the corresponding homopolypeptide. This provides *evidence of the progressive formation of a free energy funnel upon decreasing temperature*, owing to the specifics of the amino acid interactions within a heteropolypeptide.

Despite the difference in the simulated folding behavior of homo- and heteropolypeptides, we are tempted to combine our observations in a general view on the protein folding problem. Prior to that, we note that within the current model one cannot expect accurate reproduction of all the details of the tertiary structure of a protein, but rather one should consider this model as a convenient tool for analyzing the folding trends general for a vast majority of proteins. In this respect, we believe that already at this stage some valuable insights have been gained.

(1) Our results for homopolypeptides show how diverse and complex the free energy landscape can be even for a relatively simple polypeptide model that omits the side chain interaction details altogether. The general conformation behavior of real proteins, most of which are much longer than the chains considered here, should be significantly more involved, even though the specifics of interactions within a particular amino acid sequence reduce to a certain degree the roughness of the underlying free energy surface.

(2) One should also consider the diversity of real proteins in their folding habits. (i) It seems reasonable to assume that a smooth free energy funnel forms upon lowering T in the case of relatively short and fast folding proteins which readily attain their native structure, just as we observed with the two model heteropolypeptides. (ii) To the contrary, for proteins which upon denaturation cannot spontaneously refold back into their native state, one cannot as straightforwardly adopt the folding funnel view, unless effectively accounting for all the relevant external factors assisting the correct protein folding, such as the role of chaperones and the local environment within the living cell. Instead, we would presume in this case the free energy landscapes (perhaps, funnel-like but) having more in common with the profiles we obtained for homopolypeptides. (iii) There is yet another type of proteins—those resembling a random coil behavior, i.e., without an incline toward any structured fold, for which our results are not applicable at all.

In the light of the above considerations, we would like to emphasize the importance of discriminating between, at least, these three protein classes, as it appears that the folding funnel concept, albeit being very convenient, hardly suits all the diverse protein cases.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figure S1 (jp300990k_si_001.pdf): Additional free energy profiles, $F(\xi)/k_BT$, obtained for homopolypeptide with $N = 60$ at $T' = 0.21$ from simulations comprising 4×10^6 blocks. Animation S1 (jp300990k_si_002.zip): The initial stage of the simulation for homopolypeptide with $N = 80$ at $T' = 0.20$, starting from the stretched C_α zigzag initial configuration and quickly folding into an α-helix dominated conformation. Animation S2 (jp300990k_si_003.zip): The initial stage of the simulation for homopolypeptide with $N = 80$ at $T' = 0.20$, starting from the (extended) random initial configuration and quickly folding into an α-helix dominated conformation. Animation S3 (jp300990k_si_004.zip): The initial stage of the simulation for homopolypeptide with $N = 80$ at $T' = 0.20$, starting from the collapsed-coil initial configuration and quickly folding into the β-sheet dominated conformation. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: abrukhno@gmail.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We express our gratitude to Dr Krivov for providing us the trajectory analysis tools for calculating FEPs. This work was supported by the EPSRC grant EP/G026165/1.

■ REFERENCES

(1) Onuchic, J. N.; Socci, N. D.; Luthey-Schulten, Z.; Wolynes, P. G. *Folding Des.* **1996**, *1*, 441−450.

(2) Dobson, C. M.; Sali, A.; Karplus, M. *Angew. Chem., Int. Ed.* **1998**, *37*, 868−893.

(3) Freddolino, P. L.; Harrison, C. B.; Liu, Y. X.; Schulten, K. *Nat. Phys.* **2010**, *6*, 751−758.

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

(5) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689−12698.

(6) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334−350.

(7) Allen, L. R.; Krivov, S. V.; Paci, E. *PLoS Comput. Biol.* **2009**, *5*, e1000428.

(8) Best, R. B.; Hummer, G. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6732−6737.

(9) Becker, O. M.; Karplus, M. *J. Chem. Phys.* **1997**, *106*, 1495−1517.

(10) P. G. Bolhuis, C. D; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877−5882.

(11) Ensign, D. L.; Pande, V. S. *Biophys. J.* **2009**, *96*, L53−L55.

(12) Freddolino, P. L.; Liu, Y. X.; Gruebele, M.; Schulten, K. *Biohys. J.* **2008**, *94*, L75−L770.

(13) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. B.; Wriggers, W. *Science* **2010**, *330*, 341−346.

(14) Hoang, T. X.; Seno, F.; Banavar, J. R.; Maritan, A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7960−7964.

(15) Hoang, T. X.; Trovato, A.; Seno, F.; Banavar, J. R.; Maritan, A. *Biophys. Chem.* **2004**, *115*, 289−294.

(16) Hoang, T. X.; Marsella, L.; Trovato, A.; Seno, F.; Banavar, J. R.; Maritan, A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 6883−6888.

(17) Ricchiuto, P.; Brukhno, A. V.; Paci, E.; Auer, S. *J. Chem. Phys.* **2011**, *135*, 061101.

(18) Auer, S.; Dobson, C. M.; Vendruscolo, M. *HFSP J.* **2007**, *1*, 137.

(19) Auer, S.; Trovato, A.; Vendruscolo, M. *PLoS Comput. Biol.* **2009**, *5*, e1000458.

(20) Tsemekhman, K.; Goldschmidt, L.; Eisenberg, D.; Baker, D. *Protein Sci.* **2007**, *16*, 761−764.

(21) Auer, S.; Kashchiev, D. *Phys. Rev. Lett.* **2010**, *104*, 168105.

(22) Alder, B. J.; Wainwright, T. E. *J. Chem. Phys.* **1959**, *31*, 459−466.

(23) Davis, C. H.; Nie, H.; Dokholyan, N. V. *Phys. Rev. E* **2007**, *75*, 051921.

(24) Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384−2393.

(25) Iba., Y. *Int. J. Mod. Phys. C* **2001**, *12*, 623.

(26) Wang, F.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050.

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

(28) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. *Nat. Phys.* **2010**, *6*, 751−758.

(29) Krivov, S. V. *J. Phys. Chem. B* **2011**, *115*, 12315−12324.

(30) Krivov, S. V. *J. Phys. Chem. B* **2011**, *115*, 11382−11388.

(31) Berezhkovskii, A.; Szabo, A. *J. Chem. Phys.* **2004**, *121*, 9186−9187.

(32) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13841−13846.

(33) Krivov, S. V. *PLoS Comput. Biol.* **2010**, *6*, e1000921.

(34) Auer, S.; Miller, M. A.; Krivov, S. V.; Dobson, C. M.; Karplus, M.; Vendruscolo, M. *Phys. Rev. Lett.* **2007**, *99*, 178104.

(35) Magee, J. E.; Vasquez, V. R.; Lue, L. *Phys. Rev. Lett.* **2006**, *96*, 207802.