Monte Carlo studies of the thermodynamics and kinetics of reduced protein models: Application to small helical, β, and α/β proteins

Andrzej Kolinski[a)]
Department of Chemistry, University of Warsaw, ul. Pasteura 1, 02-093 Warsaw, Poland
Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

Wojciech Galazka
Department of Chemistry, University of Warsaw, ul. Pasteura 1, 02-093 Warsaw, Poland

Jeffrey Skolnick[b)]
Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

(Received 23 September 1997; accepted 30 October 1997)

Employing a high coordination lattice model and conformational sampling based on dynamic and entropy sampling Monte Carlo protocols, computer experiments were performed on three small globular proteins, each representing one of the three secondary structure classes. The goal was to explore the thermodynamic character of the conformational transition and possible mechanisms of topology assembly. Depending on the stability of isolated elements of secondary structure, topology assembly can proceed by various mechanisms. For the three-helix bundle, protein A, which exhibits substantial helix content in the denatured state, a diffusion–collision mechanism of topology assembly dominates, and here, the conformational transition is predicted to be continuous. In contrast, a model β protein, which possesses little intrinsic denatured state secondary structure, exhibits a sequential ''on-site'' assembly mechanism and a conformational transition that is well described by a two-state model. Augmenting the cooperativity of tertiary interactions led to a slight shift toward the diffusion–collision model of assembly. Finally, simulations of the folding of the α/β protein G, while only partially successful, suggest that the *C*-terminal β hairpin should be an early folding conformation and that the *N*-terminal β hairpin is considerably less stable in isolation. Implications of these results for our general understanding of the process of protein folding and their utility for *de novo* structure prediction are briefly discussed. © *1998 American Institute of Physics.* [S0021-9606(98)50606-X]

I. INTRODUCTION

Under the appropriate conditions, many small globular proteins undergo reversible thermal denaturation. Quite often the folding transition from the random coil state to the native conformation is not only highly cooperative, but is well described as an all-or-none process.[1,2] This means that at equilibrium, the population of intermediates is low, i.e., they are unstable. On the other hand, the folding process itself can be rather slow, taking from milliseconds to seconds.[3] Typically, more or less nativelike secondary structures form relatively rapidly, and it is the passage from the resulting molten globule conformation to the native structure that is the slow, cooperative step.[4,5] It is very likely that the molten globule has the overall topology of the native state, but is not as compact.[6–8] More important, at this stage of protein folding, the unique, crystal-like packing of the side chains is absent.[6] The protein interior resembles a drop of liquid with substantial conformational freedom of the side chains, and perhaps entire secondary structure elements exhibit substantial mobility as well. Actually, the fixation process of the polypeptide side chains into their more rigid native arrangement can often be the longest stage of the protein folding process. Precisely because the topology assembly process itself is more rapid, perhaps occurring on the millisecond time scale or faster, it is most difficult to study experimentally. Hence, simulations may perhaps provide some useful insights that can guide experiment. In this spirit, we describe simulations of both the thermodynamics and process of protein folding.

Over the years, various mechanisms for protein topology assembly have been proposed.[9–20] At one extreme is the ''diffusion-collision'' model,[21] which assumes that preassembled elements of secondary structure collide to form larger portions of the final conformation. Such elements need not always be present; they may fluctuate, but a key assumption is that they persist as independent, quasistable entities long enough for them to collide. An opposite view is that the secondary structure formation is coincident with the formation of loose tertiary structure. That is, topology assembly occurs by an on-site or zipper mechanism. Here, early assembled structures provide a scaffold for subsequent assembly of the remainder of the polypeptide chains.[13] A helix

a)Electronic mail: kolinski@chem.uw.edu.pl; address for correspondence: The Scripps Research Institute, Department of Molecular Biology, 10550 North Torrey Pines Road, La Jolla, California 92037.
b)Electronic mail: skolnick@scripps.edu

fragment or a β turn may serve as a nucleation site for such a mechanism. Assembly could occur from either relatively expanded conformations or from collapsed dense states that undergo subsequent rearrangement. Other mechanisms of protein assembly that could be placed somewhere between these extreme points of view have been proposed.[9,22,23]

Experimental evidence, extracted from the analysis of protein fragments, has not yet provided conclusive results for a unique topology assembly mechanism. Indeed, different proteins may assemble differently.[24] For the diffusion–collision model to be applicable, the secondary structure content in the denatured state has to be substantial. In reality, the denatured state helix content of helical proteins[25] varies from a few percent to about 15% (but sometimes it can be substantially greater[25]) for proteins having several helices. Simple statistical considerations appear to argue against a pure diffusion–collision assembly mechanism, even under the best of circumstances. It is very unlikely that when a pair of helices collide they will be exactly in-register; however, elements of this mechanism cannot be excluded. In the case of β proteins, the diffusion–collision model is more difficult to justify. According to experimental studies of denatured β proteins and their fragments, isolated β-hairpin structures are rather unstable.[26] Detectable nativelike clusters mostly consist of protein fragments involving narrow β turns.[26–28] This may suggest that an ''on-site'' sequential mechanism where β turns serve as nucleation sites for β-hairpin assembly is more appropriate. A problem with the on-site topology assembly model is that the assembly process should be faster (or at least occur on a similar time scale) than the lifetime of a fraction of the partially assembled clusters. It is unclear if such requirements can be fulfilled for complex β-type folds.

Until a few years ago, reduced models of proteins were designed in two, almost mutually exclusive, ways. Continuous models tried to reproduce the main features of protein geometry, especially the short-range correlations,[29] and usually employed an α-carbon (and sometimes side chains) united atom representation.[30–33] On the other hand, the general aspects of protein folding thermodynamics were investigated using very simple lattice homopolymers or heteropolymers that had little geometric fidelity to real proteins.[34,35] In these simplified lattice models, the protein secondary structure was very poorly defined at worst, or at best, only some classes of protein geometry could be qualitatively modeled.[12–14,20,36–39]

To retain the advantages of both approaches without their disadvantages, we have developed a series of high coordination lattice models.[39] In common with continuous models, their geometric resolution is close to that of structures determined from protein crystallography or NMR. However, in contrast to continuous space models, use of a lattice representation permits much more effective conformational sampling to occur. Lattice algorithms are at least two orders of magnitude faster than equivalent continuous models.[40] Due to discrete transitions between predefined conformations, some local energy barriers are smeared out in the lattice models. Consequently, more complex, and hopefully, more physical interaction schemes could be implemented. Thus, issues of protein folding thermodynamics, which were previously only in the venue of highly simplified models, could also be addressed; yet, closer geometric fidelity to real proteins is retained. Subsequently, the ability to fold simple motifs in the context of these models was demonstrated, although due to inadequacies in both the interaction and sampling schemes, more complex protein motifs could not be folded.[39]

Recently, within the framework of these reduced models, the origin of the cooperativity of protein folding was examined.[39,41] In agreement with much earlier work,[42,43] cooperative short-range interactions, which control the formation of secondary structure, and cooperative hydrogen bond-type interactions did not produce an all-or-none folding transition. Rather, the all-or-none conformational transition, typical of many proteins,[44] was shown to emerge from multibody interactions involving cooperative protein side chain packing. Interestingly, the lowest energy structures were essentially the same whether or not such side chain packing cooperativity was included. Thus, it is also of interest to investigate the effect of these cooperative interactions on the thermodynamics and folding mechanisms of other small globular proteins of various secondary structural classes. Preliminary work in this direction is reported here.

In this study, three protein motifs were examined. These consist of the B domain fragment of protein A,[45] which adopts a three-helix bundle fold, a computer-designed sequence,[46] which folds (on a computer) into a classical Greek key, β-barrel motif, and the $B1$ domain fragment of protein G,[47] which has a very stable, minimal α/β structure. The first two sequences have been previously folded using various realizations of the model and its potential, while the folding of protein G had very low reproducibility. In the case of the Greek key β-barrel motif, the entropy sampling Monte Carlo method[41,48–50] was also used to examine folding thermodynamics. Unfortunately, at present, this approach is computationally too expensive for protein G, but can be applied to the various domains of protein A, see below. Nevertheless, for all three molecules, to obtain a picture of the folding mechanism(s), we simulated equilibrium folding/unfolding by means of long isothermal dynamic Monte Carlo runs near the transition midpoint. This way, possible distortions of the folding pathway(s) that may result from rapid quenching of the model systems are eliminated. We explicitly examined the possible effects of protein secondary structure (all α, β type, and α/β), the cooperativity of tertiary interactions, and the relative strength of long- versus short-range interactions on the mechanism of protein structure assembly.

The outline of the remainder of this paper is as follows. In Sec. II, we summarize the salient aspects of the model. Because many features of the approach have been published elsewhere, we refer the reader to the literature for additional details.[41,46,51] Then, in Sec. III, we first present results on the thermodynamics and kinetics of protein A folding. Then, we turn our attention to the thermodynamic and kinetic behavior of our designed model β protein. Finally, results for isothermal simulations of protein G are described. In Sec. IV, the possible implications of these simulations are discussed in the context of their relationship to the folding mechanism of

real globular proteins and their implications for future developments in protein structure prediction based on this class of reduced protein models.

II. MODEL AND SIMULATION METHOD

The reduced representation of protein conformation employed in this work has been previously described in great detail.[39,46] The only change is in the side group pair interaction scale, which has been recently rederived using a larger database and a more rigorous definition of the reference state.[52] This should have no qualitative effect on the model's behavior. Thus, we merely outline the model, the conformational sampling protocols, and the interaction scheme for the reader's convenience.

A. Reduced model of protein conformation

A chain of virtual bonds between α carbons comprises a framework for the definition of the model protein's structure, which also consists of backbone carbonyl oxygens, amide hydrogens, and side chain centers of mass. The latter three are defined by the appropriate set of $C-\alpha$ virtual backbone vectors and are off-lattice.[46] The coordinates of the α carbons are restricted to a set of simple cubic lattice points, with a lattice spacing of 1.22 Å. There are 90 possible orientations of the bonds between consecutive α carbons defined as follows: $\{v\}=\{(3,1,1),\ldots\ (3,1,0),\ldots\ (3,0,0),\ldots\ (2,2,1),\ldots\ (2,2,0),\ldots\}$.[39,46] The allowed sequences of three consecutive α-carbon backbone vectors are restricted to those having close counterparts in a protein structural database.[53] A protein chain consisting of N residues is represented by $N+1$ lattice vectors connecting N α-carbon united atoms and two additional united atoms, which serve as N- and C-termini caps, respectively.

B. Monte Carlo dynamics

The dynamics of the model chains have been simulated by a Metropolis type[54] Monte Carlo, MMC, algorithm that employs random local conformational transitions and small distance motions of larger portions of the model chain. Such a model of dynamics is nonphysical for very fast, local events; however, for long time scales it constitutes a solution of a stochastic equation of motion that mimics Brownian dynamics. In principle, one could define the time scale of Monte Carlo dynamics based on the frequency of local conformational jumps. There are, however, difficulties with accounting for some correlations of various transitions and, perhaps, it is safer to adopt a more empirical approach and scale the model's time according to the total time of protein assembly.[40] It might be expected that particular models of Monte Carlo dynamics distort the time scale; however, based on comparison with Brownian dynamics simulations of very closely related idealized models, the qualitative picture of topology assembly and the predicted order of events should be correct.[40,55]

The dynamics are simulated by a long random series of attempts at local (short-range) and intermediate range conformational updates. In practice, the local transitions employed were as follows:

(1) A single rotamer random update with fixed main chain positions.
(2) Two-bond moves, where two consecutive backbone vectors are replaced by two different vectors that satisfy on-chain geometry restrictions. According to the rotamer library definition, a two-bond rearrangement requires an appropriate change of positions of three side chains. Two-bond moves also were used for random modification of the chain ends.
(3) Three-bond motions with four side groups updated.
(4) Four-bond rearrangements.

The intermediate-range conformational transitions consist of:

(5) A lateral, rigid-body type motion of a randomly selected piece of the chain, executed by a correlated replacement of two or two pairs of backbone vectors separated down the chain by several residues. The length of the displaced fragments is limited by the average secondary structure element length, and their particular span and location are randomly selected. The version for the chain ends requires only vector modifications in the middle of the model chain.
(6) A ''reptation''-type motion of a chain portion generated by a correlated replacement of two backbone vectors by a more compact four-vector fragment (and vice versa). This move requires side chain rebuilding within the entire affected fragment.

The acceptance probability of these medium-range moves by the Metropolis scheme within the relevant temperature range is rather low. Nevertheless, they allow motions of assembled fragments of protein structure. This prevents a possible bias against the diffusion–collision mechanism of assembly.[40]

The time unit of the model Monte Carlo process corresponds to N attempts at each of the local moves at randomly selected positions along the model chain. The medium-range moves are attempted less frequently (according to their spatial extent) so as to avoid severe time scale distortion.

Turning to the entropy sampling Monte Carlo, ESMC, approach,[48–50,56] we recently published an analysis of the Greek key sequence.[39,46] The technical details of the ESMC simulations employed here are exactly the same. Let us just note that ESMC samples a statistical ensemble whose relative frequency is determined by the relative entropy of various energy levels.

C. Interaction scheme

The short-range interactions are described by several terms. There are two generic terms designed to simulate the characteristic conformational stiffness of the protein chains. One term favors a proteinlike distribution of the end-to-end distance for four-vector segments of the protein backbone. The second is a bias toward the appropriate peptide bond plate correlations with respect to its second and fourth neighbors. Sequence specific secondary structure propensities were derived from a statistical analysis of a representative database of protein structures. It has been recently shown that this factorization reproduces the conformational stiffness

and secondary structure propensities of protein sequences reasonably well. The relative strength of the generic and sequence specific interactions has been previously adjusted, and here we use the same parameters.[41,46,51,57]

Tertiary interactions consist of one-, two-, and in the ''cooperative model'', four-body terms. With the minor modification described above for pairwise interactions, one- and two-body terms have been previously described.[41,46,51,57] The former depend on the location of the side chain center of mass relative to the molecular center of mass. Pair interactions involve both backbone and side chains. The explicit position of main chain atoms in our model allows for building a directional and cooperative hydrogen bond interaction scheme. Hydrogen bonds are also identical to the previous implementation, and play an important structural regularizing role.[41,46]

Recently, we have shown that to reproduce an all-or-none thermodynamic transition in model proteins, pair tertiary interactions are insufficient.[41] Unfortunately, due to inadequate statistics, for most triplets, the derivation of three-body, much less four-body, sequence specific potentials is not practical. Therefore, we used an *ad hoc* four-body potential[39,41,58,59] that reflects the side chain packing regularities seen in globular proteins.[60] A common feature of both β–β and α–α interaction patterns is that given a contact at i and j, there is very likely to be contact at $(i\pm3,j\pm3)$ and $(i\pm4,j\pm4)$. A repeat of type $(\pm1,\pm1)$ is also common for α- and β-type proteins. In helices, this reflects intrahelix interactions, while for β strands, this pattern reflects intrasheet interactions.

At present, it is unclear what would be the best choice for the strength of such interactions. An ideal solution would be to have a four-residue specific statistical potential. As mentioned above, the size of the structural database is too small to derive meaningful statistics. Thus, we assumed that the strength of these four-body interactions, E_4, is proportional to the sum of two corresponding pairwise terms,

$$E_4=\sum\ [(\epsilon_{ij}+\epsilon_{i+k,j+n})C_{ij}C_{i+k,j+n}] \quad \text{with} \ |k|=|n|. \tag{1}$$

In the above formula, ϵ_{ij} is the pair interaction term and C_{ij} is equal to 1(0) when residues i and j are (not) in contact.

As in our recent work,[41] here, we consider three tertiary interaction models. In model I, we only account for pairwise interactions, assuming E_4 equals zero. In model II, $k=\pm3$ and $k=\pm4$ correlations are included. Model III further incorporates $k=\pm1$ correlations. To retain the same balance between short- and long-range interaction in all three models, the pairwise parameters for models II and III are appropriately scaled down.

III. RESULTS AND DISCUSSION

According to the commonly accepted view of protein folding thermodynamics,[1] a reasonable potential must have its free energy minimum as well as the conformational energy minimum in the native state. To confirm that this is indeed the case, we employed ESMC for the *B* domain of protein A and the designed β protein to establish the relationship between energy and conformation. These studies also permitted a more complete analysis of the folding thermodynamics. Unfortunately, due to the prohibitive computational cost of such simulations for the somewhat longer protein *G* domain, only isothermal equilibrium Monte Carlo simulations near the transition temperature have been performed.

To explore the mechanism(s) of topology assembly for each molecule, isothermal dynamic Monte Carlo folding experiments were performed. To observe as many folding events as possible, the majority of simulations were performed near the folding transition temperature, where the assembly process is the fastest.[61] Below the folding temperature, the system could be easily trapped in metastable states. At higher temperatures, the folding intermediates and even more fully assembled structures are unstable; thus, they very rarely occur.

A. Folding of the *B* domain of protein A

1. Folding thermodynamics

An exact description of the protein energy landscape and the folding thermodynamics can be obtained from entropy sampling Monte Carlo simulations. Here, all three tertiary interaction models of the *B* domain of protein A were studied. Previously, for the designed Greek key β-barrel protein, we found that the all-or-none folding transition was obtained only when explicit cooperative terms of tertiary interactions were implemented (as in models II and III). Otherwise, the folding transition was continuous. Complementary simulations for protein A led to a different result. While the cooperativity of folding increases from model I to model III, the transition is still continuous, even for model III. To the best of our knowledge, we are unaware of any calorimetric studies of the thermal denaturation of protein A. However, Gdn HCl denaturation studies of protein A and its mutants are not inconsistent with a continuous transition.[62,63] These studies indicated that the different helices of protein A unfold at different denaturing agent concentrations. While suggestive, this cannot be considered proof of a continuous thermal transition. However, it is entirely possible that the continuous transition seen in our simulations could simply reflect the too low cooperativity of the model force field. A major cause is perhaps the unphysically large amount of helix content, $\sim$50%, in the denatured state. Indeed, recent NMR studies by Wright *et al.*,[64] which show the lack of a stable folding intermediate and very rapid fold assembly, may indicate two-state folding thermodynamics.

Regardless of the above ambiguity about the extent of folding cooperativity, the crucial requirement that the native-like state have the lowest conformational energy has been confirmed by the ESMC simulations. Furthermore, with decreasing conformational energy, the conformations of the model chain become closer to the folded state, and at very low energy, all conformations are characterized by a $C-\alpha$ rms deviation from native of about 3 Å. By way of example, in Fig. 1, the rms deviation of the $C-\alpha$'s from native is plotted against the conformational energy for model III of the tertiary interactions. Models I and II show essentially the same dependence of the rms on the energy.

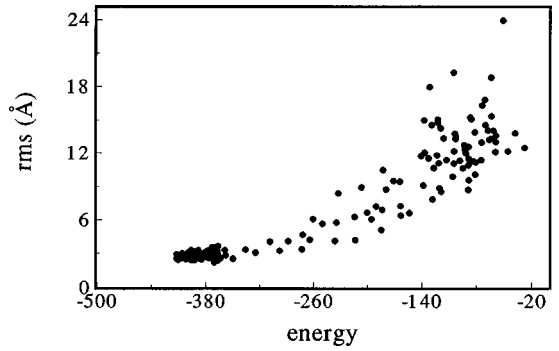

FIG. 1. The coordinate root-mean-square deviation (in Å for the α-carbon atoms) from the native conformation of the protein A fragment as a function of conformational energy for model III (see the text) of the tertiary interactions. These data have been extracted from the short final iteration of the entropy sampling Monte Carlo procedure.

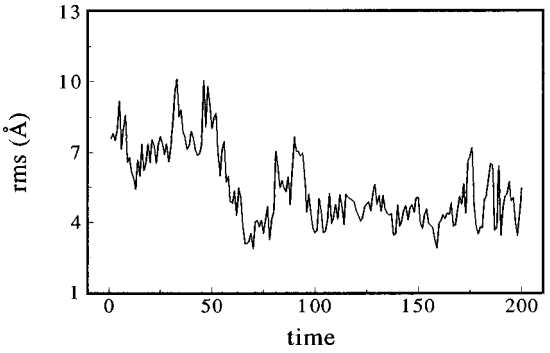

FIG. 3. The plot of coordinate root-mean-square deviation of the model $C-\alpha$ trace of protein A fragment from the native (in Å) as a function of simulation time at $T=1.9$ for model I of the tertiary interactions. The time unit corresponds to 3000 MC simulation cycles.

Figure 2 presents representative Molscript[65] snapshots of the model chain for various values of the energy. While the *N*-terminal helix of protein A could be observed at higher energies, over a somewhat lower energy range, the dominant partly folded structure is a *C*-terminal helical hairpin. Consequently, while the *N*-terminal helix appears to be most stable in isolation, tertiary interactions seem to favor formation of a *C*-terminal hairpin. An intact *N*-terminal hairpin is observed less frequently. Recent experiments suggest marginally greater stability for the *C*-terminal (helix II and helix III) hairpin.[64]

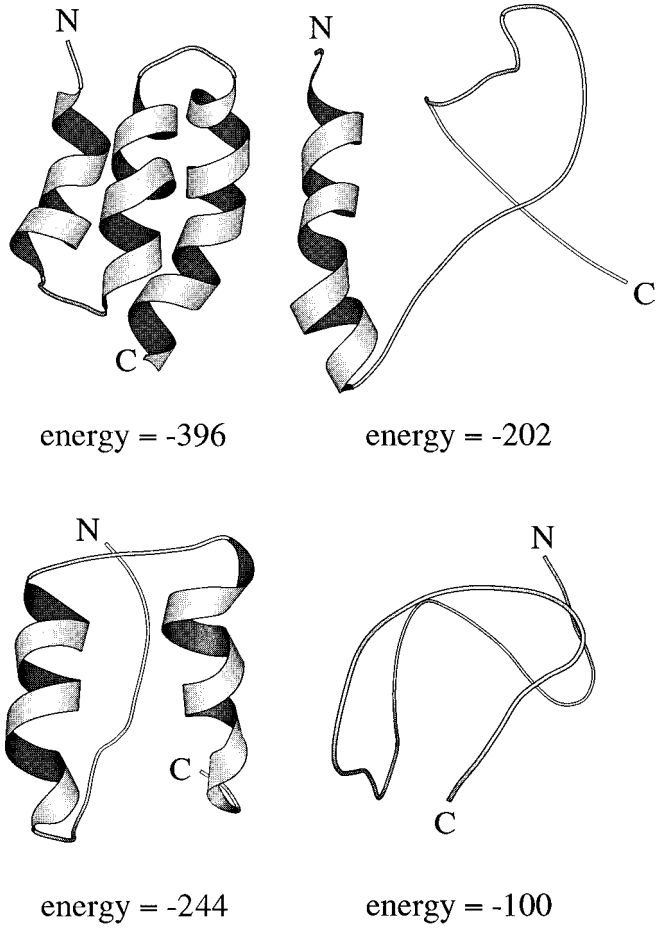

FIG. 2. Snapshots of representative conformations of the *B* domain of protein A model in the context of model III which were extracted from the short final iteration of the entropy sampling Monte Carlo procedure.

2. Mechanism of topology assembly

Experimentally, the *B* domain of protein A (residues 9–55) adopts a three-helix bundle topology. For all three of the tertiary interaction models, several long simulations were performed near the folding transition temperature or, more precisely, at the temperature where the model system samples mostly compact states. By way of illustration, we present representative results from model I, which lacks cooperative side chain packing terms. For this model, the folding temperature was slightly below a reduced temperature $T=1.9$. That is, all the energy terms are divided by this number to give the energy in kT units at the temperature of interest. In all simulations, the number of Monte Carlo time steps is equal to 600 000, and in all cases, 200 snapshots were taken every 3000 steps. In all the figures presented below, the time unit corresponds to 3000 MC steps.

At the transition temperature, the structure is not fixed, and the square radius of gyration oscillates between a nativelike value of about 55 and a value of 100, which is typical of very open conformations. Such conformational changes are very frequent, but most of the time the volume is close to that found in the native state. To determine if this really means that only nativelike compact conformations occur, in Fig. 3, we plot the rms (coordinate root-mean-square deviation from the native state for α-carbon atoms) as a function of simulation time. Clearly, some compact conformations are nativelike, such as at $t=70$ and 159, while others are not. Between $t=20$ and $t=40$, the structure is relatively compact, but the rms is large, ranging between 7 and 10 Å. In this range, the protein is by all measures unfolded. Again, the rms deviation changes rapidly, but not as frequently as the radius of gyration does.

How do the individual structural elements of protein A behave? In Fig. 4, the rms deviation from native is plotted for each of the three putative helical fragments. The *N*-terminal helix appears to be the most stable in isolation, and its assembly appears to be more cooperative than the remaining two helices. In the first 50 time steps, this helix is

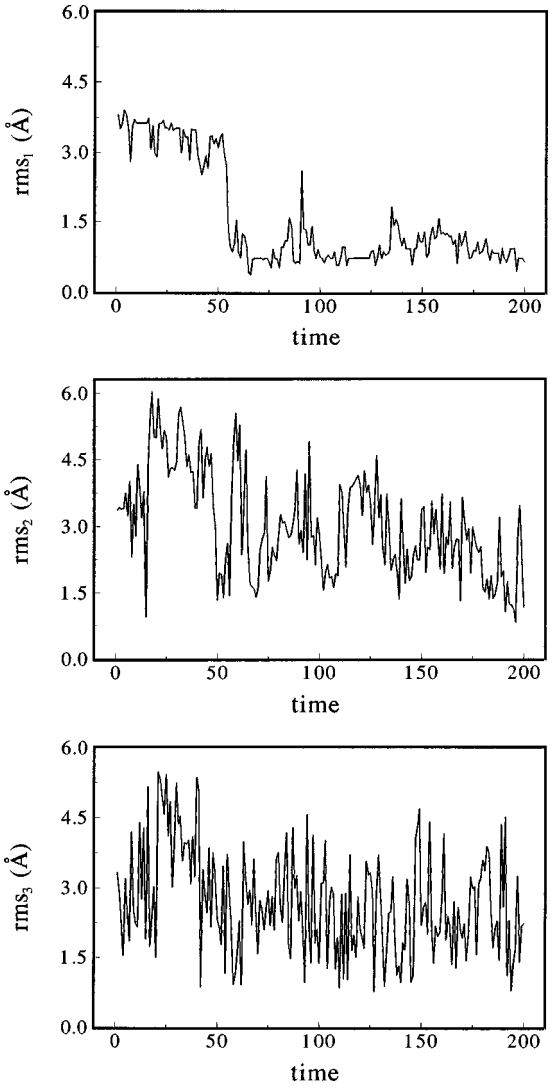

FIG. 4. Plots of coordinate root-mean-square deviation from native (in Å) for the three putative helical regions of the protein A sequence as a function of simulation time (see the caption to Fig. 3) at $T=1.9$ for model I of the tertiary interactions. (A) The first helix, residues 10–19. (B) The second helix, residues 25–37. (C) The third helix, residues 42–55.

loosely defined, with a rms from native oscillating around 3.5 Å. During the remainder of this simulation, this helix is very well defined, with a rms of about 1.0 Å, punctuated by occasional fluctuations to 2.0–2.5 Å. The second helix forms and dissolves many times during the simulations; nevertheless, a loosely defined helical conformation dominates. The third helix is even more mobile than the second. Its structure changes with the highest frequency, but, on average, it is more frequently close to the native conformation. Of course, the minimum rms for the entire structure coincides with the low rms values of all the helices. The time regions when the structural fluctuations of the second and the third helices are largest coincide with those when the first helix is poorly defined. This coincidence, when combined with a visual inspection of the chain conformations along the simulation trajectory, indicates that the first helix serves to structurally ''lock'' the other helices into place.

Sometimes structure assembly proceeds sequentially by starting from the first helix, then a helical hairpin involving the central helix forms. This is followed by docking of the *C*-terminal helix. Clearly, this mechanism cannot be described as on-site assembly. The two less stable helices often form in isolation and frequently collide to form the *C*-terminal helical hairpin. Some collisions are completely ineffective. Others lead to a substantial rearrangement of one or two helices. However, on-site assembly events also occur. These result from a collision between a better defined helical fragment and one with a loosely defined helix. Then, the helical fragments sequentially zip up. Thus, these simulation experiments reveal features of both diffusion-collision and on-site assembly mechanisms. However, the former mechanism seems to dominate.

Tertiary interaction models II and model III result in a similar picture of topology assembly. There is, however, one interesting difference seen relative to model I. At the transition temperature, the helices, especially the central one, seem be slightly more stable in isolation. Consequently, more cooperative tertiary interactions seem to favor the diffusion–collision model.[21] At first glance, this is a rather unexpected result because the cooperative terms only contribute to the tertiary interactions between the secondary structure elements (helices in this case). However, at the transition temperature, the contribution of the tertiary range interactions to the total energy of all conformations different from a three-helix structure is less than that in the model without cooperative terms. As a result, the short-range interactions are more important in the unfolded state, and thereby act to slightly increase the stability of the individual helices. Nevertheless, for all three tertiary interaction schemes, various modes of structure assembly were observed.

The assembly mechanisms described above serve to illustrate a technical problem associated with structure prediction from Monte Carlo simulations. Just below the collapse temperature, the model system frequently samples near-native states; however, these structures have no time to reach deep local (and hopefully global) conformational energy minima. (This point is even more apparent for the Greek key β-barrel, see below.) However, at lower temperatures, the escape time from the local minima becomes too long, and such structures become kinetically trapped. Thus, perhaps a different conformational sampling strategy should be employed, where lower energy states from a high temperature run (some of them presumably having a nativelike topology) are subject to lower temperature simulations. The subsequent selection of the lowest energy states should lead to well-defined native folds.

B. Folding thermodynamics and kinetics of a model Greek-key, β barrel

1. Folding thermodynamics

The sequence used in these simulations was previously designed on a computer to fold to the six-stranded, Greek-key β-barrel topology.[41,46] It is not obvious if this model sequence would really fold in nature, and the exaggerated design could well be unphysical. Previous ESMC studies[48] indicate that the conformational transition in model I was continuous, but becomes all-or-none in models II and III.[41] All three models produce the same native state, but differ in

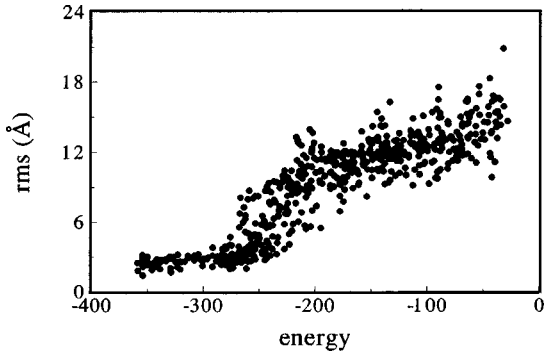

FIG. 5. The coordinate root-mean-square deviation (in Å for α-carbon atoms) from the native conformation of the Greek-key, designed protein as a function of conformational energy for model III (see the text) of the tertiary interactions. These data have been extracted from the short final iteration of the entropy sampling Monte Carlo procedure.

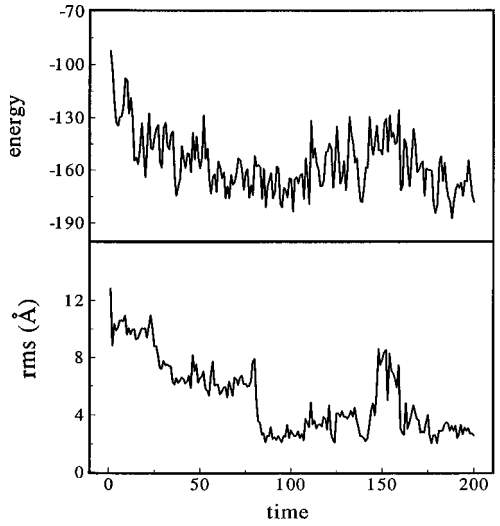

FIG. 6. The plot of conformational energy (in dimensionless kT units, upper panel) and coordinate root-mean-square deviation from the lowest energy nativelike conformation (in Å, lower panel) of the model $C-\alpha$ trace of the Greek-key barrel designed sequence as a function of simulation time (see the caption to Fig. 5) at $T=2.0$ for model III of the tertiary interactions. The time unit corresponds to 3000 simulation cycles.

the free energy barrier between the random coil and native states. For this designed sequence (perhaps due to its somewhat exaggerated design), the rms versus energy relationship was slightly stronger than that seen for the protein A sequence, i.e., the ''native'' state was better defined according to the force field of the model. This is clearly evidenced in Fig. 5 by the plot of rms versus conformational energy obtained via ESMC sampling for model III.

2. Mechanism of topology assembly

In the interest of brevity, we limit ourselves to presenting the analysis for the most cooperative model, model III. The isothermal simulations discussed below were done at $T=2.0$, which is somewhat below the folding transition temperature of 2.13. The observed fluctuations in the size of the globule range from nativelike values to volumes that are roughly twice as large as native. The conformational energy and the rms from the average folded ''native'' structure as a function of simulation time are shown in Fig. 6. The high rms regions correspond to various β-barrel-like structures having the wrong topology. The lowest energy states observed in this simulation exhibited an almost perfect nativelike Greek-key, barrel structure.

Figures 7(a)–7(c) show the rms versus time for the putative N-terminal two β strands, the central pair of β strands, and the C-terminal pair of β strands, respectively. As shown in Fig. 7(b), putative strands three and four spend most of the simulation near their native conformation, as do the first two N-terminal strands. Around $t=80$, the latter dissolve [Fig. 7(a)] for a very short period of time, and subsequently, the entire chain adopts a nativelike conformation whose rms is about 2.5 Å from the properly folded structure. Unfolding and refolding between $t=145$ and $t=160$ is accompanied by unwrapping the entire structure with the exception of the two Phe-rich strands three and four [see Fig. 7(b)], which remain intact. Here, topology assembly clearly proceeds via a sequential on-site assembly mechanism. For all simulations of this model Greek-key structure, on-site, sequential assembly strongly dominates; however, fusion of the two loosely defined hairpins involving strands one and two with strands three and four was only very rarely observed and is generally not successful.

It is interesting to compare the energy flow-chart obtained in the isothermal folding simulations with our finding from ESMC studies[41] of this model system. The lowest energy states observed in folding simulations correspond to correctly assembled structures, yet they are several tens of kT units higher than the lowest energy ''native'' conformations found in the ESMC studies.[41] The observed topology assembly occurs mainly in the region of high energy states and is on the high energy side of the conformational free energy barrier describing the thermodynamics of the conformational transition. Consequently, the large decrease of conformational energy below the transition point has to be associated with fine-tuning of the structure accompanying ''side chain fixation.'' [4–6] This process is found here to be extremely slow[6] and is not seen in any of the isothermal folding simulations. Sometimes the ''nativelike'' state could be obtained from a simulated annealing protocol; however, no simple test would prove that the obtained conformation really corresponds to the lowest energy region. As a result, many simulated annealing experiments must be performed and carefully analyzed in order to identify the native state. Finding the lowest energy state in an isothermal simulation is very difficult, and for more complex systems, as a practical matter, impossible. At temperatures above the folding transition, the native state has low thermodynamic probability, at lower temperatures the process becomes extremely slow. Thus, alternative conformational sampling protocols are clearly required.

C. Simulations of the *B*1 domain of protein G

1. Folding thermodynamics

The $B1$ domain of protein G has a very regular $\beta\beta\alpha\beta\beta$ structure[47] with a central helix on the top β sheet and the N-

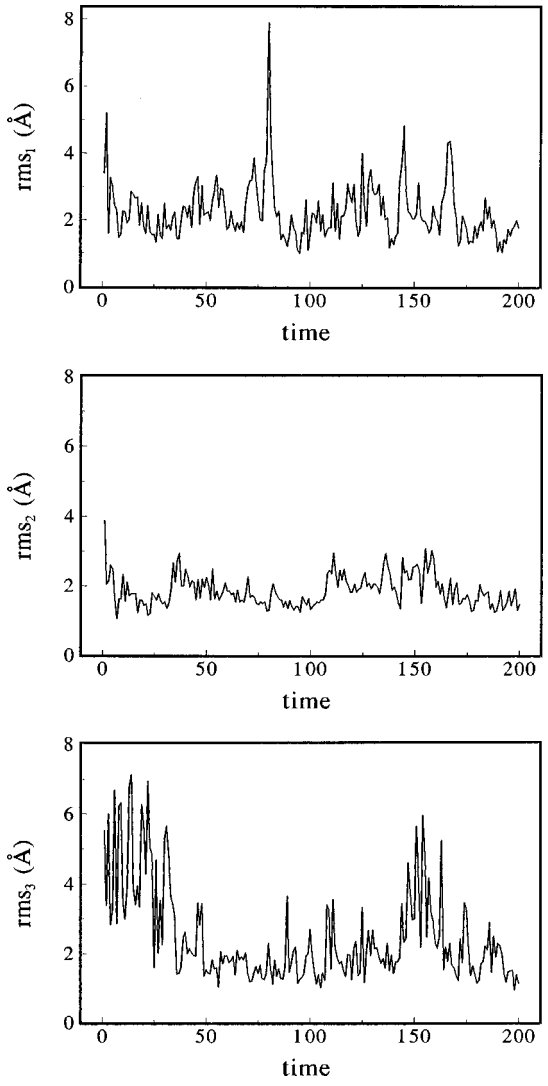

FIG. 7. Plots of coordinate root-mean-square deviation of three fragments of Greek-key barrel designed sequence from the lowest energy nativelike conformation (in Å) as a function of simulation time (see the caption to Fig. 5) at $T=2.0$ for model III of tertiary interactions. (A) N terminal, two β strands, (B) central, two β strands, (C) C terminal, two β strands.

and C-termini (parallel) strands located in the center of the β sheet. Because of computational cost, ESMC calculations were not possible for this molecule. Thus, we restricted ourselves to standard dynamic MMC sampling. In many isothermal simulations (for all three models of tertiary interactions), a loosely defined fold of protein G with a rms from native in the range of 5 Å was observed many times. The structural errors of these topologically correct folds were mostly associated with the helix packing with a wrong angle on top of the β sheet. Sometimes, a mirror image topology of the protein G fold was observed. Simulated annealing procedures for protein G were nonreproducible. The native fold was obtained in only one of 18 (relatively rapid annealing) experiments.

2. Mechanism of assembly

To illustrate typical behavior seen for protein G, the results from isothermal simulations of model III at $T=1.75$ are presented and discussed. The square radius of gyration of the

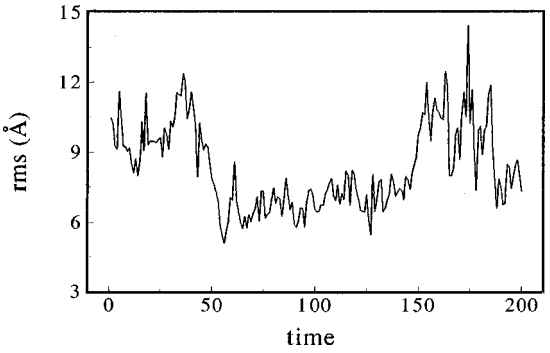

FIG. 8. The plot of coordinate root-mean-square deviation of the model $C-\alpha$ trace of the $B1$ domain of protein G from the lowest energy nativelike conformation (in Å) as a function of simulation time at $T=1.75$ for model III of the tertiary interactions. The time unit corresponds to 3000 MC simulation cycles.

polypeptide chain mostly oscillates between 70 and 100. The former is that expected for the native state. Occasionally, more open conformations are observed. As shown in Fig. 8, the observed rms from native samples a broad range of values, from 12 to 14 Å, characteristic of a compact random state, to about 5 Å, characteristic of a low resolution native topology. As before, we divide the protein into fragments corresponding to possible structural elements. These consist of the putative N-terminal β hairpin, helix, and C-terminal β hairpin. Even a brief inspection of the rms versus time flow charts given in Fig. 9 shows that the central helix assembles most rapidly and is the most stable very early folding intermediate. This is consistent with experimental studies of Serrano and co-workers.[66] Its rms from the native conformation oscillates between 0.7 and 4.5 Å, with the average rms around 2.0 Å. The C-terminal hairpin spends about a third of the simulation time around the native conformation with rms values oscillating between 1.8 and 4 Å. During the remaining time, this fragment has various random conformations with rms from native up to 10 Å.

It is interesting to note that the part of protein G structure consisting of the helix and the C-terminal β hairpin assembled into a nativelike conformation (with a rms deviation between 2 and 3 Å) many times during the simulation, and the mechanism of assembly had a predominantly sequential (on-site) character. The C-terminal β hairpin almost always remained intact when it was in contact with the helix. The behavior of the N-terminal β-hairpin fragment was more random. In only a few snapshots was the rms from native below 4 Å. Similar to the cluster formed by the C terminus, this corresponded to a hairpin helix motif close to the native one. The N-terminal fragment of the structure was the most mobile. Occasionally, the native protein G topology formed, but with large structural defects. For example, it contained a crumpled, second β-strand fragment that lacks most of the hydrogen bonding with the rest of the β sheet, or the β-sheet formed from incorrectly registered strands, with distorted helix to sheet packing. The lowest observed rms deviation from the native structure in this simulation was about 5 Å, and the lowest energy conformation had a strongly distorted protein G topology and a rms deviation from native of 7 Å.

The assembly mechanism of protein G is dominated by

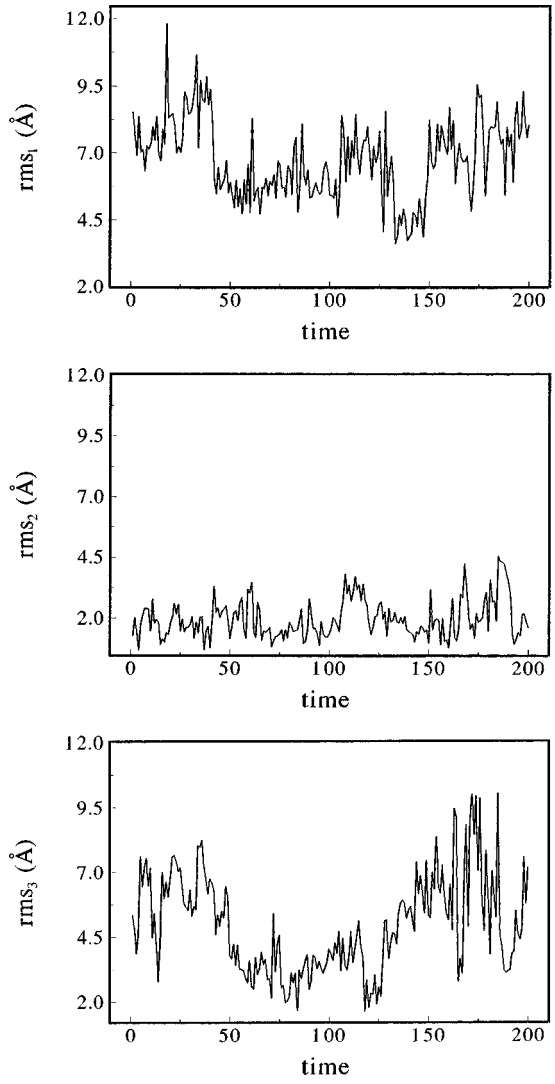

FIG. 9. Plots of coordinate root-mean-square deviation from native (in Å) for three structural fragments of the $B1$ domain of protein G sequence as a function of simulation time at $T=1.75$ for model III of tertiary interactions. (A) N terminal β hairpin, residues 2–20, (B) the central helix, residues 24–36, (C) C terminal β hairpin, residues 40–55.

the sequential ''on-site'' building of β hairpins on the helical fragment, which, in the context of our reduced model, seems be the most stable structural element of protein G. The motif consisting of the helix and C-terminal β hairpin was relatively easy to assemble and was quite stable. Interestingly, experimentally, the C-terminal hairpin fragment of protein G is stable, whereas the N-terminal hairpin is not.[66] In such situations, the completion of the fold requires a proper zipping up of the N-terminal β strand or collision with a ''prefabricated'' N-terminal hairpin. The latter is unlikely due to the very low stability of this hairpin, and the former is difficult to achieve from an entropic viewpoint. These are probably the reasons why the native protein G structure is so rarely assembled in this simulation. The question remains as to what extent such structure assembly bottlenecks are experienced by the real proteins as opposed to their just being artifacts of our reduced model and conformational sampling scheme. One simple explanation of the present results is that our sampling times are too short. However, it may also turn out that the balance between isolated fragment stability and global stability in these models is incorrectly shifted toward the former.

IV. CONCLUSIONS

Using a reduced model of protein conformation and a knowledge-based potential, we have investigated the aspects of protein folding thermodynamics and possible mechanisms of protein topology assembly for three small proteins. Previously, we have shown that this reduced model is capable of finding the native structure of very simple, small globular proteins.[39,58,59,67–73] The inclusion of cooperative side chain packing terms enabled side chain fixation to occur, but did not require it.[39,58,59,67] Interestingly, using ESMC, this class of terms in a computer-designed β protein was shown to produce an all-or-none conformational transition. However, here, also using ESMC, in the case of protein A, such terms proved insufficient to yield a two-state model. Whether this correctly mimics the actual folding thermodynamics is uncertain. However, the protein A simulations clearly suggest that two-state thermodynamic behavior results from the interplay of intrinsic secondary structure stability and tertiary interactions. If the former is too strong, then as *Go* suggested, the conformational transition is continuous. It is only when the native conformation is sufficiently stable relative to the collection of misfolded structures does an all-or-none transition emerge. This study clearly suggests that the relative balance of interactions in this class of models needs reexamination; such studies are currently underway.

Another objective of this paper was to examine the early, relatively fast folding events associated with topology assembly. For this purpose, we simulated long-time dynamics at or near the transition midpoint. Our results can be summarized as follows:

(i) For all motifs, features of the diffusion–collision mechanism and the sequential ''on-site'' mechanism of assembly are observed.
(ii) Cooperative side chain packing interactions slightly bias the assembly process toward ''prefabricated'' assembly. It should be pointed out, however, that the effect is small.
(iii) In the all α protein, the diffusion–collision model of assembly dominates, while the β protein folds predominantly via an on-site sequential mechanism.
(iv) The differential relative stabilities of various protein fragments seem to create kinetic barriers for protein topology assembly.

The above findings are not inconsistent with known experimental facts.[2,23,27,72,74] However, apart from a few cases, it has to be stressed that rather little is known about the specific sequence(s) of events that lead to the assembly of protein topologies. This is because the early folding events are very fast, and at equilibrium, the population of intermediates is very low. Apart from being of fundamental interest, understanding of the topology assembly process would also have practical consequences, such as the design of more efficient conformational search protocols that could enable a broader class of molecules to be folded. This is now under-

way. Finally, recent progress in experimental techniques that probe relatively early folding events[64,75] will yield insights into some aspects of the mechanism of protein topology assembly.

ACKNOWLEDGMENTS

This work was partially supported by NIH Grant No. GM-37408 and by University of Warsaw Grant No. BST-34/97. A.K. is an International Scholar of the Howard Hughes Medical Institute (Grant. No. 75195-543402). W.G. acknowledges support by KBN Grant No. 6PO4A02510.

[1] C. B. Anfinsen, Science **181**, 223 (1973).
[2] C. B. Anfinsen and H. A. Scheraga, Adv. Prot. Chem. **29**, 205 (1975).
[3] T. E. Creighton, Biochem. J. **270**, 131 (1990).
[4] K. Kuwajima, Proteins **6**, 87 (1989).
[5] K. Kuwajima, M. Mitani, and S. Sugai, J. Mol. Biol. **206**, 547 (1989).
[6] O. B. Ptitsyn, R. H. Pain, G. V. Semisotnov, E. Zerovnik, and O. I. Razgulyaev, FEBS Lett. **262**, 20 (1990).
[7] O. B. Ptitsyn, Curr. Opinion Struct. Biol **5**, 74 (1995).
[8] D. Eliezer, P. A. Jennings, P. E. Wright, S. Doniach, K. O. Hodgson, and H. Tsuruta, Science **270**, 487 (1995).
[9] M. Karplus and E. Shakhnovich, in *Protein Folding*, edited by T. E. Creighton (Freeman, New York, 1992), pp. 127–196.
[10] M. Karplus and A. Sali, Curr. Opinion Struct. Biol. **5**, 58 (1995).
[11] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[12] J. Skolnick, A. Kolinski, and R. Yaris, Proc. Natl. Acad. Sci. USA **85**, 5057 (1988).
[13] J. Skolnick and A. Kolinski, Annu. Rev. Phys. Chem. **40**, 207 (1989).
[14] J. Skolnick, A. Kolinski, and R. Yaris, Biopolymers **28**, 1059 (1989).
[15] J. Skolnick, A. Kolinski, and R. Yaris, Proc. Natl. Acad. Sci. USA **86**, 1229 (1989).
[16] J. Skolnick and A. Kolinski, J. Mol. Biol. **212**, 787 (1990).
[17] J. Skolnick and A. Kolinski, Science **250**, 1121 (1990).
[18] J. Skolnick, A. Kolinski, and A. Sikorski, Chem. Design Automation News **5**, 1 (1990).
[19] K. A. Dill and K. Yue, Prot. Sci. **5**, 254 (1996).
[20] A. R. Dinner, A. Sali, and M. Karplus, Proc. Natl. Acad. Sci. USA **93**, 8356 (1996).
[21] M. Karplus and D. L. Weaver, Biopolymers **18**, 1421 (1979).
[22] P. S. Kim and R. L. Baldwin, Annu. Rev. Biochem. **51**, 459 (1982).
[23] P. S. Kim and R. L. Baldwin, Annu. Rev. Biochem. **59**, 631 (1990).
[24] H. J. Dyson, G. Merutka, J. P. Waltho, R. A. Lerner, and P. E. Wright, J. Mol. Biol. **226**, 795 (1992).
[25] P. N. Lewis, N. Go, M. Go, D. Kotelchuck, and H. A. Scheraga, Proc. Natl. Acad. Sci. USA **65**, 810 (1970).
[26] H. J. Dyson, J. R. Sayre, H.-C. Shin, G. S. Merutka, R. A. Lerner, and P. E. Wright, J. Mol. Biol. **226**, 819 (1992).
[27] P. E. Wright, H. J. Dyson, and R. A. Lerner, Biochemistry **27**, 7167 (1988).
[28] J. H. Dyson and P. E. Wright, Curr. Biol. **3**, 60 (1993).
[29] E. W. Knapp, J. Comput. Chem. **13**, 793 (1992).
[30] M. Levitt, J. Mol. Biol. **104**, 59 (1975).
[31] C. Wilson and S. Doniach, Proteins **6**, 193 (1989).
[32] A. T. Hagler and B. Honig, Proc. Natl. Acad. Sci. USA **75**, 554 (1978).
[33] S. Sun, Protein Sci. **2**, 762 (1993).
[34] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).
[35] E. Shakhnovich, G. Farztdinov, and A. M. Gutin, Phys. Rev. Lett. **67**, 1665 (1991).
[36] A. Sikorski and J. Skolnick, J. Mol. Biol. **212**, 819 (1990).
[37] A. Sikorski and J. Skolnick, Proc. Natl. Acad. Sci. USA **86**, 2668 (1989).
[38] J. Skolnick and A. Kolinski, J. Mol. Biol. **221**, 499 (1991).
[39] A. Kolinski and J. Skolnick, *Lattice Models of Protein Folding, Dynamics and Thermodynamics* (Landes, Austin, 1996).
[40] A. Rey and J. Skolnick, Chem. Phys. **158**, 199 (1991).
[41] A. Kolinski, W. Galazka, and J. Skolnick, Proteins **26**, 271 (1996).
[42] N. Go and H. Taketomi, Proc. Natl. Acad. Sci. USA **75**, 559 (1978).
[43] N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).
[44] P. L. Privalov and S. J. Gill, Adv. Protein Chem. **39**, 191 (1988).
[45] H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata, and I. Shimada, Biochemistry **40**, 9665 (1992).
[46] A. Kolinski, W. Galazka, and J. Skolnick, J. Chem. Phys. **103**, 10286 (1995).
[47] A. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore, Science **253**, 657 (1991).
[48] M.-H. Hao and H. A. Scheraga, J. Phys. Chem. **98**, 4940 (1994).
[49] M.-H. Hao and H. A. Scheraga, J. Phys. Chem. **98**, 9882 (1994).
[50] M.-H. Hao and H. A. Scheraga, J. Chem. Phys. **102**, 1334 (1995).
[51] A. Kolinski, M. Milik, J. Rycombel, and J. Skolnick, J. Chem. Phys. **103**, 4312 (1995).
[52] J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik, Prot. Sci. **6**, 676 (1997).
[53] F. C. Bernstein *et al.*, J. Mol. Biol. **112**, 535 (1977).
[54] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **51**, 1087 (1953).
[55] A. Rey and J. Skolnick, Proteins **16**, 8 (1993).
[56] H. A. Scheraga, M.-H. Hao, and J. Kostrowicki, in *Methods in Protein Structure Analysis*, edited by M. Z. Atassi and E. Appela (Plenum, New York, 1995).
[57] A. Kolinski and J. Skolnick, Parameters of Statistical Potentials. Available via ftp from public directory, scripps.edu(pub/MCSP) (1997).
[58] A. Kolinski, A. Godzik, and J. Skolnick, J. Chem. Phys. **98**, 7420 (1993).
[59] A. Kolinski and J. Skolnick, Proteins **18**, 338 (1994).
[60] A. Godzik, J. Skolnick, and A. Kolinski, Protein Eng. **6**, 801 (1993).
[61] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).
[62] S. P. Bottomley, A. G. Popplewell, M. Scawen, T. Wan, B. J. Scutton, and M. G. Gore, Protein Eng. **7**, 1463 (1994).
[63] L. Cedergren, R. Andersson, B. Jansson, M. Uhlen, and B. Nilsson, Protein Eng. **6**, 441 (1993).
[64] Y. Bai, A. Karmi, H. J. Dyson, and P. E. Wright, Protein Sci. **6**, 1449 (1997).
[65] P. J. Kraulis, J. Appl. Crystallogr. **24**, 946 (1991).
[66] F. J. Blanco, A. R. Ortiz, and L. Serrano, Folding Design **2**, 123 (1997).
[67] A. Kolinski and J. Skolnick, Proteins **18**, 353 (1994).
[68] K. A. Olszewski, A. Kolinski, and J. Skolnick, Protein Eng. **9**, 5 (1996).
[69] K. A. Olszewski, A. Kolinski, and J. Skolnick, Proteins **25**, 286 (1996).
[70] M. Vieth, A. Kolinski, C. L. Brooks III, and J. Skolnick, J. Mol. Biol. **237**, 361 (1994).
[71] M. Vieth, A. Kolinski, C. L. Brooks III, and J. Skolnick, J. Mol. Biol. **251**, 448 (1995).
[72] J. Skolnick, A. Kolinski, C. Brooks III, A. Godzik, and A. Rey, Curr. Biol. **3**, 414 (1993).
[73] J. Skolnick and A. Kolinski, in *Computer Simulations of Biomolecular Systems. Theoretical and Experimental Studies*, edited by W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson (ESCOM Science Publ., 1997).
[74] R. L. Baldwin and H. Roder, Curr. Biol. **1**, 219 (1991).
[75] R. L. Baldwin, J. Biomol. NMR **5**, 103 (1995).