Dynamics and Thermodynamics of β-Hairpin Assembly: Insights from Various Simulation Techniques

Andrzej Kolinski,*# Bartosz Ilkowski,* and Jeffrey Skolnick#

*Department of Chemistry, University of Warsaw, 02-093 Warsaw, Poland, and #Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037 USA

ABSTRACT Small peptides that might have some features of globular proteins can provide important insights into the protein folding problem. Two simulation methods, Monte Carlo Dynamics (MCD), based on the Metropolis sampling scheme, and Entropy Sampling Monte Carlo (ESMC), were applied in a study of a high-resolution lattice model of the C-terminal fragment of the B1 domain of protein G. The results provide a detailed description of folding dynamics and thermodynamics and agree with recent experimental findings (Munoz et al., 1997. *Nature.* 390:196–197). In particular, it was found that the folding is cooperative and has features of an all-or-none transition. Hairpin assembly is usually initiated by turn formation; however, hydrophobic collapse, followed by the system rearrangement, was also observed. The denatured state exhibits a substantial amount of fluctuating helical conformations, despite the strong β-type secondary structure propensities encoded in the sequence.

INTRODUCTION

The folding process of single-domain globular proteins is usually very cooperative, with a small population of intermediate states (Ptitsyn, 1995) at the transition temperature (Creighton, 1993). Such an all-or-none transition has many features of a first-order phase transition. Because intermediates are sparsely populated, much less is known about the mechanism of assembly. A number of experiments, simulations (Karplus and Sali, 1995), and theoretical considerations (Friesner and Gunn, 1996) indicate that hydrophobic collapse from a random coil state (with a small amount of fluctuating secondary structure) to a dense globular state with a significant secondary structure content may be the first well-defined stage of the folding process. This so-called molten globule state has a significant fraction of native secondary structure, a volume larger than the volume of the native state, and a poorly defined pattern of tertiary interactions (Ptitsyn, 1995). Subsequently, a slow collective rearrangement of the molten globular state leads to the native structure.

Because of its complexity, studies of the folding mechanism of globular proteins are very difficult (Fersht, 1993; Baldwin, 1995); thus, investigators tend to study smaller model systems, which can be better controlled and very useful for elucidation of the most fundamental aspects of protein folding (Blanco et al., 1994; Blanco and Serrano, 1995; Dyson and Wright, 1993). It is important, however, to establish the extent to which the folding dynamics and thermodynamics of these model systems resemble that of globular proteins.

Recently, in an important study, Munoz, Thompson, Hofrichter, and Eaton (Munoz et al., 1997) published the results of experimental studies on the folding of the C-terminal fragment (residues 41–56) of the B1 domain of protein G. The B1 domain of protein G in its native state adopts a very stable structure with high regular secondary structure content (Gronenborn et al., 1991), in which the C-terminal fragment is a β-hairpin. This fragment, when excised from the entire sequence, shows a significant population of β-type structure. Munoz and co-workers applied temperature-jump kinetic spectroscopy to study the folding process of this small system. In the native structure of protein G, the tryptophan at position 43 interacts with phenylalanine at position 52 and valine at position 54, providing an internal fluorescence probe for structure formation. An additional probe was introduced by adding dansylated lysine to the C-terminus, which allowed monitoring of the thermal unfolding/folding process.

They found a sharp increase in the β-hairpin population at a critical temperature. The folding process was significantly slower than the formation of a helix of comparable size. The results of these experiments can be explained within the framework of a very simple statistical mechanical model. The model assumed a significant degeneracy of the native basin of the free energy landscape, associated with structural fluctuations of the end residues. The calculated β-hairpin content changed from below 10% at 360K to above 80% at 280K. The free energies of partly folded structures calculated from the model were 2.5–4.5 kcal/mol higher than the free energy of folded or unfolded states. This indicates a rather cooperative, all-or-none transition. A similar simplified statistical model of protein folding dynamics and thermodynamics was previously proposed by Camacho and Thirumali (1996).

Received for publication 28 May 1999 and in final form 26 August 1999.

Address reprint requests to Dr. Jeffrey Skolnick, Laboratory of Computational Genomics, Danforth Plant Science Center, CET, 4041 Forest Park Avenue, St. Louis, MO 63108. Tel.: 314-615-6931; Fax: 314-615-6924; E-mail: skolnick@danforthcenter.org; or to Dr. Andrzej Kolinski, Department of Chemistry, University of Warsaw, 02-093 Warsaw, Poland. E-mail: kolinski@chem.uw.edu.pl.

0006-3495/99/12/2942/11 $2.00

Over the last few years, we have developed a series of discretized protein models (Kolinski and Skolnick, 1996). These models employ a high coordination lattice representation of the polypeptide chain and potentials of mean force derived from the statistical regularities seen in known protein structures. Here we employ a model that uses a lattice representation of the side-chain units and a single interaction center per amino acid residue. The model has previously been used for assembly of protein structure from sparse experimental data (Kolinski and Skolnick, 1998), modeling of protein secondary structure (Kolinski et al., 1997), distant homology modeling, and *ab initio* protein structure prediction (Kolinski et al., 1998). The applicability of the model in *ab initio* protein structure prediction was tested during the CASP3 prediction contest; a fraction (about one-third) of the query protein folds were qualitatively predicted. The details of the model are given in the Methods section. Interestingly, for this β-hairpin, the same results as reported below were obtained, using a different lattice model that has two interaction centers per residue. That model has previously been employed in various studies of protein structure prediction, dynamics, and thermodynamics, including studies of the first-order transition in model polypeptides (Kolinski et al., 1996).

METHODS

Protein model

The model employed here is very similar to that described previously (Kolinski and Skolnick, 1998). Small updates to the protein representation slightly increase the geometric fidelity of the model. For the reader's convenience, the design of the model is outlined below.

The model chain consists of a string of virtual bonds connecting the interaction centers that correspond to the center of mass of the side chains, including the α-carbons. All heavy atoms have the same weight in this averaging. Thus the center of glycine coincides with its C_α, the center of alanine is located in the middle of the C_α-C_β bond, the center of valine coincides with the C_β atom, etc. For the side chains that possess internal degrees of freedom, the interaction centers correspond to the center of mass of the actual rotamer. These interaction centers (beads) are projected onto an underlying cubic lattice with a lattice spacing of 1.45 Å. This constant defines the spatial resolution of the model. Obviously, the virtual bonds resulting from such a projection are of various lengths, depending on the identities of the two successive residues, the main chain conformation and the rotameric state of the side chain. A change in any of these variables may change the corresponding virtual bonds. In proteins, these distances have a quite broad distribution, ranging from 3.8 Å between a pair of glycines to ~10 Å for some pairs of large side chains in their antiparallel orientation and expanded conformations. The corresponding set of lattice vectors covers this distribution with good fidelity. The shortest vectors are of the form of (±2, ±2, ±1) or (±3, 0, 0) vectors, including all possible permutations of the coordinates. The length of these vectors corresponds to 4.35 Å. The longest lattice vectors are of the (±5, ±2, ±1) type, and their length corresponds to 7.94 Å; thus the wings of the distribution are cut off. This should not have any noticeable effect on the model's fidelity; the small distance cutoff error is well below the resolution of the model, and the long distance cutoff error is not important, because of the very rare occurrence of distances above 8 Å. Consequently, the set of the allowed lattice bonds consists of 646 vectors. For a technical reason, sequentially adjacent vectors must not be identical. A cluster of the excluded volume points is associated with each bead of the model chain. Each cluster consists of 19 lattice points: the central one; six points at the positions (±1, 0, 0), (0, ±1, 0), and (0, 0, ±1) with respect to the central one; and 12 points at the positions (±1, ±1, 0), including all permutations. Thus the closest approach positions of another cluster with respect to a given cluster are of the form of (±2, ±2, ±1) and (±3, 0, 0) vectors as measured between the cluster centers. It could easily be calculated that here are 30 positions of closest approach. The distance of the closest approach nicely corresponds to the smallest values of the interresidue distances in real proteins. Because the average "contact distances" (see the following sections) of the model residues are somewhat larger than the distance of the closest approach, there are many more than 30 spatial orientations of two residues in contact. Consequently, such a representation of protein structure entirely avoids various anisotropy effects typically seen in the lower resolution lattice protein models. With the above outlined geometric restrictions, all PDB structures (Bernstein et al., 1977) could be represented with an average root mean square deviation, RMSD, of ~0.8 Å. Again, the accuracy of the fit does not show any systematic dependence on protein length or the orientation of the crystallographic structure with respect to the lattice coordinate system.

Model of interactions

The model force field consists of several types of potentials. The first group has the form of generic biases that penalize against non-protein-like conformations. These potentials are sequence independent. Sequence-specific contributions to the force field consist of knowledge-based two-body and multibody potentials extracted from a statistical analysis of known protein structures.

The generic protein stiffness potential and secondary structure bias

The model chain as defined above is intrinsically very flexible. A substantial fraction of its conformations that are allowed because of the assumed simplified hard-core interactions do not correspond to any accessible real polypeptide chain conformation. In particular, proteins are relatively stiff polymers. Moreover, the folded state of proteins has

very characteristic distributions of certain short-range distances. For example, the bimodal distribution of the distances between the ith and $i + 4$th residues reflects the tendency to adopt either of two types of conformations. These correspond to expanded (β-type, or expanded coil) or very compact conformations (as within helices or turns). Such generic features have to be included in the model. We proceed in a similar fashion, as described elsewhere. The details are different, because of the refined protein representation (larger number of chain vectors allowed and modified position of the center of interaction, which now also includes α-carbons).

First, let us define for all possible two-vector sequences of the model chain a direction $\mathbf{w}$ that is almost perpendicular to the plane formed by the fragment. A small systematic deviation from the exactly orthogonal direction is introduced to obtain vectors $\mathbf{w}$ that are on average parallel to a helix axis and account for an average supertwist of β-strands:

$$\mathbf{u}_i = (\mathbf{v}_{i-1} \otimes \mathbf{v}_i - \mathbf{v}_{i-1} - \mathbf{v}_i) \tag{1}$$

$$\mathbf{w}_i = \mathbf{u}_i / |\mathbf{u}_i| \tag{2}$$

where $\mathbf{v}_i$ is the ith vector (or virtual bond) of the model chain, the symbol $\otimes$ denotes the vector cross-product, and $|\mathbf{u}_i|$ is the length of vector $\mathbf{u}_i$. Consequently, these "directions of secondary structure" (the vectors $\mathbf{w}$ point along a helix or across a β-sheet) are normalized so that their length is equal to 1.

The stiffness/secondary structure bias has the following form:

$$E_{stiff} = -\epsilon_{gen}\{\Sigma \max(0, \mathbf{w}_i \bullet \mathbf{w}_{i+4})\} \tag{3}$$

where ϵ_{gen} is a constant energy parameter, common for all generic potential, and Σ means summation along the chain for helical and β-expanded states. The above formulation means that the system is energetically stabilized when pairs of "direction of secondary structure" vectors are in a parallel orientation (positive dot product). The stabilization energy increases in the range between 0° and 90° (the angle between appropriate vectors $\mathbf{w}$). The minimum value of the stiffness function per residue is equal to $-0.625\epsilon_{gen}$, and the maximum is 0. For the studied system, it was assumed a priori that the secondary structure is known in a three-letter code. This constituted a small bias toward expanded states. Because the studied polypeptide has a very strong propensity toward β-type conformations, such a bias should have a marginal effect (if any) on the qualitative behavior of the model system. It should also be mentioned that the bias does not prohibit the formation of helical states, as is discussed later. The described model allows the *ab initio* folding of protein G without any knowledge of secondary structure; however, usage of predicted secondary structure (and even more, the assumption of known native secondary structure) increases the accuracy of the predicted native state as well as its reproducibility (unpublished work).

Furthermore, a bias has been introduced toward the specific geometry of helical and β-type expanded states (however, it is quite permissively defined). All conformations are, of course, allowed; the purpose of the bias is to mimic a protein-like (average) distribution of local conformations. Symbolically, this could be written as follows:

$$E_{struct} = \Sigma\{\delta H1(i) + \delta H2(i) + \delta E1(i) + \delta E2(i)\} \tag{4}$$

with

$$\delta H1(i) = \begin{cases} -\epsilon_{gen}, & \text{for } r^2_{i,1+4} < 36 \text{ and } (\mathbf{v}_i \bullet \mathbf{v}_{i+3}) > 0 \\ & \quad \text{and } (\mathbf{v}_i \bullet \mathbf{v}_{i+2}) < -5 \\ 0, & \text{otherwise} \end{cases} \tag{4a}$$

$$\delta H2(i) = \begin{cases} -\epsilon_{gen}, & \text{for } r^2_{i,1+4} < 36 \text{ and } (\mathbf{v}_i \bullet \mathbf{v}_{i+3}) > 0 \\ & \quad \text{and } (\mathbf{v}_{i+1} \bullet \mathbf{v}_{i+3}) < -5 \\ 0, & \text{otherwise} \end{cases} \tag{4b}$$

$$\delta E1(i) = \begin{cases} -\epsilon_{gen}, & \text{for } 56 < r^2_{i,1+4} < 135 \\ & \quad \text{and } (\mathbf{v}_i \bullet \mathbf{v}_{i+2}) > 5 \\ 0, & \text{otherwise} \end{cases} \tag{4c}$$

$$\delta E2(i) = \begin{cases} -\epsilon_{gen}, & \text{for } 56 > r^2_{i,i+4} < 135 \\ & \quad \text{and } (\mathbf{v}_{i+1} \bullet \mathbf{v}_{i+3}) > 5 \\ 0, & \text{otherwise} \end{cases} \tag{4c}$$

The numerical values are in the lattice units and are selected to define a broad range of helical/turn conformations (for the δH1 and δH2 contributions) or expanded conformations (for the δE1 and δE2 contributions). Because of the exclusive character of the two subsets of the above geometrical conditions for specific chain conformations, the minimum contribution from a residue is equal to $-2\epsilon_{gen}$ (either the first two conditions or the last two conditions could be satisfied simultaneously). Let us express the last condition a bit differently. Equation 4d says that the system gains energy equal to $-\epsilon_{gen}$ for being in an expanded β-type conformation. For a four-vector fragment of the chain, the distance between the ith and $i + 4$th chain beads (centers of mass of the side-chain $+C\alpha$ unit) must correspond to a range between 10.7 Å and 16.8 Å, and the chain vectors $\mathbf{v}_{i+1}$ and $\mathbf{v}_{i+3}$ have to be oriented in a parallel-like fashion (the dot product > 5). Additional stabilization is gained when, for the same fragment, another pair of vectors is parallel (Eq. 4c). The broad ranges allow for substantial fluctuations (without any energetic penalty) around an ideal expanded state and accommodate the variations of the model chain geometry caused by differences in the side-chain size. This kind of bias has been applied to the entire chain, regardless of the secondary structure prediction. Such predictions are never exact, so the model was designed to allow for the construction of regular secondary motifs in any location. Of course, the occurrence of the additional regular fragments is moderated in this model by the outlined short- and long-range interactions.

Generic packing cooperativity

We introduce two terms that enforce some of the most general regularities of the dense packing of protein structures (Godzik et al., 1993). In all of the more regular elements of secondary structure (within helices and β-sheets, but not between helices) and, to a lesser extent, in some coil-type fragments and turns, given a contact between a pair of reference residues, there is a very strong preference for contacts (a precise definition of the "contacts" is provided later) between the preceding and the following residues. Indeed, the contact maps of globular proteins contain very characteristic strips (Godzik et al., 1993). Those near the diagonal correspond to the intrahelical contacts; those farther from the diagonal (parallel to or antiparallel to the diagonal) correspond to contacts between β-strands within β-sheets. Thus we introduce the following energetic bias toward such a mode of packing:

$$E_{\mathrm{map}} = -\epsilon_{\mathrm{gen}} \{\Sigma\Sigma\ (\delta_{\mathrm{i,j}} \cdot \delta_{\mathrm{i+1,j+1}} \cdot \delta_{\mathrm{i-1,j-1}})\delta_{\mathrm{par}} + \Sigma\Sigma\ (\delta_{\mathrm{i,j}} \cdot \delta_{\mathrm{i-1,j+1}} \cdot \delta_{\mathrm{i+1,j-1}})\delta_{\mathrm{apar}}\} \quad (5)$$

where the summations are over all pairs of residues i, j and $\delta_{\mathrm{i,j}}$ is equal to 1 (0) when residues i and j are (are not) in contact. δ_{par} is equal to 1 only when the corresponding chain fragments are oriented in a parallel manner, i.e., the chain vectors satisfy the condition $(\mathbf{v}_{\mathrm{i-1}} + \mathbf{v}_{\mathrm{i}}) \cdot (\mathbf{v}_{\mathrm{j-1}} + \mathbf{v}_{\mathrm{j}}) > 0$; otherwise $\delta_{\mathrm{par}} = 0$. Similarly, δ_{apar} is equal to 1 when the chain fragments are antiparallel and is equal to zero otherwise. For a given contact of a pair of residues, the maximum energetic stabilization due to regular side-chain packing is therefore equal to $-\epsilon_{\mathrm{gen}}$, which has the same value as in the previously defined potentials.

The packing cooperativity of the model protein is further enhanced by a term that mimics main-chain hydrogen bonds. The geometry of protein hydrogen bonds is translated into a specific range of the model chain geometry. First, let us define a vector that is likely to connect the model beads that are within motifs that represent regular secondary structure elements. Such vectors should connect beads i and $i + 3$ in a helix and the appropriate beads in a β-sheet structure. An optimization procedure leads to the following definition:

$$\mathbf{h}_{\mathrm{i}} = 3.3\ (\mathbf{v}_{\mathrm{i-1}} \otimes \mathbf{v}_{\mathrm{i}})/|(\mathbf{v}_{i-1} \otimes \mathbf{v}_{\mathrm{i}})| - \mathbf{v}_{\mathrm{i-1}}/|\mathbf{v}_{\mathrm{i-1}}| \quad (6)$$

The value of the 3.3 prefactor has been found to be optimal (or near the optimal value) for reproducing the internal main-chain hydrogen bonding in the lattice projected PDB structures. However, it should be noted that, because of the wide distribution of the model bond lengths, there are always some hydrogen bonds missed in the model. The coordinates of the vectors $\mathbf{h}_{\mathrm{i}}$ are rounded off to the nearest integer value. In a helix, the $\mathbf{h}_{\mathrm{i}}$ vectors have a length of about three lattice units in the direction perpendicular to the three-residue plane (the first term in the above sum) and are tilted back by a lattice unit (the second term of Eq. 6). The projection along the helix axis is also about three lattice units; this nicely coincides with the 1.5-Å longitudinal increment per residue in a real helix. A residue i is considered to be hydrogen bonded with residue j when the vector $\mathbf{h}_{\mathrm{i}}$ points to any of the 19 points of the excluded volume cluster of residue j. Correspondingly, vector $-\mathbf{h}_{\mathrm{i}}$ may point to another cluster. Because the excluded volume clusters never overlap, the maximum number of these "hydrogen bonds" originating from residue i is equal to 2. The total energy of the "hydrogen bond network" can be written as

$$E_{\mathrm{H\text{-}bond}} = -\epsilon_{\mathrm{H\text{-}bond}}\Sigma\ (\delta^{+} + \delta^{-} + \delta^{+,-}) \quad (7)$$

where δ^{+} (δ^{-}) = 1 when the vector $\mathbf{h}_{\mathrm{i}}$ ($-\mathbf{h}_{\mathrm{i}}$) connects with an excluded volume cluster, and $\delta^{+,-} = 1$ when both vectors connect to some clusters, respectively. Otherwise, the corresponding terms are equal to zero. The cooperative contribution, $\delta^{+,-}$, corresponds to the local saturation of the hydrogen bond network. The "long-range" ($|i - j| > 4$) hydrogen bonds between the residues predicted as helical and between helical and β-type expanded residues were ignored.

Short-range interactions

The short-range potentials were implemented in the form of energy histograms for pairwise specific distances $r(A_{\mathrm{i}}, B_{\mathrm{j}})$, with $|i - j| = 1$, 2, 3, and 4. The reference state is the average that neglects the amino acid identity. In Table 1, we demonstrate the assumed discretization of distances for a few selected interactions. The full sets of data are provided in our home pages (http://bioinformatics.danforthcenter.org or http://biocomp.chem.uw.edu/pl).

Pairwise interactions

The pairwise interactions between model residues are defined as contact potentials in the form of a square well function:

$$E_{\mathrm{ij}} = \begin{cases} \infty, & \text{for } r_{\mathrm{ij}} < 3 \\ E^{\mathrm{rep}}, & \text{for } 3 \le r_{\mathrm{ij}} < R^{\mathrm{rep}}_{\mathrm{i,j}} \\ \epsilon_{\mathrm{ij}}, & \text{for } R^{\mathrm{rep}}_{\mathrm{i,j}} \le r_{\mathrm{ij}} < R_{\mathrm{i,j}} \\ 0, & \text{for } R_{\mathrm{i,j}} < r_{\mathrm{ij}} \end{cases} \quad (8)$$

where ϵ_{ij} are the pairwise interaction parameters, r_{ij} is the distance between chain beads i and j, $E^{\mathrm{rep}} = 3kT$ is a constant repulsive term operating at very short distances, and $R^{\mathrm{rep}}_{\mathrm{i,j}}$ and $R_{\mathrm{i,j}}$ are the cutoff values that depend on amino acid type. The values of these cutoff parameters are provided in Table 2. The pairwise interaction parameters were derived from the statistics of the known protein structure, using the quasichemical approximation. These parameters are orientation dependent and are different for parallel and antiparallel contacts. Parallel contacts are those for which the dot product of the "side-chain vectors" (vectors generated as a difference of the two neighboring chain bonds) is positive. The others are antiparallel contacts. A more pre-

TABLE 1 Examples of pairwise short-range interactions

Distance	$r_{i,i+1} < 4.5$	(4.5,5.5)	(5.5,6.5)	(6.5,7.5)	>7.5 Å
Potential					
G-G	−1.61	2.0	2.0	2.0	2.0
G-T	−0.43	−1.42	2.0	2.0	2.0
A-A	−0.38	−1.26	2.0	2.0	2.0
V-V	2.0	−0.11	−1.36	2.0	2.0
I-K	2.0	1.47	−0.09	−0.85	−0.29

Distance	$r_{i,i+2} < 6$	(6,7)	(7,8)	(8,9)	(9,10)	>10
Potential						
G-G	−1.04	−0.93	0.47	2.0	2.0	2.0
G-T	−0.17	−0.90	−0.74	1.34	2.0	2.0
A-A	0.63	−1.53	0.28	2.0	2.0	2.0
V-V	0.61	−1.05	−0.70	0.67	2.0	2.0
I-K	1.51	0.81	0.16	−0.92	−0.46	0.98

Distance	$r^*_{i,i+3} < -12$	(−12,−8)	(−8,−4)	(−4,0)	(0,4)	(4,8)	(8,12)	>12
Potential								
G-G	0.58	−1.07	0.61	2.0	2.0	−0.41	−0.76	1.8
G-T	0.79	−0.83	0.58	2.0	2.0	−0.47	−0.99	0.99
A-A	0.98	−0.45	1.50	2.0	2.0	−1.61	0.37	2.0
V-V	0.09	−1.26	1.69	2.0	2.0	−0.86	0.32	1.36
I-L	0.44	−1.14	1.33	2.0	2.0	−1.18	0.55	2.0

Distance	$r_{i,i+4} < 5.5$	(5.5,7.5)	(7.5–9.5)	(9.5,11,5)	(11.5–13.5)	(13.5–15.5)	>15.5 Å
Potential							
G-G	0.54	−0.42	−0.04	−0.26	−0.81	−1.20	2.0
G-T	0.96	−0.10	0.27	−0.28	−0.75	−0.27	2.0
A-A	1.32	−1.42	0.31	0.51	0.07	1.12	2.0
V-V	1.43	−0.84	0.60	−0.09	−0.61	0.06	2.0
I-K	1.51	−0.61	−0.01	−0.05	−0.16	−0.23	0.81

All distances are given in Angstroms. $r^*_{i,i+3}$ is the "chiral" value, negative for left-handed, and positive for right-handed conformations. The listed pair of amino acids is located on the ends of the $i,i+k$ fragment. A value of 2.0 of the potential corresponds to distances not observed in proteins.

cise definition of the contact "orientation" was given in the paragraph describing the generic packing cooperativity. The values of pairwise interaction parameters are given in Table 3:

$$E_{pair} = \Sigma\Sigma E_{ij} \quad (9)$$

where the summations are over all $j > i$ pairs of residues.

Multibody potentials

The hydrophobic interactions in our model are partially accounted for by the pairwise interactions. This is not sufficient, however, so a surface exposure statistical potential was developed. The scheme is as follows. Each model residue has assigned 24 surface contact points. A specific subset of these contact points became occupied upon contact with other residues. The main-chain Cα atoms contribute separately to the coverage of a given residue. The positions of the Cα atom could be quite well approximated, given the positions of three consecutive side-chain beads (Kolinski and Skolnick, 1998). Some contact points could be multiply occupied. The fraction of the nonoccupied surface points defines the exposed fraction of a given side chain. Proper potentials could be derived from the statistical analysis of the protein structures for which the solvent exposure has been determined on the atomic level. The total surface energy can be computed as follows:

$$E_{surface} = \Sigma\, E_b(A_i, a_i) \quad (10)$$

where a_i is the covered fraction of the residue A_i and $E_b(A_i, a_i)$ is the value of statistical potential when amino acid type A has a_i of its surface points occupied, i.e., the covered fraction of its surface is equal to $a_i/24$.

Studying the distribution of interresidue contacts in globular proteins, we have found that various amino acids have different tendencies to pack in a parallel or antiparallel fashion. A contact between residues i and j is considered

TABLE 2 Compilation of pairwise cut-off distances for pairwise interactions

A_i	A_j	$R^{rep}_{i,j}$ (Å)	$R_{i,j}$ (Å)
Small*	Small	4.35[#]	5.97
Large[§]	Large	4.83	6.80
Other	Combinations[¶]	4.57	6.32

*Small amino acids in the lattice model are Gly, Ala, Ser, Cys.
[#]This value corresponds to the excluded volume radius of three lattice units; therefore, for pairs of small amino acids, the soft-core envelope does not exist.
[§]Large amino acids are Phe, Tyr, Trp.
[¶]Other combinations means the following: small-large, medium-large, or medium-small ("medium" means other than small or large).

TABLE 3 Side-group pairwise interaction parameters

	Gly	Ala	Ser	Cys	Val	Thr	Ile	Pro	Met	Asp	Asn	Leu	Lys	Glu	Gln	Arg	His	Phe	Tyr	Trp
Parallel contacts																				
Gly	0.4	0.4	0.2	−0.3	0.1	0.0	0.0	0.0	0.1	0.1	−0.1	−0.1	0.1	−0.1	0.0	−0.3	−0.4	−0.2	−0.3	−0.3
Ala	0.4	0.4	0.1	0.0	−0.2	0.1	−0.4	0.3	−0.1	0.1	0.2	−0.3	0.2	0.5	0.1	0.0	0.0	−0.2	−0.3	−0.1
Ser	0.2	0.1	−0.2	−0.4	−0.2	−0.2	0.0	0.0	−0.2	−0.4	−0.1	0.0	0.0	−0.4	−0.3	−0.1	−0.2	−0.2	−0.4	−0.3
Cys	−0.3	0.0	−0.4	−0.8	−0.5	−0.2	−0.8	−0.2	−0.4	−0.2	−0.3	−0.5	0.0	−0.1	−0.2	−0.3	−0.5	−0.9	−0.5	−0.5
Val	0.1	−0.2	−0.2	−0.5	−0.9	−0.5	−0.9	−0.1	−0.5	0.0	0.1	−1.2	−0.1	0.1	−0.3	−0.4	−0.4	−0.7	−0.9	−0.7
Thr	0.0	0.1	−0.2	−0.2	−0.5	−0.5	−0.6	−0.1	−0.1	−0.8	−0.4	−0.3	−0.2	−0.6	−0.6	−0.6	−0.4	−0.4	−0.6	−0.5
Ile	0.0	−0.4	0.0	−0.8	−0.9	−0.6	−1.0	−0.3	−0.5	0.0	0.0	−1.1	−0.2	−0.1	−0.1	−0.4	−0.2	−1.0	−0.9	−0.9
Pro	0.0	0.3	0.0	−0.2	−0.1	−0.1	−0.3	0.2	−0.2	0.2	−0.2	0.0	−0.1	0.0	−0.2	−0.4	−0.2	−0.2	−0.5	−0.4
Met	0.1	−0.1	−0.2	−0.4	−0.5	−0.1	−0.5	−0.2	−0.5	0.1	0.0	−0.8	−0.1	−0.1	−0.1	−0.5	−0.2	−0.8	−0.3	−0.3
Asp	0.1	0.1	−0.4	−0.2	0.0	−0.8	0.0	0.2	0.1	−0.2	−0.6	0.1	−0.7	0.0	−0.4	−0.9	−0.3	0.0	−0.5	−0.1
Asn	−0.1	0.2	−0.1	−0.3	0.1	−0.4	0.0	−0.2	0.0	−0.6	−0.5	−0.1	−0.5	−0.4	−0.6	−0.6	−0.2	−0.1	−0.4	−0.1
Leu	−0.1	−0.3	0.0	−0.5	−1.2	−0.3	−1.1	0.0	−0.8	0.1	−0.1	−1.1	0.0	−0.1	−0.3	−0.3	−0.4	−1.2	−1.0	−0.9
Lys	0.1	0.2	0.0	0.0	−0.1	−0.2	−0.2	−0.1	−0.1	−0.7	−0.5	0.0	0.1	−0.8	−0.6	−0.2	−0.2	0.2	−0.4	−0.2
Glu	−0.1	0.5	−0.4	−0.1	0.1	−0.6	−0.1	0.0	−0.1	0.0	−0.4	−0.1	−0.8	−0.1	−0.2	−1.2	−0.5	−0.2	−0.5	−0.3
Gln	0.0	0.1	−0.3	−0.2	−0.3	−0.6	−0.1	−0.2	−0.1	−0.4	−0.6	−0.3	−0.6	−0.2	−0.2	−0.7	−0.3	−0.6	−0.6	−0.3
Arg	−0.3	0.0	−0.1	−0.3	−0.4	−0.6	−0.4	−0.4	−0.5	−0.9	−0.6	−0.3	−0.2	−1.2	−0.7	−0.3	−0.4	−0.4	−0.6	−0.3
His	−0.4	0.0	−0.2	−0.5	−0.4	−0.4	−0.2	−0.2	−0.2	−0.3	−0.2	−0.4	−0.2	−0.5	−0.3	−0.4	−0.4	−0.2	−0.6	−0.3
Phe	−0.2	−0.2	−0.2	−0.9	−0.7	−0.4	−1.0	−0.2	−0.8	0.0	−0.1	−1.2	0.2	−0.2	−0.6	−0.4	−0.2	−0.8	−1.0	−0.6
Tyr	−0.3	−0.3	−0.4	−0.5	−0.9	−0.6	−0.9	−0.5	−0.3	−0.5	−0.4	−1.0	−0.4	−0.5	−0.6	−0.6	−0.6	−1.0	−0.6	−0.5
Trp	−0.3	−0.1	−0.3	−0.5	−0.7	−0.5	−0.9	−0.4	−0.3	−0.1	−0.1	−0.9	−0.2	−0.3	−0.3	−0.3	−0.3	−0.6	−0.5	−0.3
Antiparallel contacts																				
Gly	0.3	0.4	0.3	−0.3	0.1	0.1	0.0	0.2	0.0	0.4	0.2	−0.2	0.4	0.3	−0.2	0.1	0.1	−0.2	−0.2	−0.4
Ala	0.4	0.2	0.4	−0.3	−0.2	0.1	−0.5	0.2	−0.2	0.5	0.1	−0.3	0.5	0.3	0.1	0.2	0.1	−0.3	−0.5	−0.1
Ser	0.3	0.4	0.1	0.0	0.4	0.2	0.2	0.3	0.0	0.3	0.0	0.1	0.2	0.2	0.3	0.2	−0.1	−0.2	−0.1	−0.2
Cys	−0.3	−0.3	0.0	−1.3	−0.3	−0.4	−0.5	−0.2	−0.4	−0.3	−0.1	−0.5	−0.2	−0.1	−0.1	−0.2	−0.3	−0.6	−0.4	−0.5
Val	0.1	−0.2	0.4	−0.3	−0.6	0.0	−0.9	0.0	−0.4	0.3	0.3	−1.2	0.4	0.2	0.0	0.0	−0.1	−0.8	−0.6	−0.5
Thr	0.1	0.1	0.2	−0.4	0.0	0.2	−0.2	0.1	−0.3	0.3	0.2	−0.2	0.4	0.2	0.1	0.2	−0.1	−0.1	−0.3	−0.2
Ile	0.0	−0.5	0.2	−0.5	−0.9	−0.2	−1.0	−0.2	−0.8	0.3	0.1	−1.3	0.3	0.1	−0.3	0.0	0.0	−1.2	−0.8	−0.6
Pro	0.2	0.2	0.3	−0.2	0.0	0.1	−0.2	0.0	0.3	0.0	−0.2	−0.2	0.3	0.2	−0.2	0.2	−0.1	0.0	−0.5	−0.4
Met	0.0	−0.2	0.0	−0.4	−0.4	−0.3	−0.8	0.3	−0.4	0.1	0.1	−0.6	0.1	0.0	0.0	−0.1	−0.2	−0.6	−1.0	−0.4
Asp	0.4	0.5	0.3	−0.3	0.3	0.3	0.3	0.0	0.1	0.2	0.2	0.4	0.0	0.3	0.1	−0.2	−0.1	0.0	−0.2	−0.1
Asn	0.2	0.1	0.0	−0.1	0.3	0.2	0.1	−0.2	0.1	0.2	−0.1	0.2	0.0	0.1	0.0	0.1	−0.1	−0.1	0.0	−0.4
Leu	−0.2	−0.3	0.1	−0.5	−1.2	−0.2	−1.3	−0.2	−0.6	0.4	0.2	−1.3	0.2	0.2	−0.1	−0.2	−0.1	−1.3	−1.0	−0.6
Lys	0.4	0.5	0.2	−0.2	0.4	0.4	0.3	0.3	0.1	0.0	0.0	0.2	0.2	0.2	0.3	0.3	0.1	−0.1	0.1	0.2
Glu	0.3	0.3	0.2	−0.1	0.2	0.2	0.1	0.2	0.0	0.3	0.1	0.2	0.2	−0.1	0.2	−0.4	−0.1	−0.1	−0.2	−0.2
Gln	−0.2	0.1	0.3	−0.1	0.0	0.1	−0.3	−0.2	0.0	0.1	0.0	−0.1	0.3	0.2	0.0	0.1	0.1	−0.1	−0.1	−0.2
Arg	0.1	0.2	0.2	−0.2	0.0	0.2	0.0	0.2	−0.1	−0.2	0.1	−0.2	0.3	−0.4	0.1	0.1	0.1	−0.3	−0.3	−0.1
His	0.1	0.1	−0.1	−0.3	−0.1	−0.1	0.0	−0.1	−0.2	−0.1	−0.1	−0.1	0.1	−0.1	0.1	0.1	−0.2	−0.6	−0.5	−0.3
Phe	−0.2	−0.3	−0.2	−0.6	−0.8	−0.1	−1.2	0.0	−0.6	0.0	−0.1	−1.3	−0.1	−0.1	−0.1	−0.3	−0.6	−0.8	−0.8	−0.8
Tyr	−0.2	−0.5	−0.1	−0.4	−0.6	−0.3	−0.8	−0.5	−1.0	−0.2	0.0	−1.0	0.1	−0.2	−0.1	−0.3	−0.5	−0.8	−0.6	−0.7
Trp	−0.4	−0.1	−0.2	−0.5	−0.5	−0.2	−0.6	−0.4	−0.4	−0.1	−0.4	−0.6	0.2	−0.2	−0.2	−0.1	−0.3	−0.8	−0.7	−0.4

"parallel" when $(\mathbf{v}_{i-1} - \mathbf{v}_i) \cdot (\mathbf{v}_{j-1} - \mathbf{v}_j) > 0$, and "antiparallel" otherwise. Moreover, there are strong correlations between the number of parallel and antiparallel contacts, given the total number of contacts of a given residue. Because of the reduced character of our model, the other contributions to the force field do not properly account for such effects. Therefore, the model force field has been supplemented by the following multibody potential:

$$E_{\text{multi}} = \Sigma\, E_{\text{m}}(A, n_{\text{p}}, n_{\text{a}}) \tag{11}$$

where $E_{\text{m}}(A, n_{\text{p}}, n_{\text{a}})$ is the value of statistical potential for residue type A having n_{p} parallel and n_{a} antiparallel contacts. The reference state is a random distribution of contacts. The values along particular diagonals ($n_{\text{p}} + n_{\text{a}} = n_{\text{c}}$) have been renormalized in such a way that the lowest energy for a diagonal was exactly equal to the value of the corresponding statistical potential derived from the distribution of the total number of contacts n_{c} for a given type of residue. Examples of such a potential are given in Table 4. The numbers in the head row and in the first column correspond to the number of parallel and antiparallel contacts, respectively.

The total internal conformational energy of the model chain was equal to

$$E = 1.25(E_{\text{pair}} + E_{\text{stiff}} + E_{\text{map}} + E_{\text{struct}}) + 0.875E_{\text{H-bond}} + 0.75E_{\text{short}} + 0.5(E_{\text{surface}} + E_{\text{multi}}) \tag{12}$$

with the value of generic parameter $\epsilon_{\text{gen}} = 1\ kT$.

The relative scaling of various potentials has been adjusted by trial and error in *ab initio* folding experiments performed for a few small proteins. The objective was to

TABLE 4 Examples of multibody, orientation-dependent interaction parameters*

	0	1	2	3	4	5	6	7	8
Alanine									
0	0.1	−0.3	−0.4	−0.1	0.7	1.7	2.0	2.0	2.0
1	0.8	−0.2	−0.5	−0.3	0.2	1.0	2.0	2.0	2.0
2	0.7	−0.5	−0.7	−0.6	0.1	1.3	2.0	2.0	2.0
3	0.1	−0.6	−0.8	−0.4	0.8	2.0	2.0	2.0	2.0
4	0.4	−0.4	−0.2	0.6	2.0	2.0	2.0	2.0	2.0
5	1.3	0.7	1.3	2.0	2.0	2.0	2.0	2.0	2.0
6	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
7	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
8	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
Lysine									
0	0.6	0.1	−0.3	−0.4	0.1	0.8	2.0	2.0	2.0
1	1.8	0.6	−0.3	−0.3	0.2	1.1	2.0	2.0	2.0
2	2.0	0.6	0.0	−0.1	0.5	1.3	2.0	2.0	2.0
3	2.0	1.2	0.6	0.6	1.2	1.9	2.0	2.0	2.0
4	2.0	2.0	1.4	1.3	1.8	2.0	2.0	2.0	2.0
5	2.0	2.0	2.0	1.8	2.0	2.0	2.0	2.0	2.0
6	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
7	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
8	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
Phenylalanine									
0	1.6	1.0	0.5	0.3	0.3	0.9	1.2	2.0	2.0
1	2.0	1.2	0.6	0.1	0.2	0.3	1.3	1.2	2.0
2	2.0	1.0	−0.1	−0.3	−0.2	0.1	0.2	0.4	1.6
3	1.8	0.8	−0.2	−0.6	−0.5	−0.8	−0.7	0.1	2.0
4	1.4	0.5	−0.3	−0.8	−1.1	−1.0	−0.6	0.6	2.0
5	2.0	0.8	−0.2	−1.1	−1.1	−1.0	−0.1	0.8	2.0
6	2.0	1.1	−0.1	−0.7	−0.7	−0.5	0.2	2.0	2.0
7	2.0	1.8	0.2	−0.1	0.3	0.8	2.0	2.0	2.0
8	2.0	2.0	1.5	1.0	1.8	2.0	2.0	2.0	2.0

*The top row and left-hand column indicate the number of parallel and antiparallel contacts, respectively.

maintain a low secondary structure content in the random coiled state and dense packing with a proper level of secondary structure in the collapsed globular state. The model is not sensitive to small variations of these scaling parameters.

Sampling procedures

MCD was performed using a standard asymmetrical Metropolis scheme. The set of local moves involved two-bond moves, chain end moves (two-bond), and three-bond moves as described elsewhere. To study some aspects of local dynamics, larger scale moves were not applied in the scheme.

ESMC simulations were performed in the same fashion as described previously. The interval of the generated energy histogram was equal to $1kT$, and the observed range of the model internal energy was from about -115 to about $+20$.

RESULTS

The sequence used in the present studies is GEWTYDDATKTFTVTE; it consists of the G41-E56 fragment of the B1 domain of protein G. A reduced protein model is used. Two different sampling techniques were employed in these studies: Monte Carlo dynamics (MCD) at various temperatures and Entropy Sampling Monte Carlo (ESMC), which provides a full thermodynamic description of the model system.

Folding thermodynamics

Standard Monte Carlo simulations allow an estimation of the system's configurational energy and heat capacity at a given temperature (note that by temperature we really mean a reduced temperature, expressed in dimensionless kT units, where k is Boltzmann's constant and T is absolute temperature). To obtain the average energy and to identify the transition temperature, long simulations (MCD) were performed at several temperatures covering a wide range that certainly contains the folding temperature. The resulting estimates of the system energy and the heat capacity (computed from the energy fluctuations) provide sufficient data for a rough identification of the transition midpoint.

A relatively new Monte Carlo sampling technique (ESMC) allows for the simultaneous statistical estimation of the energy and entropy in a single simulation series (Scheraga and Hao, 1999). Such simulations are quite expensive, but the obtained data are valid for all temperatures. Furthermore, from ESMC, one obtains an estimate of the partition function, and therefore thermodynamic quantities are calculated from analytical expressions.

The fact that the same results are obtained from the two simulation techniques provides a strong validation of the methodology and indicates that there is no kinetic frustration in the model and that the results provide "a true" description of the model system. In Fig. 1, the energy and heat capacity are plotted as a function of the system temperature. The data from the ESMC are plotted in the continuous solid curves. The data from MCD (at various spe-

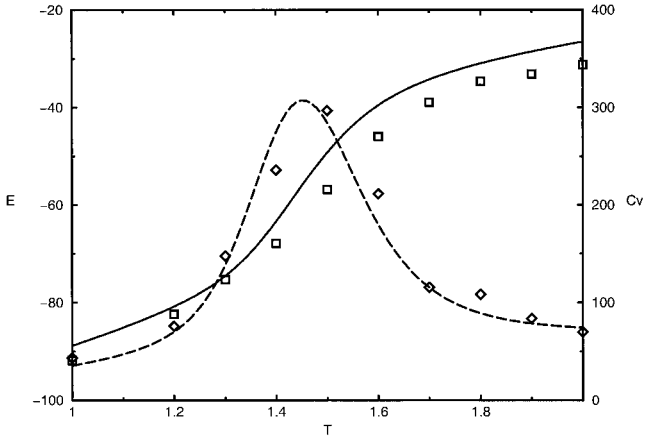

FIGURE 1 Thermodynamic properties of the C-terminal hairpin of protein G. The solid line (*dashed line*) corresponds to the system conformational energy (heat capacity) obtained from ESMC calculations. Squares and diamonds represent data from Metropolis Monte Carlo sampling at various temperatures.

cific temperatures) are plotted in the dashed line. The heat capacity has a higher statistical error than the configurational energy. Data from the two simulation techniques are in good agreement. A small systematic deviation at high temperatures apparently results from a trick used to speed up the ESMC sampling; namely, the population of very high-energy conformations (in the upper part of the random coil part of energy spectrum) was artificially suppressed. ESMC allows the calculation of free energy profiles (as a function of the configurational energy) at various (arbitrarily chosen) temperatures. At the transition midpoint, the free energy of low-energy and high-energy states is the same. From the free energy profile (see Fig. 2) at the transition temperature, one can extract the value of the free energy barrier between the folded and unfolded states. The height of the barrier is $\sim 0.75kT$. This indicates that the system exhibits a weakly cooperative transition. The population of intermediates at the transition temperature is therefore low and is $\sim$20% of all conformations. It is interesting to observe the structural properties of representative states at various values of the energy. Analysis of the low-energy states (near the left-hand minimum of the free energy profile) presents folded conformations that differ from each other with a root mean square deviation (RMSD) of less than 1 Å. The manifold of unfolded conformations corresponds to the free energy minimum at high energy. Conformations that correspond to the free energy barrier are rather diverse; however, a large fraction have a native-like turn region. Fig. 3 shows snapshots of representative conformations for various internal energy levels. This defines the energy landscape of the model that could be studied in detail. The low conformational energy states have a well-defined β-hairpin structure and a well-defined pattern of side-chain contacts and hydrogen bonding.

In Fig. 4, we plot the distribution histogram of the number of native contacts per conformation at three distinct temperatures. Indeed, at the transition temperature, the distribution of the number of contacts is bimodal, indicating the preference for either folded or unfolded states. At higher temperatures, the most probable number of contacts is typical of the unfolded state, whereas the native pattern dominates at lower temperatures. The same can be observed for the pattern of model hydrogen bonds.

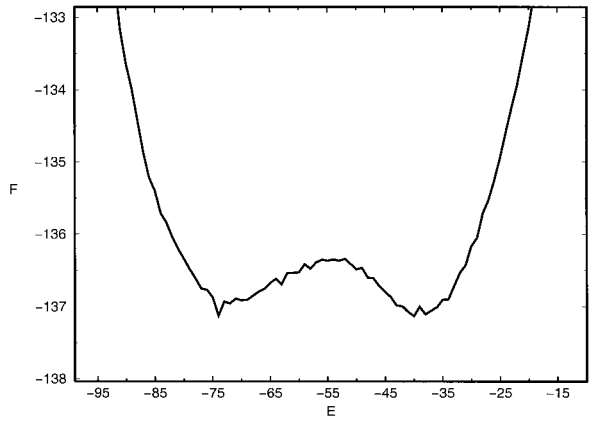

FIGURE 2 Free energy as a function of conformational energy at $T = 1.456$, obtained from ESMC. The existence of a free energy barrier indicates a weakly cooperative transition.

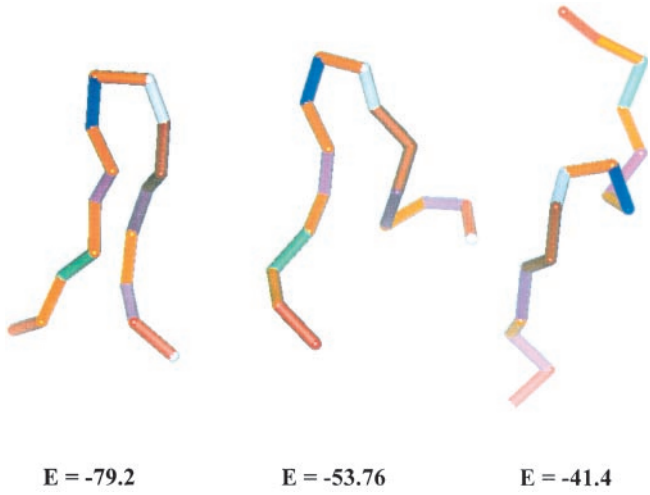

FIGURE 3 Representative conformations of the model peptide at various conformational energy levels extracted from ESMC simulations. From left to right: an example of the folded state (at the low energy free energy minimum), a typical intermediate (at the top of the free energy barrier), and a high-probability unfolded state.

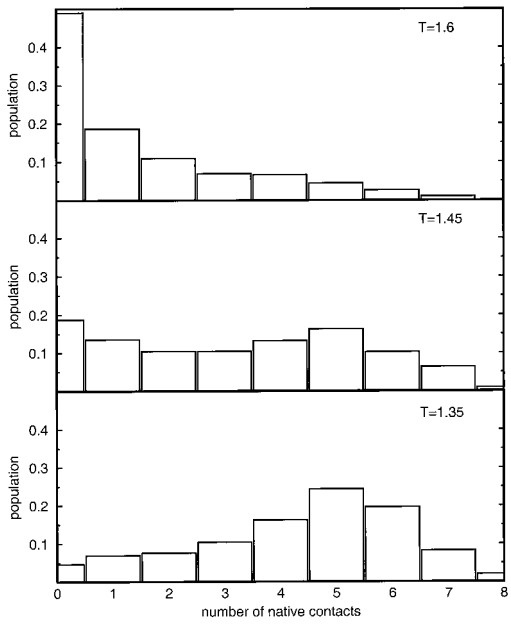

FIGURE 4 Population of various states (according to the number of native contacts) at three temperatures, above the transition (*top*), near the transition, and below the transition (*bottom*). At the transition, the histogram is bimodal, indicating some features of an all-or-none transition. The maximum of five contacts below the transition temperature reflects the mobility of the end segments (and some additional small fluctuations) in the folded state. Data were extracted from long MMC runs.

Folding mechanism

MCD simulations at the transition temperature and near the transition provide a detailed description of the folding pathway. Analysis of successful folding events shows that in the vast majority of cases, folding initiates by the formation of the β-turn, which is followed by successive formations of the remaining contacts along the hairpin. In many cases, the turn forms in the wrong place. Such folding attempts are usually unsuccessful. A competing, less frequent mechanism involves the formation of a hydrophobic cluster involving the F and V residues in the first strand and the Y and W residues in the second putative strand of the β-hairpin. The assembly of the rest of the hairpin follows. The end residues (G and E) are mobile even well below the folding temperature. This is further illustrated in Fig. 4, which shows the distribution of the number of native contacts observed at various temperatures. The folded state is therefore quite degenerate. Eventually, at a much lower temperature, the end residues become frozen in the hairpin structure.

Fig. 5 shows snapshots of a very typical folding pathway extracted from a high-density trajectory near the folding temperature. Fig. 6 shows flow charts from high-density trajectories. The points represent various native contacts in the hairpin. The highest line displays the D-K contacts near the turn, the second one the Y-F contacts, and the lowest one the W-V contacts, as a function of time. The top panel shows a short time window extracted from the longer time data displayed in the bottom panel. Inspection of these flow charts confirms our observation that typical folding events start from the putative turn. The native contacts usually form by starting from the turn as well. Nucleation near the turn is frequently, but not always, followed by a rapid rearrangement that leads to the folded structure. Inspection of several folding/unfolding events near the transition temperature shows that unfolding is somewhat slower than folding. The bottom panel demonstrates the cooperativity of the process. The majority of the snapshots correspond to either a folded or unfolded state, and the population of intermediates is low.

What is the nature of the unfolded state? Inspection of the MCD trajectories shows very high chain mobility at temperatures above the transition. Here essentially all possible conformations characteristic of a semiflexible polymeric random coil could be observed. However, very mobile partially helical conformations contribute noticeably to the unfolded state. This is quite interesting because the sequence has a strong β-type secondary propensity. As suggested by experiment, the coil-helix transition is much faster than β-sheet formation. Moreover, short helical conformations can provide easy access to locally compact structures. Thus perhaps a low helical content in the denatured state is not so unusual.

As mentioned before, the folded state contains an ensemble of structures; however, the level of structural degeneracy is orders of magnitude less than in the denatured state. The most visible fluctuations involve the end residues. In our force field, the Gly-Glu interactions are slightly repulsive, which is rather physical. The cooperative terms of the interaction scheme (see the Methods section) are not sufficiently strong to provide structural fixation at the transition temperature. There are also other structural fluctuations. While the majority of the native contacts and the hydrogen bond network (except for the above-mentioned two end residues) are fixed in the native state, some additional fluctuations persist. A trivial one involves small fluctuations of the dihedral angles that maintain the interaction pattern of a β-hairpin. More interestingly, the F-W contact breaks and forms quite frequently, even below the folding temperature. Under these conditions, the remaining contacts within the "hydrophobic core" of the hairpin are essentially fixed.

Folding of modified sequences

The explanation provided by Munoz and co-workers suggests that the hydrophobic cluster's long distance from the turn is the main factor responsible for a slower folding rate and higher folding cooperativity of the β-hairpin with respect to helical sequences. If so, a mutation that shifts the location of the hydrophobic cluster should change the fold-

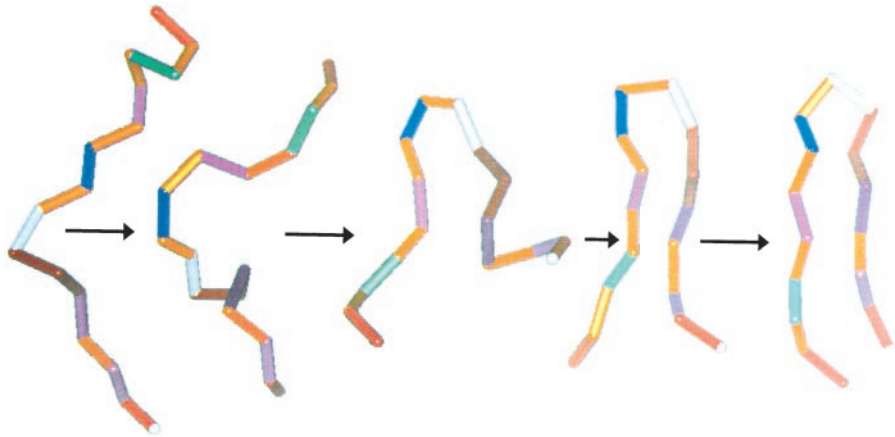

FIGURE 5 A typical folding pathway near the transition temperature extracted from a high-density (short time) trajectory of MMC simulation. This particular sequence of events corresponds to the folding times between $t = 575$ and 580 of the flow chart shown in Fig. 6 (the time t is counted as the number of elapsed Monte Carlo cycles).

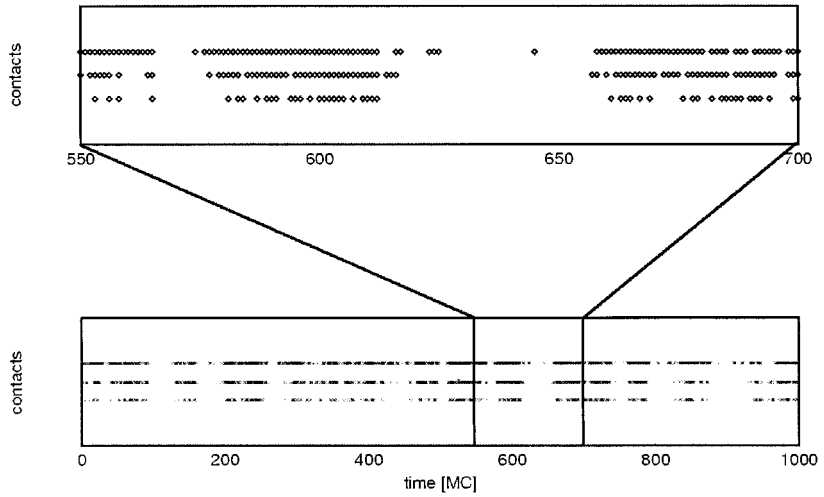

FIGURE 6 Flow charts illustrating the formation of some native contacts during the MMC simulations near (slightly below) the transition temperature. The highest row in each panel corresponds to D-K contacts near the turn, the second row is for Y-F contacts, and the lowest row represents the W-V contacts. The upper panel shows a short-time window of simulations extracted from the relatively short trajectory illustrated in the lower panel. Two complete unfolding/folding events can be observed in the upper panel.

ing cooperativity. For this reason, we also studied two modified sequences. The first sequence (s1) has the hydrophobic residues shifted toward the chain ends and reads as follows: GWTYEDDATKTTFTVE. The second sequence (s2) has the hydrophobic residues closer to the turn: GEDWTYDATKFTVTTE. It is assumed that the resulting modification of the hairpin face itself should have no effect on the folding process because the hairpin is isolated. The two sequences folded into very similar hairpin structures. Surprisingly, the cooperativity of the transition increases slightly from sequence s1 through the original sequence s to sequence s2, and the estimated free energy barriers are 0.52, 0.75, and 0.80 kT. The slight change in the transition temperature indicates a small increase in the hairpin stability with the shift of the hydrophobic cluster toward the turn. In the series s1, s, s2, the folding temperatures are 1.485, 1.456 and 1.426. Thus the effect is consistent, but small. The observed changes are only a few times larger than the error of the method.

DISCUSSION

The results of simulations described in this work show qualitative agreement with recent experimental studies. In agreement with experiment, these simulations indicate that the C-terminal β-hairpin of the B1 domain of protein G is capable of folding into a unique native-like state. The transition is cooperative and has the features of an all-or-none folding transition. The level of cooperativity observed in the simulations is lower than that suggested by experimental studies. It should be noted that the specific value of the free energy barrier prescribed to experiment has been deduced from a simplified statistical mechanical model that was fitted to the experimental data. Because a number of possibly competing interactions were omitted, the actual value of the barrier might be lower. On the other hand, the hairpin population versus temperature observed in these simulations is qualitatively the same as that deduced from experiment. Fig. 7 shows the hairpin population as a function of the reduced temperature of the model. The hairpin population is computed from the number of observed native contacts. To allow for the previously mentioned higher mobility of the chain ends, it was assumed that those conformations having four or more native contacts (including the two contacts in the hydrophobic cluster and two contacts near the turn) are in the folded state. To compare the curves obtained in experiment with those from our simulations, the dimensionless reduced temperature has to be converted into degrees Kelvin by multiplying our temperature scale by a factor equal to the ratio T_{exp}/T_{MC}, where T_{exp} is the experimental folding temperature (in degrees Kelvin) and T_{MC} is the reduced dimensionless transition temperature determined from simulations. The data obtained in our simulations closely match the experimental results. The solid line is scanned from the plot given from the work of Munoz and

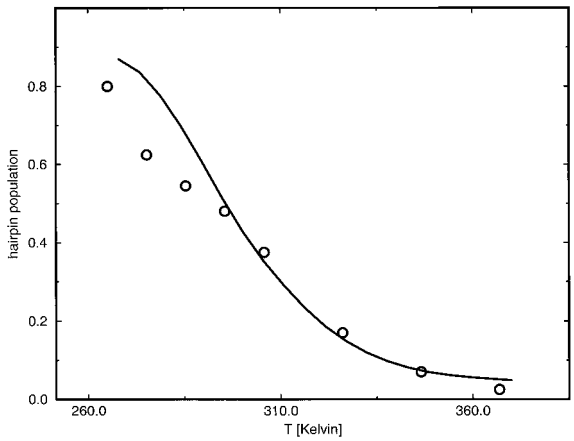

FIGURE 7 Comparison of the experimental data (*solid line*) on the thermal unfolding of the C-terminal hairpin of protein G with the results of the MMC simulations at various temperatures. The experimental data were derived from an interpretation of tryptophan fluorescence and scanned into this plot from Fig. 2 of the work by Munoz et al. (1997). The circles represent simulation results. The hairpin population was estimated by the fraction of conformations having four or more native contacts. The MMC dimensionless reduced temperature translated into Kelvin (see the text).

co-workers; the circles are from the present work. The temperature width of the transition and the content of secondary structure at various temperatures are qualitatively the same.

Although the free energy barrier to folding is found to be smaller than that suggested by an analysis of the experimental data, it may still lead nevertheless to exponential folding kinetics. As a function of temperature, these simulations provide a very similar population of folded states, as seen in the experimental situation. This strongly suggests that the thermodynamics of the real system is very well described by the proposed model. Furthermore, many aspects of the kinetics of assembly are reproduced as well.

Munoz and co-workers propose that the most probable way to initiate folding is from the β-turn. In our simulations, we also observed such a folding pathway as the statistically dominant fraction of successful folding events. However, a noticeable fraction of folding sequences started from the formation of the hydrophobic cluster in the center of the putative hairpin. After such a nucleation event, the rest of the chain frequently readjusted into the hairpin structure. These and other details of the folding pathway are provided by the simulations. Of course, our results could be somewhat biased by the specific design of the model and its force field. However, the qualitative agreement with the "hard" experimental facts encourages us to believe that the other observations should be qualitatively true.

Interestingly, our modifications of the original sequence show that the location of the hydrophobic cluster with respect to the hairpin turn has some effect on protein stability and the cooperativity of the process. It was expected that being closer to the chain end locations, the hydrophobic cluster would increase the protein cooperativity of the process. An opposite effect was observed in our simulations. A possible explanation is that a large fraction of the random coil entropy loss is associated with the formation of the turn region. Formation of the subsequent hairpin segments requires a relatively small change in the system entropy. Thus strong stabilizing interactions near the turn may decrease the number of sampled intermediate states, thereby increasing the cooperativity of the process.

CONCLUDING REMARKS

A reduced high-resolution lattice model of protein structure and dynamics was used in a simulation study of the folding of the C-terminal hairpin of the B1 domain of protein G. In agreement with recent experiments, these simulations show that this short polypeptide has many of the features of globular proteins and folds cooperatively into a well-defined β-hairpin structure. The simulations provide a detailed picture of the folding dynamics and thermodynamics. In particular, there is a free energy barrier separating the manifold of denatured states and a folded state that exhibits some level of structural degeneracy. Folding was usually initiated by formation of the β-turn, while folding initiated by hydrophobic collapse to generate the hydrophobic cluster was less frequent.

Finally, we note that the model employed here allows for simulations of much larger systems of the size of typical single-domain globular proteins. The good agreement with the experimental results for the small system examined here suggests that the proposed methodology could be employed in meaningful simulation studies of the globular protein folding process.

This work was partially supported by KBN (Poland) grant 6PO4A-1413 and National Institutes of Health grant P41 RR12255. AK is an International Scholar of the Howard Hughes Medical Institute.

REFERENCES

Baldwin, R. L. 1995. The nature of protein folding pathways: the classical versus the new view. *J. Biomol. NMR.* 5:103–109.

Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr, M. D. Brice, J. R. Rodgers, O. Kennard, T. Simanouchi, and M. Tasumi. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.

Blanco, F., G. Rivas, and L. Serrano. 1994. A short linear peptide that folds into a native stable β-hairpin in aqueous solution. *Struct. Biol.* 1:584–590.

Blanco, F. J., and L. Serrano. 1995. Folding of protein G B1 domain studied by the conformational characterization of fragments comprising its secondary structural elements. *Eur. J. Biochem.* 230:634–649.

Camacho, C. J., and D. Thirumalai. 1996. A criterion that determines the fast folding of proteins. A model study. *Europhys. Lett.* 35:627–632.

Creighton, T. E. 1993. Proteins: Structures and Molecular Properties. W. H. Freeman and Company, New York.

Dyson, J. H., and P. E. Wright. 1993. Peptide conformation and protein folding. *Curr. Biol.* 3:60–65.

Fersht, A. R. 1993. Protein folding and stability: the pathway of folding of barnase. *FEBS Lett.* 325:5–16.

Friesner, R. A., and J. R. Gunn. 1996. Computational studies of protein folding. *Annu. Rev. Biophys. Biomol. Struct.* 25:315–342.

Godzik, A., J. Skolnick, and A. Kolinski. 1993. Regularities in interaction patterns of globular proteins. *Protein Eng.* 6:801–810.

Gronenborn, A., D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore. 1991. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science.* 253:657–660.

Karplus, M., and A. Sali. 1995. Theoretical studies of protein folding and unfolding. *Curr. Opin. Struct. Biol.* 5:58–73.

Kolinski, A., W. Galazka, and J. Skolnick. 1996. On the origin of the cooperativity of protein folding. Implications from model simulations. *Proteins.* 26:271–287.

Kolinski, A., L. Jaroszewski, P. Rotkiewicz, and J. Skolnick. 1997. An efficient Monte Carlo model of protein chains. Modeling the short-range correlations between side group centers of mass. *J. Phys. Chem.* 102: 4628–4637.

Kolinski, A., P. Rotkiewicz, and J. Skolnick. 1998. Application of high coordination lattice model in protein structure prediction. *In* Monte Carlo Approaches to Biopolymers and Protein Folding. P. Grassberger, G. T. Barkema, and W. Nadler, editors. World Scientific, Singapore. 110–130.

Kolinski, A., and J. Skolnick. 1996. Lattice Models of Protein Folding, Dynamics and Thermodynamics. R. G. Landes, Austin, TX.

Kolinski, A., and J. Skolnick. 1998. Assembly of protein structure from sparse experimental data: an efficient Monte Carlo Model. *Proteins.* 32:475–494.

Munoz, V., P. A. Thompson, J. Hofrichter, and W. A. Eaton. 1997. Folding dynamics and mechanism of β-hairpin formation. *Nature.* 390:196–197.

Ptitsyn, O. B. 1995. Structures of folding intermediates. *Curr. Opin. Struct. Biol.* 5:74–78.

Scheraga, H. A., and M. H. Hao. 1999. Entropy sampling Monte Carlo for polypeptides and proteins. *Adv. Chem. Phys.* 105:243–272.