PROTEINS: Structure, Function, and Genetics 26:271–287 (1996)

On the Origin of the Cooperativity of Protein Folding: Implications From Model Simulations

Andrzej Kolinski,[1,2] Wojciech Galazka,[2] and Jeffrey Skolnick[1]
[1]Department of Molecular Biology, The Scripps Research Institute, La Jolla, California; [2]Department of Chemistry, University of Warsaw, Warsaw, Poland

ABSTRACT **There is considerable experimental evidence that the cooperativity of protein folding resides in the transition from the molten globule to the native state. The objective of this study is to examine whether simplified models can reproduce this cooperativity and if so, to identify its origin. In particular, the thermodynamics of the conformational transition of a previously designed sequence (A. Kolinski, W. Galazka, and J. Skolnick, J. Chem. Phys. 103: 10286–10297, 1995), which adopts a very stable Greek-key β-barrel fold has been investigated using the entropy Monte Carlo sampling (ESMC) technique of Hao and Scheraga (M.-H. Hao and H.A. Scheraga, J. Phys. Chem. 98: 9882–9883, 1994). Here, in addition to the original potential, which includes one body and pair interactions between side chains, the force field has been supplemented by two types of multibody potentials describing side chain interactions. These potentials facilitate the proteinlike pattern of side chain packing and consequently increase the cooperativity of the folding process. Those models that include an explicit cooperative side chain packing term exhibit a well-defined all-or-none transition from a denatured, random coil state to a high-density, well-defined, nativelike low-energy state. By contrast, models lacking such a term exhibit a conformational transition that is essentially continuous. Finally, an examination of the conformations at the free-energy barrier between the native and denatured states reveals that they contain a substantial amount of native-state secondary structure, about 50% of the native contacts, and have an average root mean square radius of gyration that is about 15% larger than native.**

Key words: protein folding, protein thermodynamics, entropy sampling Monte Carlo, reduced protein models, folding intermediates

INTRODUCTION

A remarkable feature of naturally occurring globular proteins and some properly designed or redesigned proteins is that, under suitable conditions of temperature and solvent, they adopt a unique, densely packed, three-dimensional structure. For single-domain proteins, the folding process has some of the features of a first-order phase transition and can be described by an all-or-none model.[1,2] The low-energy native structure seems to be separated from the manifold of denatured and/or molten globule states by a substantial free-energy gap. Consequently, the population of intermediates is very small, and folding is cooperative. However, while valuable insights into the cooperativity of folding have been provided by thermodynamic measurements[1,2] and theoretical investigations of protein-like model systems,[3–14] the physical origin of this cooperativity is not very well understood. What is particularly interesting is that for a number of systems the cooperativity arises on passage from the molten globule to the native state rather than from the collapse to compact intermediates from the denatured state.[15,16] Such compact intermediates appear to have a substantial amount of secondary structure, have a volume about 50% larger than native, and have side chains that are not yet locked into place. Thus, the fixation of side chains accompanying the transition to the native state appears involved in the cooperativity observed in protein folding; however, the molecular origin of this process is still not understood.

During the past few years, in order to develop the ability to predict the native structure of proteins, we have developed a series of lattice models of increasing resolution.[11–14] In contrast to the very idealized lattice systems commonly used in studies of protein folding thermodynamics,[6,9,10] these high coordination lattice models, while still computationally tractable, can quite accurately reproduce the geometry of real polypeptide chains.[17] The lattice models used in these studies enable protein main chain atoms to be represented with an accuracy in the range of 0.7 Å RMS (root mean square deviation) from the crystallographic coordinates of the Cα trace.[17] However, most importantly from the point of view of the stud-

Received May 9, 1996; accepted May 9, 1996.
Address reprint requests to Dr. Jeffrey Skolnick, Department of Molecular Biology, The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, CA 92037.

ies described here, the model side chains are comprised of rotamers, and not just a single site rigidly attached to the backbone. Use of a single united atom, multiple rotamer representation of side groups, permits some details of side chain packing to be treated with an accuracy that is estimated to be on the level of 2–3 Å RMS from the native state. This estimate emerges from an analysis of the model geometry, from the quality of the predicted structures of small globular proteins, and from dynamic stability tests, or target folding, for larger globular proteins. Thus far, the best structures obtained in the de novo folding simulations[11–14] have an RMS in the range of 2–3 Å. Target folding driven by the model force field and a small to moderate number of NMR-type restraints leads to a 3–4 Å (Cα RMS) accuracy in larger proteins.[18] Unfortunately, these moderate resolution models are too complex for the detailed analysis of the thermodynamics of the protein folding process by exact enumeration of all (or at least all compact) conformational states, as was done for more highly idealized models (reviewed in Dill et al.[6]).

Recently, Hao and Scheraga[19,20] undertook a series of very interesting studies on the thermodynamics of protein folding. The model they employed was substantially more complex than those for which the exact enumeration of compact states is possible. Nevertheless, a version of multicanonical Monte Carlo sampling,[21,22] the so-called entropy Monte Carlo sampling (ESMC) method proposed in their work, allows for a quite exact thermodynamic description of the model system.[19,20] In these models, there is a single rotamer for each side chain. They demonstrated that some sequences, having a well-defined pattern of hydrophilic and hydrophobic residues, exhibit a cooperative all-or-none transition to a well-defined nativelike state. Other sequences, whose amino acid pattern is less consistent with the global fold, undergo a smooth collapse to compact globular states. They also demonstrated that the cooperativity of the folding transition results from proper long-range interactions, while the short-range interactions contribute substantially to the stability of the native state.

Inspired by these studies, in this paper we examine the thermodynamics of the folding process of an even more complex and, we hope, an even more realistic, protein model. The model is based on a high coordination lattice discretization of the conformational space of globular proteins and employs various potentials of mean force that may mimic a variety of physical interactions that control protein folding.[11–14,23] Among the questions addressed in the present studies are the following: Is the transition from the manifold of random coil conformations to the globular state all-or-none? What is the free-energy gap separating these states? What is the nature of the transition state? If the resulting transition is not all-or-none, how can the cooperativity of the folding process be augmented in these model systems to make it similar to real proteins? With regard to the last question, we explicitly examine the effect of multibody side chain interactions. These multibody interactions are expressed in the form of knowledge-based potentials of mean force that reflect the side chain packing preferences seen in real proteins. What is the physical meaning of these explicit cooperative terms? The first possibility is that the reduced, single sphere representation of the model protein side chains introduces a bias toward more liquidlike, nonspecific packing. If so, then the multibody term would work as a correction to the pairwise interactions, regularizing the side chain packing patterns; that is, it is necessary to fix problems with a reduced model. Another possibility is that the regular packing of protein side chains seen in the native state cannot be simply achieved in any model (even in one with atomic detail) of protein dynamics and structure without the inclusion of higher order multibody interactions. We discuss these possibilities in detail in the following sections of this paper.

The ESMC computations were done for a designed sequence that adopts a six-stranded (two-sheet), Greek-key minimal β-barrel fold for which the factors influencing the design of this sequence have been previously described.[23] The sequence is as follows (amino acid sequences of particular strands are listed on separate lines):

1. Gly-Val-Asp-Val-Asp-Val-
2. Gly-Gly-Gly-Val-Asp-Val-Asp-Val-
3. Gly-Gly-Phe-Arg-Phe-Arg-Val-
4. Gly-Gly-Gly-Val-Arg-Phe-Arg-Phe-
5. Gly-Gly-Val-Asp-Val-Asp-Val-
6. Gly-Gly-Gly-Val-Asp-Val-Asp-Val

Strands 1, 4, and 5 are expected to form the first β sheet, while strands 2, 3, and 6 form the second sheet. The loop/turn regions are composed of flexible Gly connectors. Val and Phe residues are expected to form the hydrophobic core, while the hydrophilic sides of the globule contain charged amino acids. In previous work,[23] we demonstrated that the model protein reproducibly folded to the desired native state. Between independent folding simulations, the resulting structures differ by about 3 Å RMS for the Cα trace. Using a simulated thermal annealing protocol, the folding process was simulated by a Metropolis scheme. In this work, we examine the thermodynamics of the folding process of the original model, as well as two updated models which include higher order multibody side chain interactions. The results of these investigations address, at least partially the questions raised above, and thus provide additional insights into the nature of the protein folding phenomenon.

The outline of the remainder of this paper is as

follows. We start with a summary of the protein model, which includes a discussion of the multibody potentials. Then, we outline our realization of the ESMC technique that essentially follows that given by Hao and Scheraga.[19,20] However, we introduce one important technical update that makes this relatively complex system computationally tractable and which involves the use of simulated annealing trajectories as a conformational pool for the ESMC procedure. Then, in the Results section, we analyze various aspects of the thermodynamics of the folding process and the uniqueness of the low-energy state. The Conclusion section summarizes our results and describes the physical picture for the origin of protein folding cooperativity that emerges from this study.

METHOD

For the reader's convenience, we begin with a brief overview of our geometric realization of a protein. Next, three models of the force field are introduced. The first one, model I, is essentially the same as that used previously, except for a somewhat different treatment of the hard core repulsions necessitated by the ESMC procedure.[23] More important are the updates of the two remaining models (model II and model III) of the force field, where we introduce explicit cooperative terms into the long-range interactions. This section concludes with the outline of the ESMC sampling method.

High Coordination Lattice Protein Model

The lattice model of proteins employed in this work has been previously described in detail.[23] Very similar models have also been used in test predictions of the native folds of small globular proteins[11,13] and small multimeric proteins.[14] The Cα trace of the model polypeptide is confined to an underlying simple cubic lattice with a mesh size equal to a = 1.22 Å. The Cα backbone is a chain composed of vectors a·**v** with **v** belonging to the following set {**v**} = {(1,1,1), ..., (3,1,0), ..., (3,0,0), ..., (2,2,1), ..., (2,2,0), ...}. Allowing all possible permutations of coordinates, the set consists of 90 basis vectors. However, due to the angular limitations of successive virtual bond vectors introduced to reproduce those seen in real proteins, the number of possible continuations of the Cα trace, given a conformation of the preceding residue, is about three times smaller.

In Figure 1, we illustrate the main geometrical features of the model employed here. The lattice-confined Cα trace provides a convenient reference frame for the definition of peptide bond atoms and pseudoatoms representing the side chains. The peptide bond coordinates of a given Cα–Cα chain segment are quite well defined by the geometry of three consecutive Cα trace vectors (the average error is on the level of 0.1–0.3 Å) and are stored in a large array. Consequently, during the simulation and after a conformational update, the reconstruction of the backbone atoms requires just a few arithmetical operations.[24,25] Moreover, one can easily define the model of main chain hydrogen bonds employing the peptide bond atoms.[23] Similarly, the side chain pseudoatoms (having a single interaction center for the entire side group) can adopt a discrete set of positions with respect to the main chain. This way the model mimics the various rotational isomers of the side groups. The centers of interaction for the glycines are located near the corresponding Cα carbons. The alanine side chain always has one rotamer. Rotamers for longer side chains are represented with an accuracy of 1 Å with respect to the center of mass of the side chain rotamers seen in the database. The centers of interaction for the side groups are not confined to the lattice; however, the lattice backbone is used as a reference frame. This geometric organization of the model decreases the cost of geometrical transformations by two orders of magnitude with respect to an equivalent off-lattice model and is the reason why a lattice representation is used.

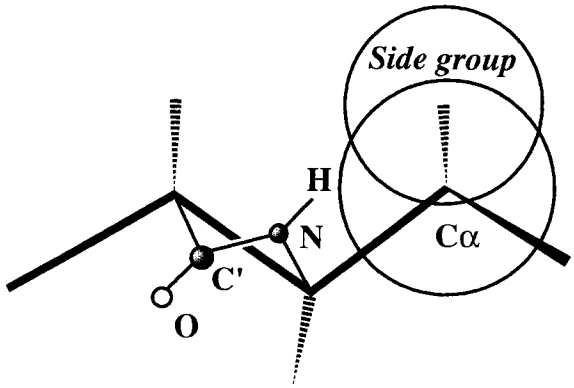

Fig. 1. Schematic illustration of the model chain geometry. The Cα trace vectors (*thick lines*) are confined to a high coordination lattice. The position of the peptide bond plate is defined by the geometry of three consecutive, virtual Cα backbone vectors. The *dashed lines* indicate the vectors from the Cα center located at the repulsive core to the center of the side group united atom (for a given rotational isomeric state).

The conformational updates of the model chain involve side group rotamer modifications with a fixed main chain, as well as local and global modifications of the chain geometry. There are three types of local backbone rearrangements: two Cα bond moves, three Cα bond moves, and four Cα bond moves. All are accompanied by the corresponding rotamer updates. Global chain rearrangements are constructed by random changes of a single Cα-Cα bond (which is accompanied by the appropriate shift of all coordinates of the N-terminal or C-terminal part), collective motions of a randomly selected larger part of the chain by application of a series of three bond rearrangements, and rotations (subject

to lattice restrictions) of a randomly selected part of the chain. The most global updates, specific to the ESMC sampling method (see the section on ESMC procedure below), involve the periodic reading of the coordinates of the entire chain from a conformational pool that contains a large set of representative conformations ranging from random coils to completely folded states.

Contributions to the Potential in Models I, II, and III

Here, we outline the various components of the model force field present in all three models. Particular terms in the interaction scheme simulate the internal conformational energy of the side chains, short-range interactions reflecting conformational stiffness of polypeptides and sequence-specific secondary structure propensities, hydrogen bonding energy and cooperativity of the hydrogen bond network, and a centrosymmetric burial energy of the side groups and their pairwise contact energy. This scheme is very similar to that used previously.[23] The only important update is related to the treatment of the excluded volume of the chain.[19] Instead of strong repulsive interactions on the order of $4k_BT$ (k_B is Boltzmann's constant, and T is the absolute temperature) that were used in previous work, here, we implement a hard core (infinite) repulsion at the previously used cutoff distances. The parameters of the potentials defining the model force field are available via anonymous ftp[26], or upon request to the authors. Another small update entails the neglect of one of the generic terms involving consecutive triplets of virtual Cα-Cα bond vectors (see equation 3 in ref. 23). We found that this term could be eliminated by somewhat stronger contributions from other generic terms of the force field. All contributions to the model force field are listed below.

First, there is rotamer energy (E_r) that reflects the average frequency of particular rotameric states. Rotamers are simulated in the model by a discrete set of (single united atom) side chains whose positions are dependent on the backbone conformation.[12,23] Then there are two sequence-specific contributions that simulate the short-range interactions, that is, secondary structure propensities and the local conformational stiffness of polypeptide chains, and depend on the pairs of amino acids A_i, A_j. The first term depends on the local geometry of the chain expressed by backbone vectors, E_s,

$$E_s = \Sigma\epsilon(A_i, A_{i+1}, r^{2*}_{i-1,i+2}) \quad (1)$$

with

$$r^{2*} = \text{sign}\,[(\mathbf{v}_{i-1} \times \mathbf{v}_i) \cdot \mathbf{v}_{i+1}]r^2_{i-1,i+2} \quad (2)$$

and

$$r^2_{i-1,i+2} = (\mathbf{v}_{i-1} + \mathbf{v}_i + \mathbf{v}_{i+1})^2 \quad (3)$$

and on the angular correlations between the side chains

$$E_{\text{sg-local}} = \Sigma\epsilon_k[A_i, A_{i+k}, \cos(\Theta_{i,i+k})]\ k = 1, 2, 3, 4 \quad (4)$$

$\Theta_{i,j}$ is the angle between $\mathbf{b}_i$ and $\mathbf{b}_j$, where $\mathbf{b}_i$ is the vector from the ith Cα to the ith side chain center of mass. In all equations, the summation, Σ, is along the chain.

Additionally, there is a generic term that "normalizes" the distribution of intermediate distances in the chain, thereby providing a bias toward compact helical states and expanded β type or coil states.

$$E_\eta = \Sigma\eta_i(r^2_{i-2,i+2}) \quad (5)$$

where

$\eta_i = -1$ for $r^2_{i-2,i+2} < 35$

$\eta_i = -1$ for $r^2_{i-2,i+2} > 75$

$\eta_i = 0$ otherwise

All distances are in lattice units. One lattice unit equals 1.22 Å.

Local conformational stiffness is additionally enforced by the preferred mutual orientation of the peptide bond plates (here we assume planar peptide bonds).

$$E_p = \Sigma\{\cos(\mathbf{h}_i, \mathbf{h}_{i+2}) + \cos(\mathbf{h}_i, \mathbf{h}_{i+4})\} \quad (6)$$

with $\cos(\mathbf{h}_i,\mathbf{h}_j)$ denoting the cosine between the ith and jth vectors defining the orientation of the peptide bond plates (the vectors between the amide hydrogen and the carbonyl oxygen). As was previously discussed (see also ref.23), the positions of the backbone atoms are well defined by the local geometry of the Cα chain. It should be pointed out that the above described factorization of local secondary structure propensities and the proper conformational "stiffness" of polypeptides reproduces the secondary structure (measured by the local geometrical criteria) with reasonable accuracy, and is comparable to the accuracy of standard secondary structure prediction methods (see ref.24 and references cited therein).

The readily accessible explicit positions of the peptide bond atoms enable the straightforward modeling of the hydrogen bond interactions[23,25]:

$$\varepsilon_{\text{H-bond}} = q_H(1 - f_H)/(r_{O,H} + 2\exp(-r^2_{O,H})) \quad (7)$$

where f_H, the angular factor, is of the following form:

$$f_H = [0.77 - \cos(\mathbf{r}_{Oi,Hj}, \mathbf{r}_{Oi,Hi})]^2 + [0.77 - \cos(\mathbf{r}_{Oi,Hj}, \mathbf{r}_{Oj,Hj})]^2. \quad (8)$$

q_H is an arbitrary scaling factor for the strength of the hydrogen bond that implicitly accommodates partial charges, local dielectric constant, and so on;

$\mathbf{r}_{Oi,Hj}$ is the vector between the oxygen in peptide plate j and the hydrogen in peptide plate i; the numerical value of 0.77 corresponds to the most probable angular geometry of the main chain hydrogen bonds in polypeptides. The model for hydrogen H bonds coincides nicely (almost all main chain bonds are the same in the context of both methods) with the DSSP definition.[27] Our model of the H-bond network is cooperative, and the strength of the network increases by 0.20 q_H every time two bonds form a parallel pattern that is characteristic of either β sheets or α helices. The H bond contribution is doubled when the number of H bonds per residue is equal to two per peptide unit; this further enhances the cooperativity of the network. We note that if the H bond indices are associated with peptide bonds rather than with the corresponding residues, then one obtains the same hydrogen patterns for both parallel and antiparallel sheets. This purely technical trick[23] allows for the above uniform treatment of the cooperativity of all regular structural elements in globular proteins.

The sequence-specific long-range interactions arise from a one-body centrosymmetric potential and from a pairwise potential. All the one-body and two-body interaction parameters are derived from the statistics of the protein structural database[28,29] and are available via anonymous ftp.[26] The one-body contribution reads:

$$E_1 = \Sigma\epsilon_1[r(A_i)/S_0] \tag{9}$$

with

$$<S>_0 = 1.8n^{0.38} \text{ (in lattice units)} \tag{10}$$

where $<S>_0$ is the expected radius of gyration of the native state of a single-domain protein consisting of n amino acids. Equation 10 has been derived from the statistics of known structures of single-domain proteins. $r(A_i)$ is the distance of the center of mass of the ith side group from the center of mass of the entire chain. The potential is used in the form of amino acid-specific histograms.

The pairwise interactions of the side chains are expressed in the form of a square-well contact potential. For attractive pairs of side groups, the interaction strength is moderated by an angular factor, f, that reflects the preferred packing angles of interacting polypeptide fragments.

$$E_{ij} = \begin{cases} \infty, \text{ for } r_{ij} < R_{ij}^{rep} \\ \epsilon_{ij}, \text{ for } R_{ij}^{rep} < r_{ij} < R_{ij}, \text{ and } \epsilon_{ij} > 0 \\ f\epsilon_{ij}, \text{ for } R_{ij}^{rep} < r_{ij} < R_{ij}, \text{ and } \epsilon_{ij} < 0 \end{cases} \tag{11}$$

$$f = 1.0 - \{\cos^2(\mathbf{u}_i, \mathbf{u}_j) - \cos^2(20°)\}^2 \tag{12}$$

where R_{ij}^{rep} and R_{ij} are the cutoff values for hard core excluded volume interactions and for square-well, soft pairwise interactions, respectively. R_{ij}^{rep} equals the average contact distance (obtained from the statistics in a structural database for amino acids i and j) minus two standard deviations of this value. R_{ij} is equal to the average contact distance plus three standard deviations. The vectors $\mathbf{u}_i$ define the local direction of the main chain. For amino acid i, $\mathbf{u}_i = \mathbf{r}_{i+2} - \mathbf{r}_{i-2}$, where $\mathbf{r}_k$ is the coordinate of the kth Cα.

The scaling of the various terms is necessary due to the statistical origin of the potential. The strength of the particular interactions ensures a reasonable balance between the short- and long-range interactions and must reproduce the low level of secondary structure in the denatured state characteristic of real proteins.[11-14,23,24] The scaling factors of the particular contributions to the force field are as follows: rotamers (1), peptide plate correlations (1), sequence-specific, backbone short-range correlations (2), angular side chain correlations (1), generic backbone regularizing term (1), hydrogen bond network (2.5), one-body long-range (1), and pairwise long-range interactions (0.5). The different (stronger 2 versus 1) scaling with respect to the previous work[23] of the sequence-specific backbone short-range correlations results from omitting a generic term of the same type.

Four-Body, Cooperative Side Chain Interactions

The force field described above has been used in the series of computations in model I. For the two other series of calculations related to model II and model III, explicit cooperative terms for side chain interaction have been introduced. These terms are of the type:

$$E_4 = \sum_i \sum_{j>i} (\epsilon_{ij} + \epsilon_{i+kj+n}) C_{ij} C_{i+kj+n} \tag{13}$$

with $|k| = |n|$. $C_{ij} = 1$, when side groups i and j are in contact; otherwise $C_{ij} = 0$.

For model II, the value of n was assumed to be equal to 3 and 4. This means that if there are simultaneously contacts between residues i and j and between residues $i + n$ and $j + n$ (also $i + n$ and $j - n$, or $i - n$ and $j - n$, or $i - n$ and $j + n$), then the system has additional energy contribution equal to the sum of the corresponding pairwise interactions. The idea is illustrated in Figure 2A. The $n = 3$ and 4 repeat is characteristic of the helix-to-helix pattern of side group contacts as well as the pattern seen in the β sheets (and less frequently between the sheets). Because our pairwise interaction parameters are negative for contacts of similar residues (hydrophilic residues with hydrophilic residues, and hydrophobic with hydrophobic residues), in native proteins, these sums are usually negative.

Previously, it has been shown that this kind of cooperative term facilitates (when appropriate) a process similar to "side chain fixation" that is characteristic of the transition from the molten globule state to the native state.[11] It should also be pointed out that this contribution to tertiary interactions

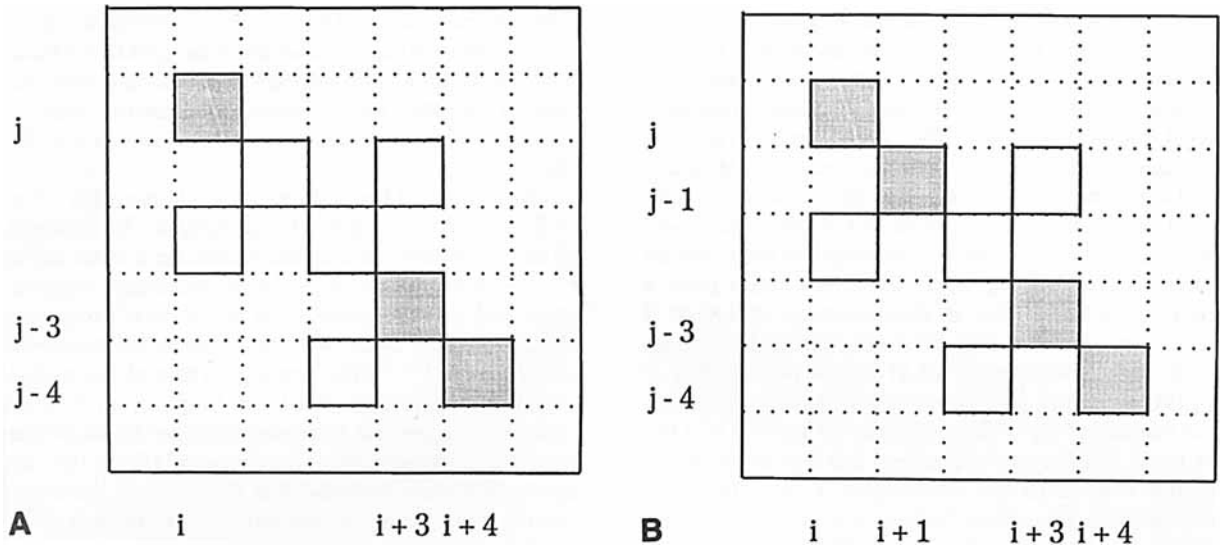

Fig. 2. Illustration of cooperative four-body side chain interactions for model II (**A**) and model III (**B**). The *shaded squares* represent the reference contact $i - j$ and the set of contacts that contribute to cooperative interactions with pairwise contact $i - j$. The underlying pattern of contacts corresponds to the ideal antiparallel β-type packing of side groups.[38]

does not automatically enforce nativelike packing and side chain fixation. For example, in two sequences designed by DeGrado and coworkers[30–32] that were previously studied using a similar model[11] in agreement with experiment, we found that one of these sequences adopted unique side chain packing, while the other remained in the molten globular state. Rationalization of such behavior emerges from the inspection of these two designed sequences. The sequence with a uniform, leucine-based, hydrophobic core forms a four-helix bundle, which, while very stable, does not exhibit side chain fixation. Indeed, in such a case, there are many helix-to-helix packing patterns that give the same cooperative contribution. For the sequence in which some leucines were replaced by other hydrophobic residues, the packing arrangements involving different side chains possessed different energies, and the degeneracy of side chain packing is broken. Consequently, the native state packing becomes unique. This empirical success argues that the choice of the magnitude of cooperative multibody contribution to the tertiary interaction as a sum of corresponding pairwise interactions is not an unreasonable first-order approximation to what is undoubtedly a more complex process.

In model III, additional cooperative terms are added, see Figure 2B. Besides patterns involving $n = 3$ and $n = 4$ repeats, an $n = 1$ repeat was also introduced into the four-body interactions. The last repeat is rather specific for the side chain packing patterns seen in β-type proteins, but it is rarely seen in helical structures.

In order to maintain the same balance between the short-range and long-range tertiary interactions in all the models, the scaling of pairwise interactions (and consequently, the four-body interactions) in model II and model III must be modified. Instead of a 0.5 scale factor (see the previous section) for the pair terms used in model I, the corresponding scaling is 0.25 and 0.2 from model II and model III, respectively, for both the pair and multibody components. This way, the thermodynamic characteristics of the three particular models could easily be compared. The magnitude of the tertiary interactions in all cases is the same, while the cooperative contribution changes substantially.

Entropy Monte Carlo Sampling Method

In the well-known Metropolis Monte Carlo (MMC) sampling methods, conformational space is randomly sampled according to the equilibrium Boltzmann distribution of (distinguishable) conformations:

$$P_i = \exp(-E_i/k_B T) \tag{14}$$

The resulting transition probability $p_{i,j}$ from the "old" conformation i to the "new" conformation j (for the asymmetric scheme) is controlled by the energy difference $\Delta E_{i,j} = E_j - E_i$:

$$p_{i,j} = \min\{1, \exp(-\Delta E_{i,j}/k_B T)\}. \tag{15}$$

Thus, the method is sensitive to the presence of energetic barriers.

The ESMC method was originally proposed by Lee[33] in the context of a simple Ising model, and more recently, it has been applied to the study of protein models by Hao and Scheraga.[19,20] ESMC is similar in spirit to the multicanonical MC technique of Berg and colleagues[21] that was recently used by Hansmann and Okamoto[22] in their studies of small peptides. Here, since the Hao-Scheraga formulation

appears to be the most straightforward, and their work deals with models similar (however, of lower resolution and with a simpler force field) to those used in this work, we follow their formalism.

In contrast to the MMC method, the ESMC method generates an artificial distribution of states that is controlled by the conformational entropy as a function of the energy of particular conformations:

$$P_i = \exp[-S(E_i)/k_B]. \quad (16)$$

In practice, E_i corresponds to a small energy interval (later called "a state"). The transition probability can be formally written as:

$$p_{ij} = \min\{1, \exp(-\Delta S_{ij}/k_B)\} \quad (17)$$

with ΔS_{ij} being the entropy difference between states i and j, respectively.

Of course, the entropy is usually not known at the beginning of the simulation. It is easy, however, to show that one can iteratively find an estimate, $J(E)$, of the entropy $S(E)$, using a density of states (energy histogram), $H(E)$. The kth iteration consists of an ESMC simulation run with $S(E)$ approximated by $J_{k-1}(E)$. The simulation produces a histogram, which is subsequently used to update $J_k(E)$ by

$$J_k(E) = J_{k-1}(E) + \ln[\max\{1, H_k(E)\}]. \quad (18)$$

After a sufficient number of runs, all the states (energy intervals) become sampled with the same frequency. This produces a flat histogram of $H(E)$, and the curve of $J(E)$ + const approaches the true $S(E)$ curve.

The energy interval used for the definition of the histogram $H(E)$ is assumed here to be equal to 4.2 k_BT. This value is relatively large in comparison with previous work.[19] On the other hand, the energy of the system studied here varies over a range that roughly spans -350 k_BT to 0 k_BT. Moreover, as shown in the Results section, states separated by few k_BT in many cases have very similar conformations. Thus, this coarse-grained histogram accelerates sampling with respect to a finer grid, without a substantial penalty in conformational resolution.

It is clear from the above formulation that the method can easily surmount local energy barriers; however, in ESMC, entropic barriers become important. The best way to avoid the situation where the sampling process is trapped by entropic barriers (these mainly result from the different conformational degeneracy of various energy levels) is to use a "conformational pool" to randomly shift the system between distant energy levels. This way the sampling process surmounts possible entropic barriers. The conformational pools used in the present work come from the simulated annealing MMC simulation described elsewhere.[23] Various sparse trajectories from the folding experiments were combined in such a way that the ESMC procedure converged after a long series of computations. The convergence is assumed when the change of the relative entropy becomes independent of the energy within a tolerance of 0.3 k_B. This condition obtains except over the very narrow range of the lowest energy states of very low degeneracy, where large fluctuations persist, and obviously, here the simulation has not fully converged.

The sampling procedure for model I took several weeks of computation on a dedicated HP-735 workstation. The cost of computation for models II and III was substantially lower; here, the ESMC procedure uses as the first approximation to the entropy the final result for model I. This way the entropy estimate, $J(E)$, in Eq. 18 was not constructed from scratch. The models are sufficiently similar (similar structure of the force field and similar strength of the long-range interactions) to enable such a procedure to work.

RESULTS

Effect of Cooperative Side Group Interactions on Folding Thermodynamics

The entropy-driven Monte Carlo sampling method provides the relative numerical values of the entropy of the system as a function of the energy. The constant term (see Eq. 18) has no physical meaning and is dependent on the number of ESMC steps in all iterations and the details of implementation of the method. The simulations performed for all three models were very long due to the relatively complex models and the size of the system. In Figure 3, we compare the final profiles of J (the entropy estimate) for the three models studied. The constant term is the smallest for model I. This does not mean that the simulations for this model were shorter than for the remaining models. The difference is due to the procedure we used to lower the computational cost. For models II and III, the initial estimates of J were taken from the final results for model I. Hence, since additional iterations are performed, one might expect the constant term to be larger in models II and III.

The $J(E)$ curves of the model systems show a number of interesting features. First, in all cases there is no single ground (nativelike) state. Instead, there are manifolds of low-energy folded conformations (we will examine the problem of conformational uniqueness separately below) of differing entropy. These low-energy regions are indicated by the very steep regions of the curves. Second, while the curve for model I is almost of the same slope over the entire relevant conformational energy range, for model II, a slight concave region is clearly visible. The plot for model III is more concave, suggesting the existence of two different, well-defined states at equilibrium. However, the most direct way to assess this is to plot the free-energy versus energy profiles; these are presented below.

Other thermodynamic characteristics of the model

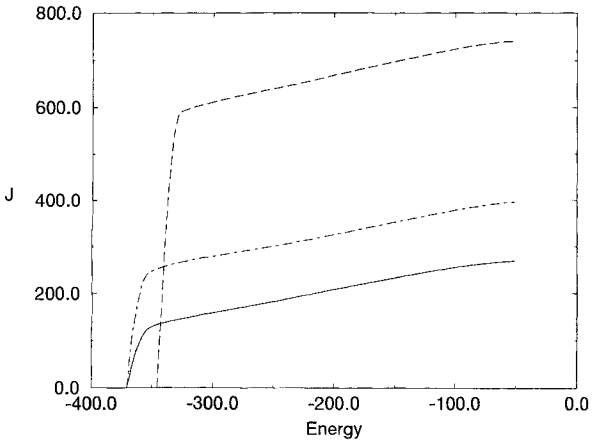

Fig. 3. Final estimates of the relative entropy J as a function of configurational energy E for three models (in model II, model III, and model I, from the top curve to the bottom curve, respectively) of side chain interactions. Different shifts (constant term) for various models reflect different initial guesses as for the entropy (see the text for more details).

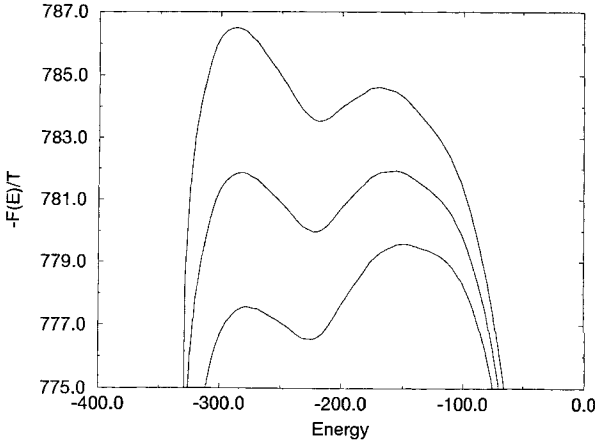

Fig. 4. In model II, free-energy, $F(E, T)/T$, plots versus energy, E, at various temperatures.

are derived easily from the $S(E)$ versus E profiles. At any temperature, one may obtain the free energy of the system as a function of energy from

$$F(E, T) = E - TS(E) \quad (19)$$

Consequently, an average property, Q (e.g., the average energy $<E>$ or the heat capacity C_v) as a function of temperature could be calculated.

$$<Q(T)> = \sum Q(E) \exp(-F/k_BT) \Big/ \sum (-F/k_BT) \quad (20)$$

where the summation is over the various states characterized by the given energy interval δE.

Figure 4 shows the $F(E, T)$ profiles obtained for model II at a series of T values. At a certain temperature, T_c, the free energies of the low-energy and the high-energy states have the same values. This is the folding temperature T_c. The estimated numerical values of T_c for all three models are given in Table I. Figure 5A-C shows the free-energy profiles at $T = T_c$ for the three models of tertiary interactions. Since ESMC introduces a constant term into the estimate of the entropy, the numbers on free-energy axes only have a relative meaning. It is clear from inspection of Figure 5 that the two state model of the folding transition is very well pronounced in model III, is slightly weaker for model II, while the folding of model I is almost continuous. For model I, the free-energy barrier (taking into consideration the estimated transition temperature), which separates the folded and denatured states is rather small, and is 0.9 k_BT_c. This value is only a couple of times larger than the estimated error of the simulations. The free-energy barrier increases to about 2.0 k_BT_c in model II and is about 4.7 k_BT_c in model III, respectively. Thus, model I, lacking cooperative side chain packing interactions, has a quasi-continuous transition. On increasing the possibility of cooperative interactions, the conformational transition assumes an all-or-none character.

In addition to the presence of a free-energy barrier between the folded state and non folded states (we defer a discussion of the structural characteristics in various regions of the energy spectrum for a moment), the transition is much sharper in those models possessing the possibility of cooperative tertiary interactions. This is clearly demonstrated in Figure 6A-C where the energy and the heat capacity are plotted as a function of temperature. On passing from model I to model III, the energy curves become much steeper in the transition region, and the heat capacity peak becomes much sharper.

The analysis of the folding thermodynamics of these three model systems shows that the transition from unfolded states to compact globular states depends on the assumed form of the tertiary interactions. Although model I adopts (as do the remaining models) a well defined, folded state topology (see the work describing MMC simulations of the folding process[23]) that is the same as the Greek-key topology seen in many β type globular proteins, the cooperativity of the transition is low. This occurs in spite of the cooperative hydrogen bond scheme incorporated into the model and the exaggerated pattern of hydrophilic and hydrophobic residues.

As was suggested in our previous studies,[23] it is possible that the compact state of model I has more features of the molten globule than the unique native state. This could be responsible for the small free-energy barrier, and consequently, the low cooperativity of the folding transition. Nevertheless, it is the qualitative conclusion that cooperativity emerges when multibody interaction terms are included in the potential (as in models II and III), which is the most important result of these model simulations.

At this point, it seems worthwhile to point out that in order to obtain a stable folded state for model I by simulated annealing MMC sampling,[23] it was

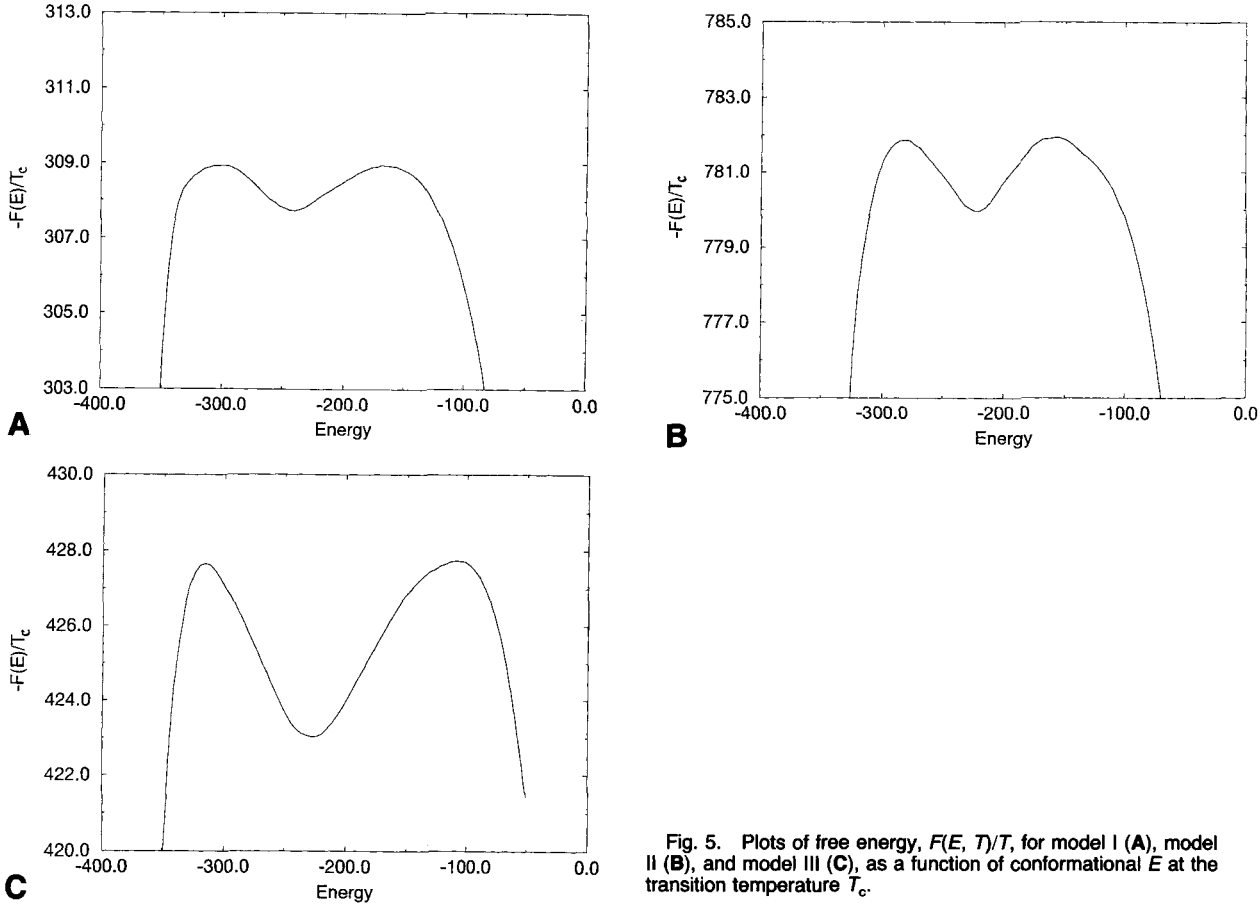

Fig. 5. Plots of free energy, $F(E, T)/T$, for model I (**A**), model II (**B**), and model III (**C**), as a function of conformational E at the transition temperature T_c.

TABLE I. Comparison of the Thermodynamic Characteristics of the Three Models for Side Chain Interactions

	Model I	Model II	Model III
$\epsilon_{ij}/\epsilon^0{}_{ij}$*	0.5	0.25	0.2
E_{min}†	$-365.6\ k_BT$	$-345.8\ k_BT$	$-366.3\ k_BT$
T_c‡	2.01	1.79	2.13
E_{min}/T_c§	−181.9	−193.2	−172.0
$\Delta F/T_c$¶	0.9	2.0	4.7

*Scaling of pairwise interactions of the side chains.
†Minimum conformational energy seen in the ESMC procedure.
‡Folding temperature.
§Reduced minimum energy.
¶Reduced free-energy barrier.

necessary to cool down the model system to temperatures below $T = 1.5$. This is considerably lower than the transition temperature $T_c = 2.01$ found by the ESMC procedure. This is the temperature at which the values of the system's free-energy in the two minima are the same (see Fig. 4). The different realization of the excluded volume interaction (presence of hard core instead of a strong repulsive force) used here cannot be responsible for such a large discrepancy in apparent folding behavior. The more likely explanation is that, due to an almost continuous transition, the temperature range where folded and unfolded conformations coexist is very broad. Consequently, to lock the system in the low-energy basin, one needs to substantially quench the molecule below its true transition temperature. This is again indicative of the low cooperativity of the folding transition exhibited by model I.

The situation changes dramatically for models with explicit cooperative potentials (model II and model III) of interactions between the side groups. The folding thermodynamics for these cases is well described by a very sharp, all-or-none transition that is consistent with experimental findings for globular proteins. Thus, these simulations indicate that in more realistic protein models, in agreement with experiment,[15] the cooperativity of the folding transition arises from the fixation of side chains. This fixation is only possible in these models upon introduction of cooperative side chain packing interactions.

Behavior of the Models at Various Energies

The two free-energy minima in all three models correspond to a spectra of random coil, denatured high-energy states and to low-energy, folded states. Figure 7 shows model polypeptide conformations at four values of the energy. The representative snapshots, with only the Cα trace displayed for the sake

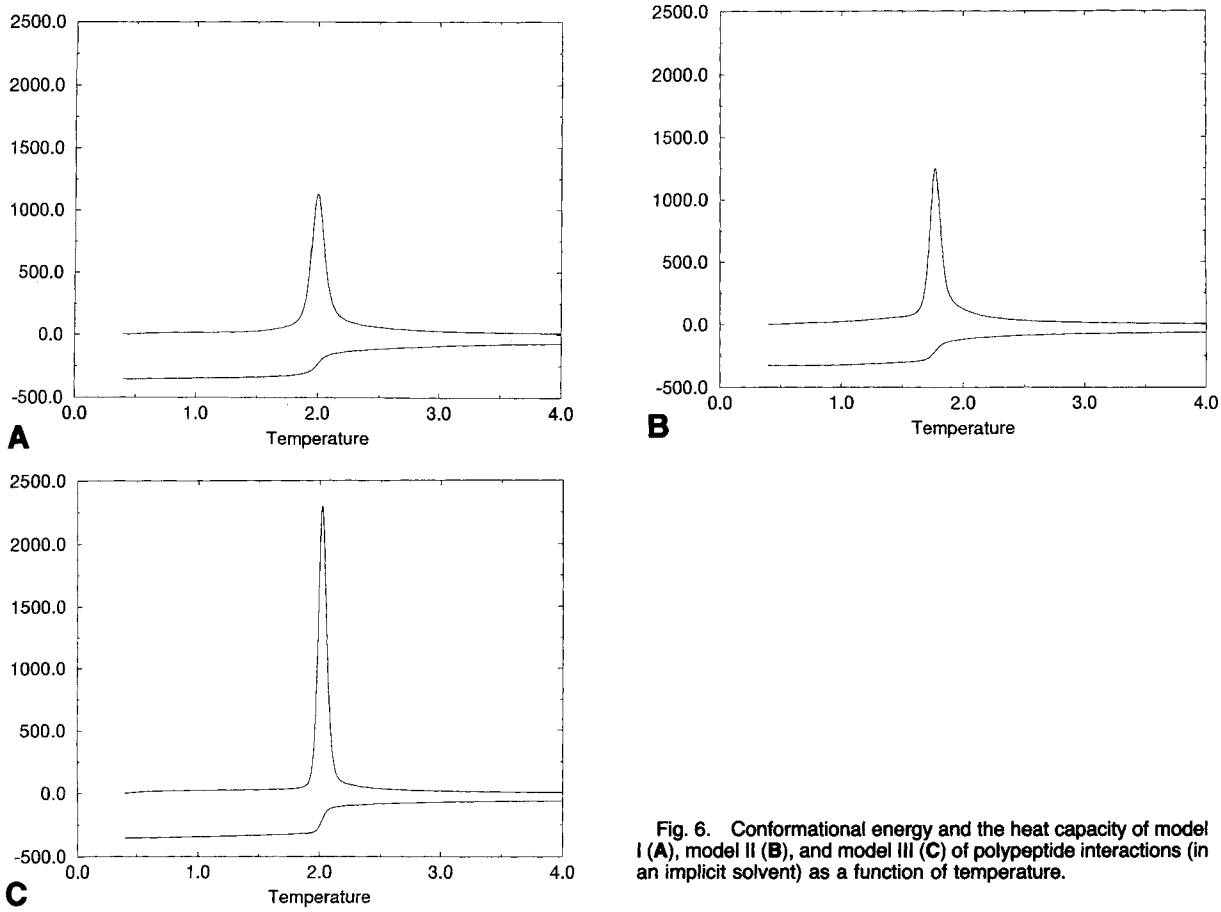

Fig. 6. Conformational energy and the heat capacity of model I (**A**), model II (**B**), and model III (**C**) of polypeptide interactions (in an implicit solvent) as a function of temperature.

of clarity, were taken from the ESMC procedure for model I. The high-energy conformations (Fig. 7A) have the character of an expanded random coil, with very little secondary structure and rather few, essentially random, contacts between the side chains. Decrease of the energy to the point that corresponds to the free-energy maximum changes the picture. Conformations at this energy (B) have a higher level of secondary structure, and there are more contacts. The global conformations are rather random, but are mostly inconsistent with the final fold. A different behavior is seen in models II and III, see below. Over a very broad energy range, the low-energy states have mostly correct folds, and dense packing of the side chains. Typical conformations at energies around -300 k_BT (C) differ from the lowest energy state (D) only by a slight perturbation of the hydrogen bond pattern and the side chain packing. Nevertheless, in energy region C, some misfolded states (in particular, those folds having the mirror image topology) could be seen, although rarely. When the energy approaches the minimum observed energy, the fraction of misfolded conformations diminishes to zero. Within a 15–20 k_BT energy range near the lowest energy state, misfolded structures were not detected. Above this range of energies, the energy spectra for globally correct folds (perhaps, some of whose conformations correspond to the molten globule state) and the energy spectra of misfolded compact structures partially overlap. We believe this is a general feature of the energy landscape of proteins and is not an artifact of this model.

A similar collection of various conformations, with decreasing energy of the systems, could be seen for model II and model III. In spite of the very similar features of unfolded and folded states, there are, however, substantial differences. The energy range for the folded structures becomes narrow on increasing the contribution of cooperative interactions. Moreover, states with intermediate energies, which correspond to partly folded structures, have a very low thermodynamic probability due to their different free-energy versus energy profiles. We now turn to a more detailed analysis of the properties of models II and III at the free-energy versus energy maximum.

Nature of the Transition State

The behavior of a number of properties of model II and III as a function of energy is further examined in Figures 8A–C and 9A–C, where we plot the average (±1 standard deviation) fraction of native contacts, native secondary structure and mean square radius of gyration $<S^2>$ in dashed (solid lines). As would be

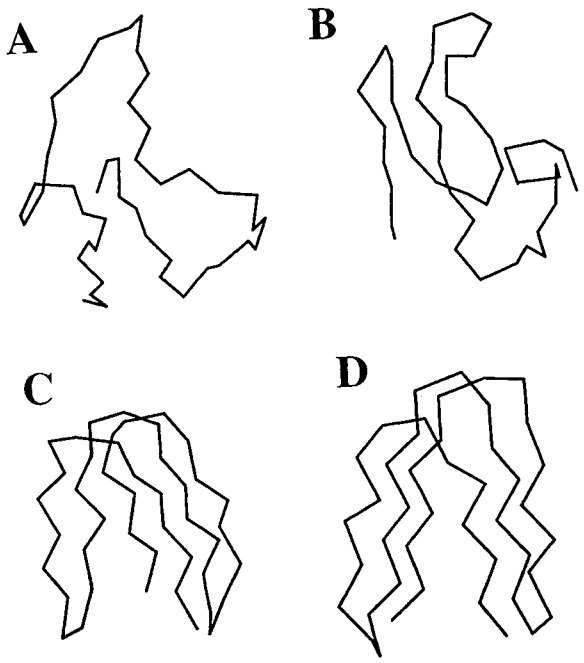

Fig. 7. Snapshots of representative conformations of the Greek-key sequence obtained from ESMC simulations of model I. **A**: at $E = -100.1$; **B**: at $E = -200.8$; **C**: at $E = -304.4$; and **D**: the lowest energy state at $E = -365.6\ k_BT$. For the sake of clarity, the side chains are omitted.

expected, these properties vary smoothly with decreasing energy. Note that the plots of native contacts and fraction of native secondary structure do not plateau at a level corresponding to 100%. The reason is the variability of the low-energy states, where the contact map overlap and the overlap of secondary structure (which is defined based on the particular binning of the main chain local geometry) are on the level of 70–80% (see also Table II). This is further discussed in the next section on the character of the folded conformation. Of particular interest is the behavior of these conformational properties at the transition state which occurs roughly at −225 kT in both models II and III. Of course, by the term transition state we refer to those conformations located at maximum of the free-energy versus energy curve. In model II, the transition state is comprised of structures having about 45% of the total number of native contacts, about 60% of the native state's secondary structure and$(\langle S^2\rangle^{\dagger}/\langle S^2\rangle_{native})^{3/2} = 1.5$. Similarly, in model III, in the transition state, there are structures with about 50% native contacts, 60% of the secondary structure and $(\langle S^2\rangle^{\dagger}/\langle S^2\rangle_{native})^{3/2} = 1.5$. This range of native like properties is qualitatively consistent with Kuwajima's description of the physical properties of the transition state of α-lactalbumin and Ca^{2+} binding parvalbumin.[15] Thus, the behavior of these models in the transition state also supports a critical substructure model. The activated state has a partial amount of native state secondary structure formed, but there is a manifold of such partially folded conformations having some, but not all, of the native state's structure.

Character of the Folded Conformations

Although we have demonstrated that the model possesses the same qualitative character as real systems, there remains the essential question: How unique is the very low-energy state? Does it exhibit the features of the native states of globular proteins? First, let us note that in all models the very low-energy state is not frozen, it still undergoes small fluctuations. These involve some small packing rearrangements, as well as the breaking and forming of one or two model hydrogen bonds. For a set of 10 low-energy states for each model, we performed an analysis of the fluctuations of the Cα backbone as well as the side chain packing. This set has been randomly selected from a large number of low-energy conformations generated during the ESMC procedure. In Table II, we compare the pairwise RMS (Cα trace) distances between pairs of these low-energy structures for all three models under consideration. In all models, the Cα backbone fluctuations for the low-energy states are small, within 3 Å RMS. Due to the relatively broad square well of the potential describing the interactions of the side groups, and due to the relatively permissive definition of the hydrogen bond interactions, the resolution of the model is on the level of 2–3 Å. The conformational fluctuations of the very low-energy states, characterized by the numerical values for the average RMS between pairs of independent structures, are exactly of this magnitude. Within the resolution of the model, all low-energy conformations of the main chain are the same for all three models of side chain interactions. It should be also pointed out that for the assumed energy range for the tested low-energy structures, there is no significant correlation between the pairwise (between two structures) RMS and the corresponding energy.

The side chain packing is relatively well defined in all three models. In Figure 10A–C, we compare representative side chain contact maps for the low-energy states for the three models. The maps show clear β-type contact patterns within the two sheets. Some additional contacts that result from the single sphere representation of the side chains contaminate the ideal pattern for antiparallel β structure, which is shown in Figure 2. Within a given model, the reproducibility of the contact maps (derived with the cutoff distances defined for soft pairwise interactions between the side chains) (see Eq. 11) is on the level of 70%, with a slightly larger standard deviation in model I as compared to models II and III. These contact maps represent very similar structures, and it should be pointed out that this level of

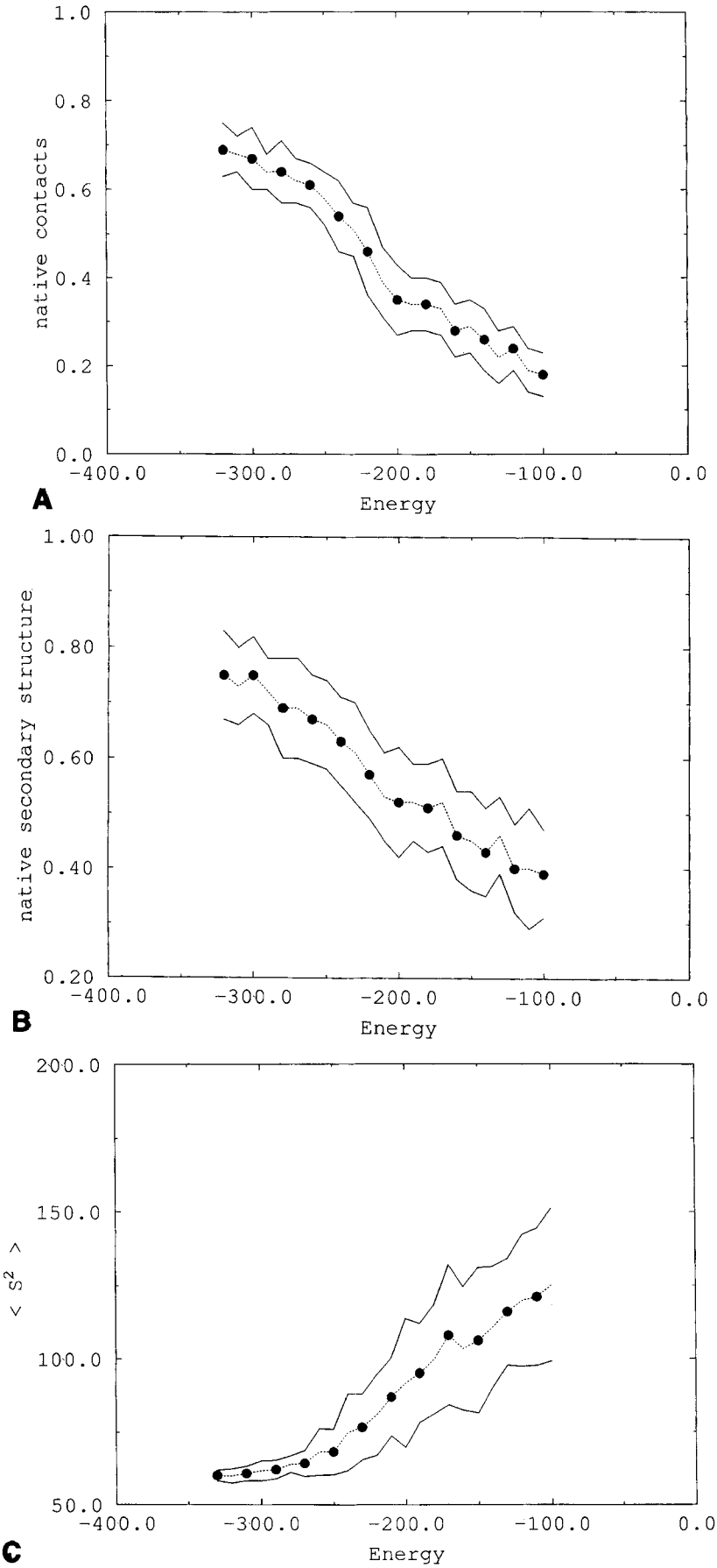

Fig. 8. Plot of the average (±1 standard deviation). **A**: Fraction of native contacts. **B**: Fraction of native state secondary structure. **C**: Mean square radius of gyration. $\langle S^2 \rangle$ in the *dashed* (*solid*) *curve(s)* versus energy for model II. The plateau value of the fraction of native contacts corresponds to the level of reproducibility of the contact maps for the low-energy states (see Table II). Similarly, the plateau value seen for secondary structure reflects the reproducibility of the model backbone conformations in the native free-energy basin.

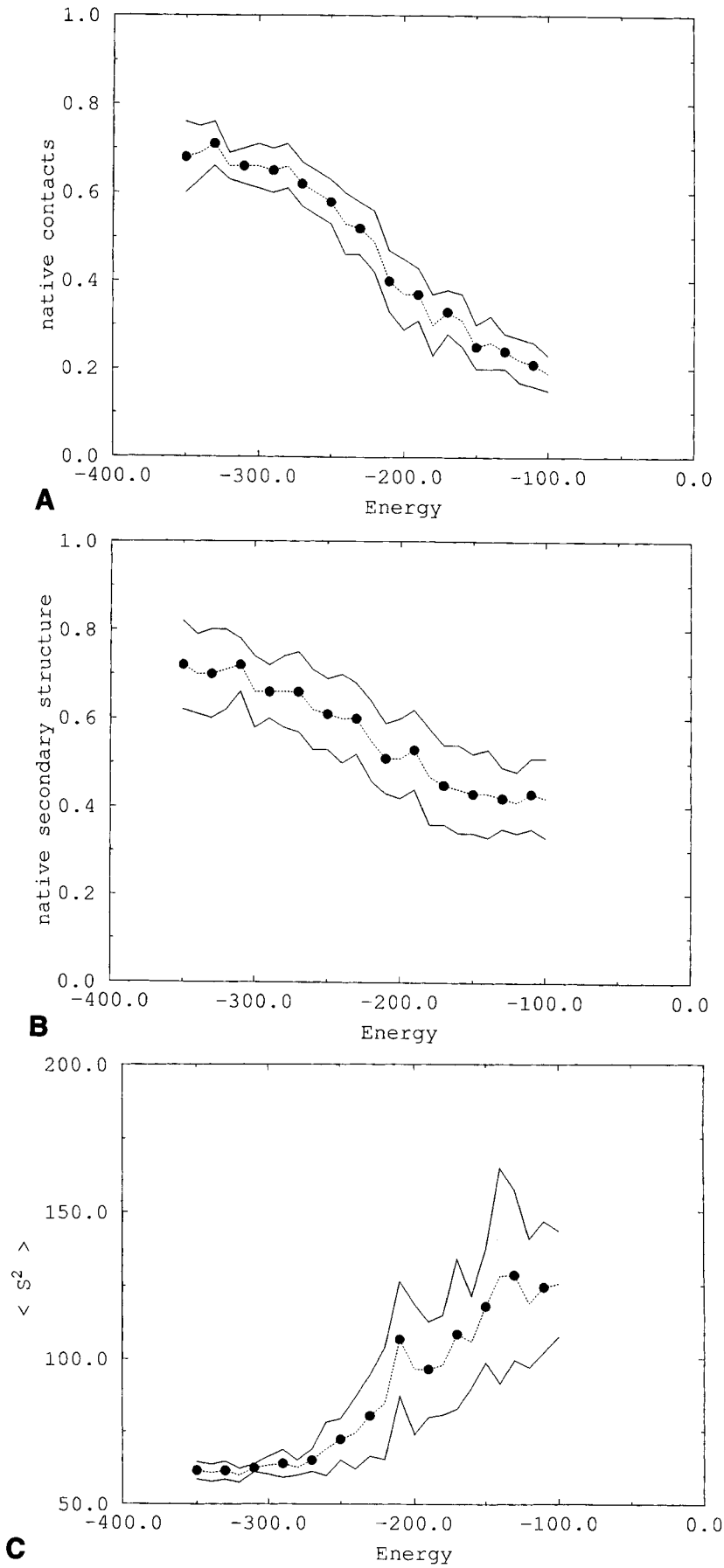

Fig. 9. Plot of the average (±1 standard deviation). **A**: Fraction of native contacts. **B**: Fraction of native state secondary structure. **C**: Mean square radius of gyration. $\langle S^2 \rangle$ in the *dashed* (*solid*) *curve(s)* versus energy for model III. The plateau value of the fraction of native contacts corresponds to the level of reproducibility of the contact maps for the low-energy states (see Table II). Similarly, the plateau value seen for secondary structure reflects the reproducibility of the model backbone conformations in the native free-energy basin.

TABLE II. Structural Characteristics of the Low-Energy States*

Model	Energy range	Cα RMS (in Å)	Contact map overlap (%)†
I	−366 to −358	1.89 ± 0.73	71 ± 11
II	−346 to −329	2.41 ± 0.65	67 ± 8
III	−366 to −351	2.40 ± 0.72	68 ± 8
I + II + III		2.23 ± 0.74	69 ± 9

*The statistics are based on 10 independent low-energy structures extracted from the ESMC simulations for each model. The average RMS and its standard deviation are based on comparison of all possible pairs.

†The contact map overlap is defined as $2n_{12}/(n_1 + n_2)$, where n_{12} is the number of common contacts and n_1 and n_2, are the numbers of contacts of the two maps that are being compared.

contact map agreement is also found in two structures of the same protein solved by two different groups or under different solvent conditions.

In all cases, there are well-defined hydrophobic clusters of the side chains, with Phe side chains having the same interaction pattern. The differences between contact maps mostly arise from the fluctuating numbers of contacts experienced by the loop Gly residues. The fluctuations in the number of contacts for Val residues and for the polar side chains are much less; however, some differences could be noticed, especially for contacts between the two sheets of the barrel. Consequently, the pairwise comparison of low-energy states (Cα trace RMS and the contact map overlaps) within the particular models and between the models indicates the existence of a relatively well defined compact structure that could be associated with the native state of real proteins.

Most importantly, in spite of the different treatment of cooperative side chain packing interactions, the lowest energy structures are the same in all three models. However, in the cooperative models, the higher energy, partly unfolded and misfolded states are characterized by a much higher free-energy in comparison to model I. Consequently, at equilibrium, the energetical and structural fluctuations in model I are much larger than in models II and II. In other words, ESMC indicates that the lowest energy states of model I are nativelike, and essentially the same as in the more cooperative models. However, there is a manifold of states of almost the same free-energy that has higher conformational energy and nonnative conformations. In contrast, the free-energy basin of nativelike conformations from the cooperative models is well defined. Due to the large free-energy barrier between native and (partially) unfolded states, the equilibrium population of intermediates, or partially folded states, becomes very small. This, then, is the origin of the different thermodynamic behavior of the folding transition observed in these models.

CONCLUSION

In this work, we applied the ESMC method to study the thermodynamics of protein folding on an example of a theoretically designed sequence, which adopts a Greek-key β barrel fold. The studies employed a high coordination lattice model of the protein and various models of side chain interactions, including explicit cooperative terms. The cooperative side chain packing terms incorporate various four-body correlations reflecting structural regularities seen in real proteins in their native state. We find that the inclusion of a cooperative side chain packing term significantly changes the thermodynamics of the folding transition from a more or less continuous transition to one having all-or-none character, and concomitantly, the low-energy "nativelike" state becomes better defined. Within this native free-energy basin, the backbone geometry and side chain packing fluctuations are small, and on the level of the resolution of the model, there is a unique native state.

At this point, it seems worthwhile to compare the present study with the previous ESMC study by Hao and Scheraga[19,20] on a somewhat simpler model. For exaggerated sequences, Hao and Scheraga have demonstrated a similar level of cooperativity such as that seen here for model II or model III. However, they did not employ any explicit cooperative side chain packing term. Our sequence is also exaggerated. Why then does our model require the cooperative side chain interaction in order to reproduce the cooperativity of the model polypeptide folding? The answer is not due to the somewhat larger number of conformational degrees of freedom in our representation of the main chain, nor to the more complex folding motif studied here. Rather, the qualitative difference between the two models that is responsible for the above effects is related to the uniqueness of the secondary structure preferences and to differences in side chain representation. Hao and Scheraga assume that there is a preferred extended conformational state for those residues that form β strands in the putative structure. Here, we use a statistical potential to encode for intrinsic secondary preferences. Even when an extended state is statistically preferred, the spectrum of backbone conformations compatible with such a state is much broader. Perhaps even more importantly, they employ a fixed orientation of the side groups with respect to the main chain backbone. By contrast, our model allows for multiple side group conformations, which corresponds to the spectrum of rotational isomeric states of the side chains of real polypeptides. Thus, because of the more constrained specification of the native backbone geometry and position of the side groups, the conformational entropy accessible to the folded conformation is much smaller in the Hao-Scheraga model than in the present case. This

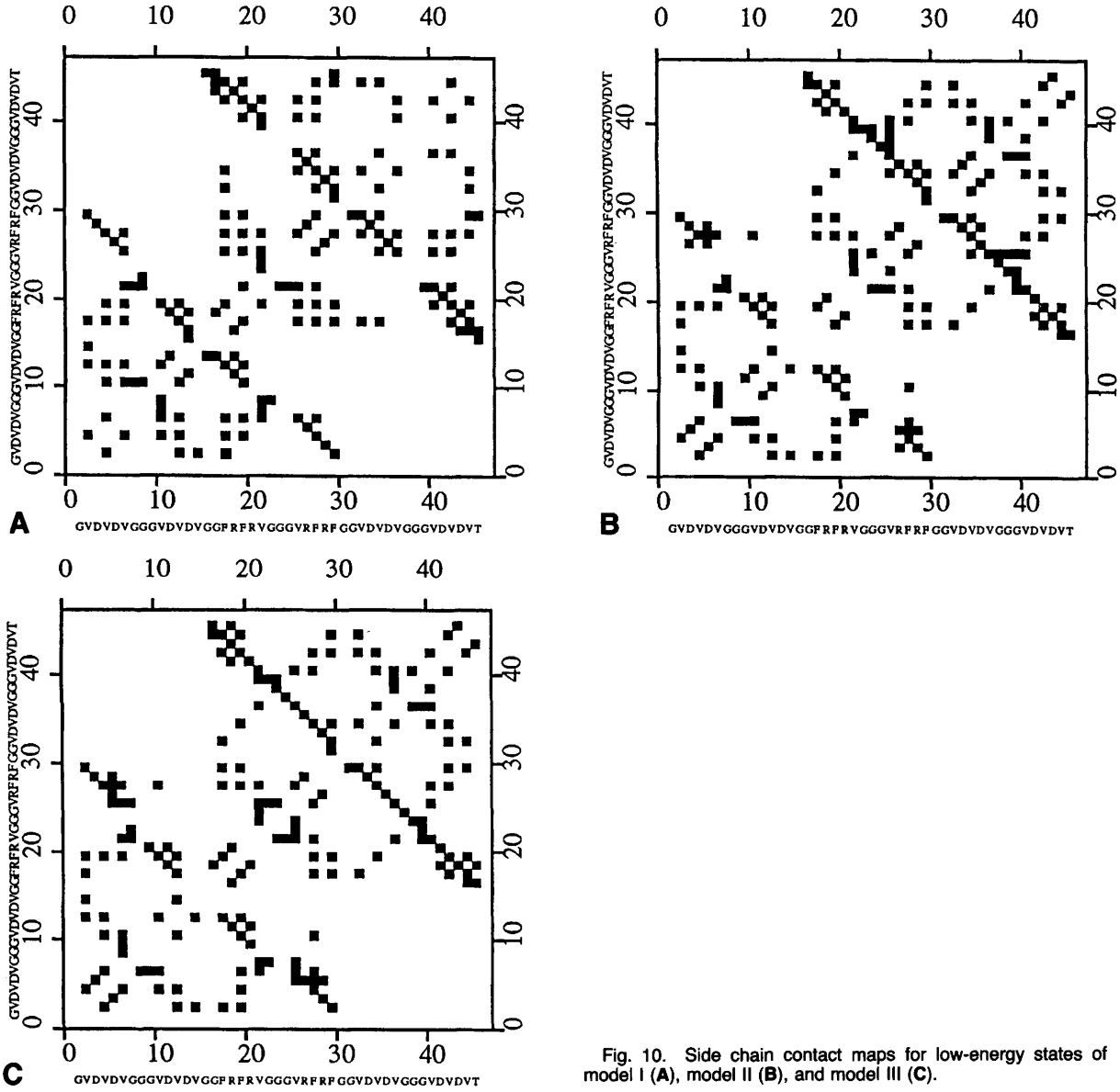

Fig. 10. Side chain contact maps for low-energy states of model I (**A**), model II (**B**), and model III (**C**).

is the essential difference between our model and more simplified models.

The compact states of our model as in real proteins can possess considerable conformational entropy. Upon the transition from the molten globule[15,16] state to the native state, real proteins undergo a complex structural rearrangement that accompanies side chain fixation. This process is commonly associated with a large entropy and energy change for this transition. In fixed single rotamer side chain representations, the molten globule to native transition involving side chain fixation is not possible. Rather, all such simplified models miss an important physical feature. However, in the framework of our reduced model, the explicit cooperative potential of side group interactions facilitates the fixation of side groups to a collection of rotameric states that, when accompanied by a small rearrangement of the main chain, leads to a proteinlike pattern of interactions.

Is the use of a multibody potential to permit side-chain fixation simply the result of the fuzzy (single interaction center with the square well contact potential) side chain representation used in the model? To some extent, this must be true. Then the explicit cooperative terms provide regularizing corrections to protein packing that are poorly defined by the pairwise interactions. On the other hand, in molecular dynamics simulations of full atom models of proteins, it has also been observed that the starting crystallographic structures tend to diffuse their initial regular side chain packing toward a more liquidlike arrangement.[34] These studies demonstrated that even detailed models of proteins allow alterna-

tive, nonproteinlike, but comparably dense packing of *the side chains*. Apparently, since full atom models have similar difficulties, the problem of producing native like packing arrangements of the side chains is not just an artifact of a reduced protein model. Perhaps, in order to reproduce the cooperative transition from the molten globule to native state of globular proteins, some kind of multibody interaction potentials have to be incorporated into any force field, regardless of the level of detail of the side group description.[35] Such a viewpoint is very much in the spirit (however, on the level of tertiary interactions) of the cooperativity of helix-coil[36] or β sheet-coil[37] transitions of statistical mechanical models. The present work suggests that, while the specific patterns of packing of the protein side chains[4,15,16,38,39] could be well characterized in the context of static properties, the modeling of the cooperative folding transition, using molecular models, may require explicit higher order multibody interactions.

In summary, the dramatic increase of the cooperativity of the folding process seen in model II and model III seems to be consistent with the presently accepted picture of protein folding. Thus, these simulations suggest the origin of the molecular mechanism of the transition from the denatured state through the molten globule state to the native state. In model I, the native state is rather poorly defined (as far as the thermodynamics is concerned), and its almost continuous transition appears to be closer to a random coil to molten globule transition. The conformations corresponding to the basin of free energy of the folded state have some features of a molten globule and some features of a native state. In model II (and even more so in model III), the folding, in spite of a comparable conformational energy change, is much more cooperative than that observed in model I. This larger free-energy gap between the native and higher energy structures, and consequently a more structurally unique low-energy state, reproduces the currently accepted picture of folding thermodynamics. The cooperativity of the folding process is associated with the passage from the molten globule to the native state; such cooperativity arises from cooperative side chain packing interactions.

ACKNOWLEDGMENTS

This work was partially supported by NIH grant GM-37408. A. Kolinski acknowledges partial support from University of Warsaw grant BST-502/34/95. A. Kolinski is an International Research Scholar of the Howard Hughes Medical Institute (HHMI grant 75195-543402). Stimulating discussions with Professor Harold Scheraga are gratefully acknowledged. Assistance of S. Wynant in the preparation of this manuscript is greatly appreciated.

REFERENCES

1. Anfinsen, C.B., Scheraga, H.A. Experimental and theoretical aspects of protein folding. Adv. Protein Chem. 29:205–300, 1975.
2. Ptitsyn, O.B. Protein folding: Hypotheses and experiments. J. Protein Chem. 6:273–293, 1987.
3. Grosberg, A.Y., Khokhlov, A.R. Physics of phase transition in solutions of macromolecules. Sov. Sci. Rev. A 8:147–252, 1987.
4. Shakhnovich, E.I., Finkelstein, A.V. Theory of cooperative transitions in protein molecules. I. Why denaturation of a globular protein is a first-order phase transition. Biopolymers 28:1667–1680, 1989.
5. Karplus, M., Sali, A. Theoretical studies of protein folding and unfolding. Curr. Opin. Struct. Biol. 5:58–73, 1995.
6. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D., Chan, H.S. Principles of protein folding: A perspective from simple exact models. Protein Sci. 4:561–602, 1995.
7. Frauenfelder, H., Wolynes, P.G. Biomolecules: Where the physics of complexity and simplicity meet. Phys. Today 47:58–64, 1994.
8. Bryngelson, J.D., Onuchic, J.N., Socci, N.D., Wolynes, P.G. Funnels, pathways and the energy landscape of protein folding: A synthesis. Proteins 21:167–195, 1995.
9. Shakhnovich, E.I., Gutin, A.M. Engineering of stable and fast-folding sequences of model proteins. Proc. Natl. Acad. Sci. USA 90:7195–7199, 1993.
10. Sali, A., Shakhnovich, E.I., Karplus, M. Kinetics of protein folding. A lattice model study of requirements for folding of the native state. J. Mol. Biol. 235:1614–1636, 1994.
11. Kolinski, A., Godzik, A., Skolnick, J.A. General method for the prediction of the three dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. J. Chem. Phys. 98:7420–7433, 1993.
12. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 18:338–352, 1994.
13. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. Proteins 18:353–366, 1994.
14. Vieth, M., Kolinski, A., Brooks III, C., Skolnick, J. Prediction of the folding pathways and structure of the GCN4 leucine zipper. J. Mol. Biol. 237:361–367, 1994.
15. Kuwajima, K. The molten globule state as a clue for understanding the folding and cooperativity of globular protein structure. Proteins 6:87–103, 1989.
16. Ptitsyn, O.B., Pain, R.H., Semisotnov, G.V., Zerovnik, E., Razgulyaev, O.I. Evidence for a molten globule state as a general intermediate in protein folding. FEBS Lett. 262: 20–24, 1990.
17. Godzik, A., Kolinski, A., Skolnick, J. Lattice representation of globular proteins: How good are they? J. Comp. Chem. 14:1194–1202, 1993.
18. Skolnick, J., Kolinski, A., Ortiz, A.R. MONSSTER: A method for folding globular proteins with a small number of distance restraints. J. Mol. Biol., in press.
19. Hao, M.-H., Scheraga, H.A. Monte Carlo simulations of a first-order transition for protein folding. J. Phys. Chem. 98:4940–4948, 1994.
20. Hao, M.-H., Scheraga, H.A. Statistical thermodynamics of protein folding: Sequence dependence. J. Phys. Chem. 98: 9882–9893, 1994.
21. Berg, B.A., Neuhaus, T. Multicanonical ensemble: A new approach to simulate first-order phase transitions. Phys. Rev. Lett. 68:9–12, 1991.
22. Hansmann, U.H.E., Okamoto, Y. Prediction of peptide conformation by a multicanonical algorithm: New approach to the multiple minima problem. J. Comput. Chem. 14:1333–1338, 1993.
23. Kolinski, A., Galazka, W., Skolnick, J. Computer design of idealized β-motifs. J. Chem. Phys. 103: 10286–10296, 1995.
24. Kolinski, A., Milik, M., Rycombel, J., Skolnick, J. A reduced model of short-range interactions in polypeptide chains. J. Chem. Phys. 103:4312–4323, 1995.
25. Milik, M., Kolinski, A., Skolnick, J. An algorithm for rapid reconstruction of a protein backbone from alpha carbon coordinates. J. Comput. Chem. In press, 1996.

26. Kolinski, A., Skolnick, J. Parameters of statistical potentials. Available by ftp from public directory, scripps.edu (pub/andr/MCSP)
27. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
28. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Simanouchi, T., Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
29. PDB, Quarterly Newsletter, No. 71, January 1995.
30. Handel, T., DeGrado, W.F. A designed 4-helical bundle shows characteristics of both molten globule and native state. Biophys. J. 61:A265, 1992.
31. Raleigh, D.P., DeGrado, W.F. A de novo designed protein shows a thermally induced transition from a native to a molten globule like state. J. Am. Chem. Soc. 114:10079–10081, 1992.
32. Betz, S.F., Bryson, J.W., DeGrado, W.F. Nativelike and structurally characterized designed α-helical bundles. Curr. Biol. 5:457–463, 1995.
33. Lee, J. New Monte Carlo algorithm: Entropic sampling. Phys. Rev. Lett. 71:211–214, 1993.
34. Elofsson, A., Nilsson, L. How consistent are molecular dynamics simulations? Comparing structure and dynamics in reduced and oxidized Escherichia coli thioredoxin. J. Mol. Biol. 223:766–780, 1993.
35. DeBolt, S., Skolnick, J. Protein Eng. 9:637–655, 1996.
36. Zimm, B.H., Bragg, J.K. Theory of the phase transition between helix and random coil in polypeptide chains. J. Chem. Phys. 31:526–535, 1959.
37. Mattice, W.L. Annu. Rev. Biophys. Biophys. Chem. 18:93, 1989.
38. Milik, M., Kolinski, A., Skolnick, J. Neural network system for the evaluation of side chain packing in protein structures. Protein Eng. 8:225–236, 1995.
39. Ponder, J.W., Richards, F.M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.