PROTEINS: Structure, Function, and Genetics 16:8–28 (1993)

RESEARCH ARTICLES

Computer Modeling and Folding of Four-Helix Bundles

Antonio Rey and Jeffrey Skolnick
Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037-1093

ABSTRACT **In the context of simplified models of globular proteins, the requirements for the unique folding to a four-helix bundle have been addressed through a new Monte Carlo procedure. In particular, the relative importance of secondary versus tertiary interactions in determining the nature of the folded structure is examined. Various cases spanning the extremes where tertiary interactions completely dominate to that where tertiary interactions are negligible have been explored. Not surprisingly, the folding to unique four-helix bundles is found to depend on an adequate balance of the secondary and tertiary interactions. Moreover, because the simplified model is composed of spheres representing α-carbons and side chains, the geometry of the latter being based on small real amino acids, the role played by the side chains, and the problems associated with packing and hard-core repulsions, are considered. Also, possible folding intermediates and their relationship with the experimentally observed molten globule state are explored. From these studies, a general set of rules is extracted which should aid in the further design of more detailed protein models adequate to more fully investigate the protein folding problem. Finally, the relationship between our conclusions and experimental work with specifically designed sequences is briefly discussed.**

Key words: protein folding, four-helix bundles, sequence design, side-chain packing, folding pathways, computer simulation

INTRODUCTION

In spite of increasing attention over the last several years, the general features of globular protein folding and stability remain controversial. Among the salient questions is the relative importance of local versus tertiary interactions in stabilizing the folded structure. Should one view protein folding as simply a collapse where local preferences for helix, extended states and turns are mainly involved in fine turning of the structure,[1] or perhaps local preferences, while small, play a crucial role in the early initiation events and in the partial elimination of liquid-like states on collapse?[2] Furthermore, what features should be considered in the design of a globular protein? How crucial is the side chain packing in the core versus the choice of amino acid sequences that favor a particular element of secondary structure? Turning to the process of protein folding itself, how do proteins avoid the Levinthal paradox? What is the nature of the early and late intermediates in folding? What exactly is the molten globule state?[3] Do proteins fold via a large manifold of pathways or are there a relatively small number?[4] Clearly, these and other questions have been subjected to increasing experimental and theoretical scrutiny. Our objective here is to examine these questions in the context of simplified theoretical models of globular proteins. These afford the advantage that one can perform a series of numerical experiments to examine the consequences of a given series of assumptions about the physics of protein folding. The aim is to identify a set of conditions under which model proteins uniquely fold to the native state, chosen here, because of its structural simplicity, to be a four-helix bundle.

The use of computer simulations to investigate protein folding has recently seen considerable development and has resulted in a wide diversity of approaches.[5] Perhaps, the most powerful use of simulations of model systems is that they permit one to ask questions of the "what if" variety. For example, while experimentally it has not proven possible to fold proteins devoid of side chains, it is possible to do so on a computer, therefore explicitly checking the role of side chain packing in the folding process. Similarly, at least for a representative range of parameters, the extremes of the relative importance of

Received July 21, 1992; revision accepted September 18, 1992.

Address reprint requests to Dr. Jeffrey Skolnick, Department of Molecular Biology, The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, CA 92037-1093.
Permanent address of A. Rey: Departamento de Química Física, Facultad de Ciencias Químicas, Universidad Complutense de Madrid, E-28040 Madrid, Spain.

secondary versus tertiary interactions in protein folding can be explored. One can turn off all the secondary preferences and see what happens, and then go to the other limit and have tertiary interactions weak and secondary preferences strong. Furthermore, computer simulations allow one to design numerical experiments which focus on properties that are not accessible to traditional experimental techniques. At present, it is very difficult to get experimental evidence to support or dismiss the different generic mechanisms proposed for the folding pathways of globular proteins. On the other hand, it is easier (at least conceptually) to design a computer simulation model adaptable to this kind of study. It is clear that the formulation of the model may influence the results, therefore severely limiting their significance. Our aim is, then, to design a simple, computationally tractable model, which can provide real insights into the mechanisms of protein folding.

For globular proteins having a medium to a relatively large number of residues, probably dynamic Monte Carlo (MC) algorithms constitute the most promising approach to the protein folding problem.[5] They allow for different levels of accuracy in the representation of the protein, are able to avoid kinetic trapping in local energy minima, and can represent, at least qualitatively, a real (though usually unknown) and long time scale, totally unreachable by other simulation methods such as Brownian dynamics (BD) or molecular dynamics.

In this paper, we describe an off-lattice dynamic MC method developed for this purpose, and have studied the folding pathways of idealized model four-helix bundles.[6] The model includes both α-carbons and side chains, and therefore the role of side chain packing and its general influence on the folding pathways can be specifically addressed. In addition, an off-lattice realization of the MC procedure can shed some light on the influence that the discretization of the configurational space (i.e., the use of an underlying lattice) employed in previous simulations[7] can have on the observed pathways. Previous studies have shown that, at least for very simplified models without side chains, the general features of the folding pathways do not depend upon the use of the lattice, nor on the simulation technique.[8] However, the use of low coordination lattices in which the elements of regular secondary structure have a very limited set of allowed orientations can pose some problems. We shall see, however, that our off-lattice results are quite consistent with folding pathways observed in high coordination lattices,[9] where these problems are greatly reduced or entirely eliminated.

The organization of the remainder of the paper is as follows. In the next section, we describe the model. This description includes some technical details of the geometry and dynamics of the model, which can be readily skipped by the reader mainly interested in the qualitative conclusions. However, we call attention to the subsection on Design Requirements, dedicated to the design requirements of a four-helix bundle from the perspective of our model. Indeed, this is precisely the aim of intense experimental and computer research currently under way[10,11] in the de novo design of globular proteins. Also, in the subsection on Potential Functions, we present the potential function employed in the model, as a crude—but not unphysical—representation of the energy surface one might expect for a simple globular protein.

The third section analyzes the trajectories which constitute the outcome of the MC simulations. First, we discuss the possibility of getting a unique folded conformation for a given sequence (and model), as a function of the different assumptions about the relative importance of various terms in the potential function. Then, we discuss the folding pathways exhibited by our model proteins. We focus in particular on the nature of the intermediate conformations observed during the folding process, and the problems related to the adequate packing of the interacting elements comprising the structure. These points find a direct parallelism to some proposed states derived from experimental evidence, e.g., molten globules,[3] which could be relevant to understanding protein folding.

Finally, the last section presents the final comparison of the models investigated, a summary of the conclusions extracted from our study, and their implications for further studies about globular protein structure and folding.

DESCRIPTION OF THE MODEL

Geometry

The protein model considered in this work includes a spherical representation of both α-carbons and side chains (one sphere for the α-carbon and one single sphere for every side chain). The spheres representing the α-carbons have a diameter of 3 Å, and behave as impenetrable spheres for any other α-carbon. The radii of the spheres representing the side chain are determined by their repulsive cores and are described below. The distance between contiguous α-carbons in the primary sequence of the protein is constant, equal to 3.8 Å, and constitutes the unit length for all the distances employed in the model.

The virtual "bond angles" between α-carbons are allowed to fluctuate quite freely, but avoid the extremes which are not found in real protein backbones. Thus, these angles can have any value in the range from 78 to 143°.

Side chains

The position of the spheres representing the side chains with respect to the α-carbon backbone is based on a statistical analysis of well resolved crystal structures. In this analysis, the centers of mass of

the side chains in a series of real protein structures were obtained, and their distances and orientations with respect to the backbone were classified in a histogram, as a function of the backbone conformation. In this initial study employing the off-lattice algorithm, we have only used models with small, rather symmetrical side chains (valine and serine), which are present as a single rotamer in the library. Therefore, the side chain is rigidly attached to the backbone, its orientation being defined for residue i by the coordinates of the three α-carbons C^{α}_{i-1}, C^{α}_{i}, and C^{α}_{i+1}. Since one of these atoms is always missing in the terminal residues, these are always defined with the geometry of glycines (i.e., without sidechains).

Double primary sequence

The position of the side chains with respect to the backbone, for a given local conformation of the latter, is dictated by the primary sequence of amino acids. In addition, the model includes a second primary sequence, in which the residues are classified, according to their interaction nature, as hydrophilic or hydrophobic. Therefore, this second sequence defines the amphipathic pattern of the protein, whose design is of importance in order to get the desired fold.

Design Requirements

An advantage of the double primary sequence in the present model is the possibility of defining sequences of residues whose side chains' geometric properties are uncoupled from the interaction pattern. For example, we can define a sequence composed of glycines (only the α-carbon representation) and still retain an amphipathic sequence of hydrophobic and hydrophilic residues. This permits the role of side chain packing in folding to be explored by examining the differences between a model with and without side chains. Our objectives are to find a pattern of hydrophobic and hydrophilic residues that (1) gives a unique folded structure, (2) is highly regular, and (3) is still representative of actual folding topologies appearing in real proteins. Both from previous lattice MC simulations[7,9] and experiments in protein design,[10,12] the four-helix bundle seems to be a reasonabled candidate.

General considerations

Taking into account the requirement for structural uniqueness, the turns connecting the four helical fragments should be as short as possible; this will tend to avoid alternative stable topologies. Also, one has to arrange the distribution of hydrophobic and hydrophilic residues to maximize the number of favorable (attractive) interactions in the desired structure. This is actually a nontrivial problem, due to the deviations from the perfectly regular topology seen in natural four helix bundles. Therefore, we employ simple structures based on the diamond lattice models studied previously.[7] The main disadvantage of this choice is that it only allows for the possibility of square helices, i.e., helices with four residues per helical turn, instead of the 3.6 in real α-helices.

Model without side chains

In a model without side chains, the simplest way of forming a compact four-helix bundle on a diamond lattice is sketched in model **A** of Figure 1. This model has the advantage that all residues are included in helical fragments, and each turn between them is just formed by one virtual bond joining α-carbons, almost precluding any large deviation from the desired structure. Longer turns result in open structures requiring tertiary interactions that are too long ranged, yielding compact globules devoid of secondary structure. This structure has a 3+1 pattern of hydrophobic and hydrophilic residues (three hydrophobic and one hydrophilic residues per helical turn). However, this structure does not leave any room for side-chain packing, since the distance between α-carbons of different helices in the hydrophobic core is the same as the distance between neighbor α-carbons in the primary sequence.

Model with side chains

In order to introduce side chains, the structure must be opened up. Two possibilities, also built on a diamond lattice, are depicted in Figure 1, models **B** and **C**. Both models have one residue and two virtual bonds per turn, which gives them a larger flexibility than model **A**. Therefore, the choice of the amphipathic pattern now has to be done very carefully. Because the distance between α-carbons is quite large, it is better to account for the interactions between the side chain positions directly rather than between the α-carbons. Otherwise their interactions would again have to be very long ranged, yielding very compact globules, with little secondary structure.

The schematic representation of the models is shown in Figure 2. Glycines have been positioned in the turns. Hydrophobic residues are represented by side chains having the geometry of valine with respect to the backbone, while hydrophilic residues have the serine geometry. Again, we note that the choice of these residues is based exclusively on geometric criteria (small size and single orientation with respect to the backbone), and consequently it is not a contradiction, in the context of this model, to have chosen valine, a β-forming residue, as a constituent of the helical fragments.

The most important points to be noticed in Figure 2 are the orientations of the side chains and their interdigitation in the protein interior. Since both α-carbons and side chains have a spherical represen-

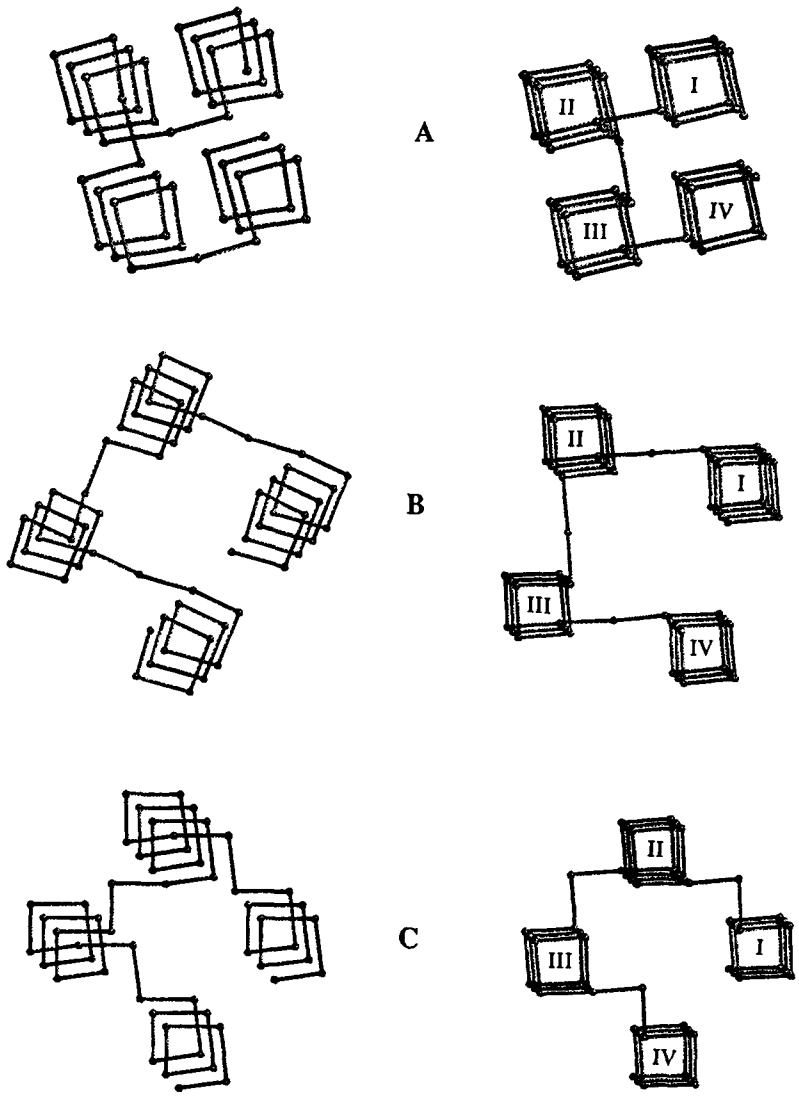

Fig. 1. α-carbon traces of different model four-helix bundles, in top and side views. The top view allows for a good guess of the amphipathic pattern (see text), while in the side view, the disposition of the helices and turns is better appreciated.

tation, space is much more densely filled than might appear from these simple line drawings.

Amphipathic patterns for the side chains

In model **B**, one side-chain in each helical turn is clearly oriented toward the protein interior, and therefore is defined as hydrophobic. Also, the side chain diametrically opposed to this one in every helix is pointing toward the exterior, and is defined as hydrophilic. For the other two side chains per helical turn, both possibilities exist, but neither is adequate. If they are hydrophobic, giving a 3 + 1 pattern of residues, then the resulting structure has very broad hydrophobic faces. This produces a manifold of globular stable compact structures, which in the best case are distorted bundles. The most frequent of these is the Z-topology of the bundle, instead of the desired U-topology (see scheme Fig. 3a). On the other hand, when we opted for the 1 + 3 pattern (only one hydrophobic side chain per helical turn), the hydrophobic faces are too narrow, and the system has a very difficult time finding the correct structure. With tertiary interactions soft enough to avoid quenching the system, the intermediate structures that appear along the folding pathway (discussed below) are not stable enough for the protein to fold. If, however, the magnitude of the interactions is increased (or the temperature decreased), the structure becomes prematurely frozen in alternative solutions. These are mainly triangular three-helix bundles, which lock most of the hydrophobic faces for three helices (see Fig. 3b), and expose only

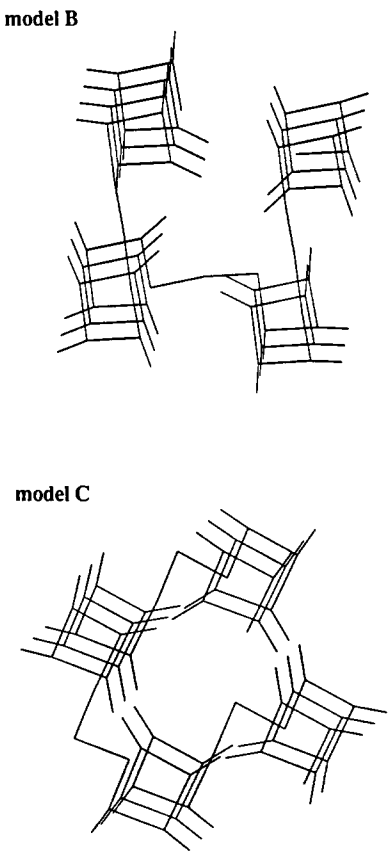

Fig. 2. Full representation of the models employed, including α-carbon and side-chain positions. Model **B:** pattern 1 + 3. Model **C:** pattern 2 + 2 (see text for explanation).

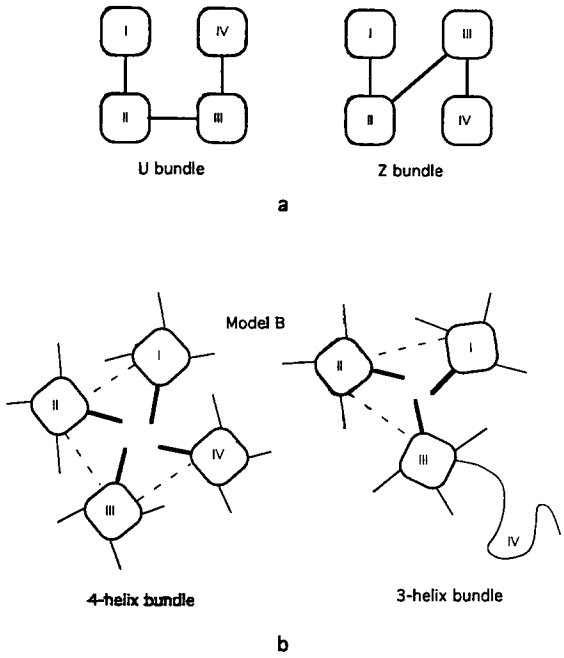

Fig. 3. (**a**) Schematic representation of two alternative possibilities for a four-helix bundle. (**b**) For the 1 + 3 pattern in model **B** (with the hydrophobic residues represented with thick side chains), three helices can fully stabilize a bundle, making it impossible for the fourth strand to assemble onto it.

hydrophilic residues to the exterior. Thus, it becomes virtually impossible for the fourth strand to assemble the last helix in the desired position, at least in a reasonable amount of computer time.

The most successful model corresponds to model C in Figures 1 and 2. This arrangement of the helices imposes a 2 + 2 amphipathic pattern, with two contiguous hydrophobic residues and two contiguous hydrophilic residues per helical turn. The possibility of alternative structures such as Z-bundles or locked three-helix bundles still exists, but these can be eliminated or overcome by adequate definition and tuning of the potential parameters.

Amphipathic patterns and tertiary interactions

The relationship between the structure of the models and their tertiary interactions is depicted in Figure 4. Hydrophobic residues have been represented in bold type, while dashed arrows represent the expected native contacts (between α-carbons in model A and between side chains in model C). These contacts correspond to the distances encountered in the rigid lattice representation of idealized structures, as depicted in Figures 1 or 2. However, in our off-lattice simulations, the folded structures are much more mobile than Figure 4 suggests, and therefore some of the contacts indicated by the arrows can be absent from a given instantaneous conformation in the folded state, while other contacts arise. As an example of the mobility within the folded state, Figure 5 shows representative folded conformations resulting from our calculations, and their corresponding contact maps. In the contact maps, the left upper triangle corresponds to the rigid ideal lattice representation. In the lower right triangle, we show the map for one of the conformations in the folded structures resulting from the folding trajectories. The perfect square arrangement of helices I, II, III, and IV is lost, resulting in contacts between helices I and III, on some occasions, and contacts between helices II and IV in others. The backbone representation of model A clearly reflects this shifting. Also observe the angle between neighboring helices in model C, which does not appear in model A; this also appears in real proteins, and is a direct consequence of the packing of side-chains.

Relation to experimental sequences

Interestingly, the distribution of hydrophobic and hydrophilic residues defined in model C is rather similar to that used by DeGrado and co-workers in their incremental design of proteins with the four-helix bundle topology.[10] The real helices, composed mainly of leucine, glutamic acid, and lysine, show

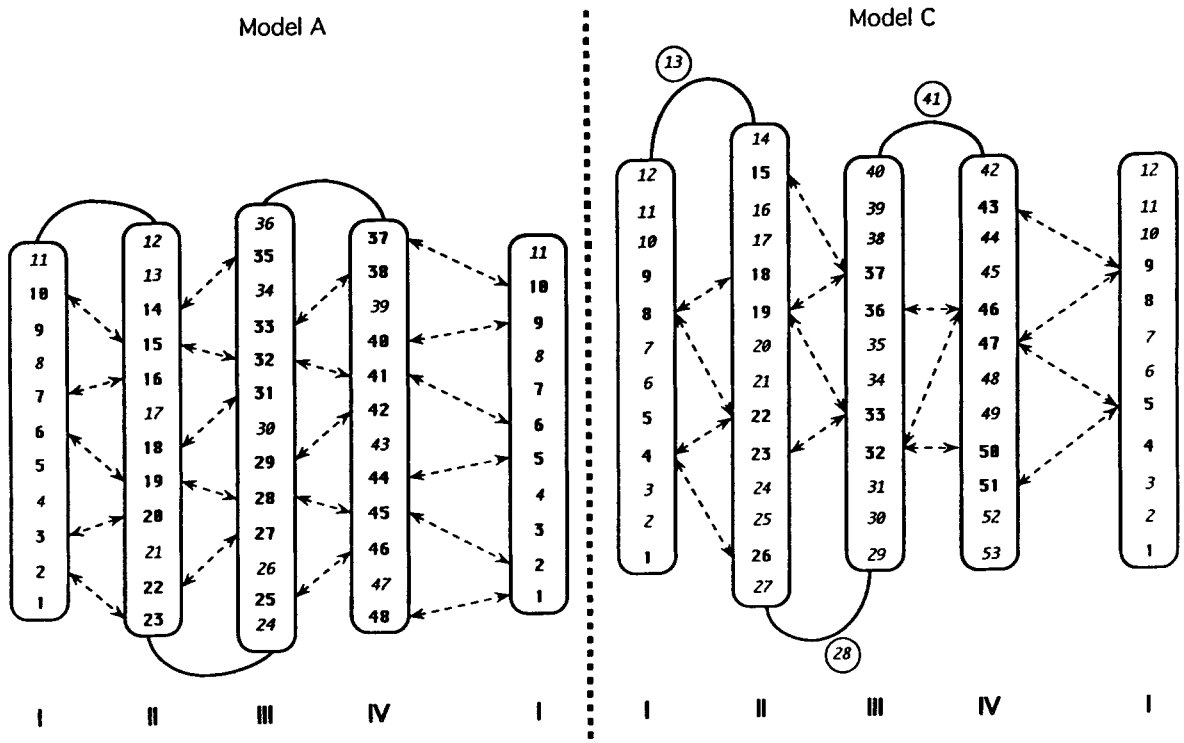

Fig. 4. Schematic representation of the four-helix bundle for the models under consideration. Residues defined as hydrophobic have a boldface number, while residues considered as hydrophilic are in italics. Helix I has been repeated at the right side of the figures, without any bonded connection, to show the tertiary interactions existing between this helix and helix IV when the bundle is assembled.

clearly two faces, distributed with approximately 50% hydrophobic residues, which form the buried core of the folded structure, and another 50% charged residues, exposed to the solvent. It is clear that our model is far simpler than these real, yet highly idealized, model proteins. The interactions between charged residues are only schematically considered in our scale for the tertiary interactions, and the question of the helix dipole is completely ignored. Nevertheless, some features of the model, such as the choice of a compatible sequence, the packing problems, the definition of turns between helices, and others, retain a close resemblance to those encountered in experiment.

Potential Function

The potential function employed in the description of the model energetics is composed of several terms

$$E_{\text{Total}} = E_{\text{b.a.}} + E_{\text{tors}} + E_{\text{helix}} + E_{\text{long}} \qquad (1)$$

whose definition is as follows.

Bond angles

$E_{\text{b.a.}}$ is the contribution to the potential of the virtual bond angle θ defined by three contiguous α-carbons, and is given by a harmonic potential for every backbone bond angle as

$$E_{\text{b.a.}} = K_{\text{b.a.}} (\cos \theta - \cos \theta_0).^2 \qquad (2)$$

In this equation, $K_{\text{b.a.}}$ is the force constant, with a rather small value in our simulations (about $10k_BT$ units, k_B being the Boltzmann constant and T the absolute temperature; we shall use their product as the reduced unit for the energy terms). θ_0 is the equilibrium angle, taken arbitrarily as the angle corresponding to a tetrahedral lattice (109.5°). Actually, this value is close to the average virtual bond angles found in α-helices in real proteins. This contribution to the potential, however, has an almost negligible effect on the dynamics of the folding process.

Torsional angles

E_{tors} is the contribution to the potential arising from the values of the torsional angles defined by the α-carbon trace. There is no straightforward way of formulating a continuous function which reproduces the different torsional states appearing in real proteins. Therefore, we have greatly simplified this contribution to the potential assuming that the torsional states are close to those defined in the rota-

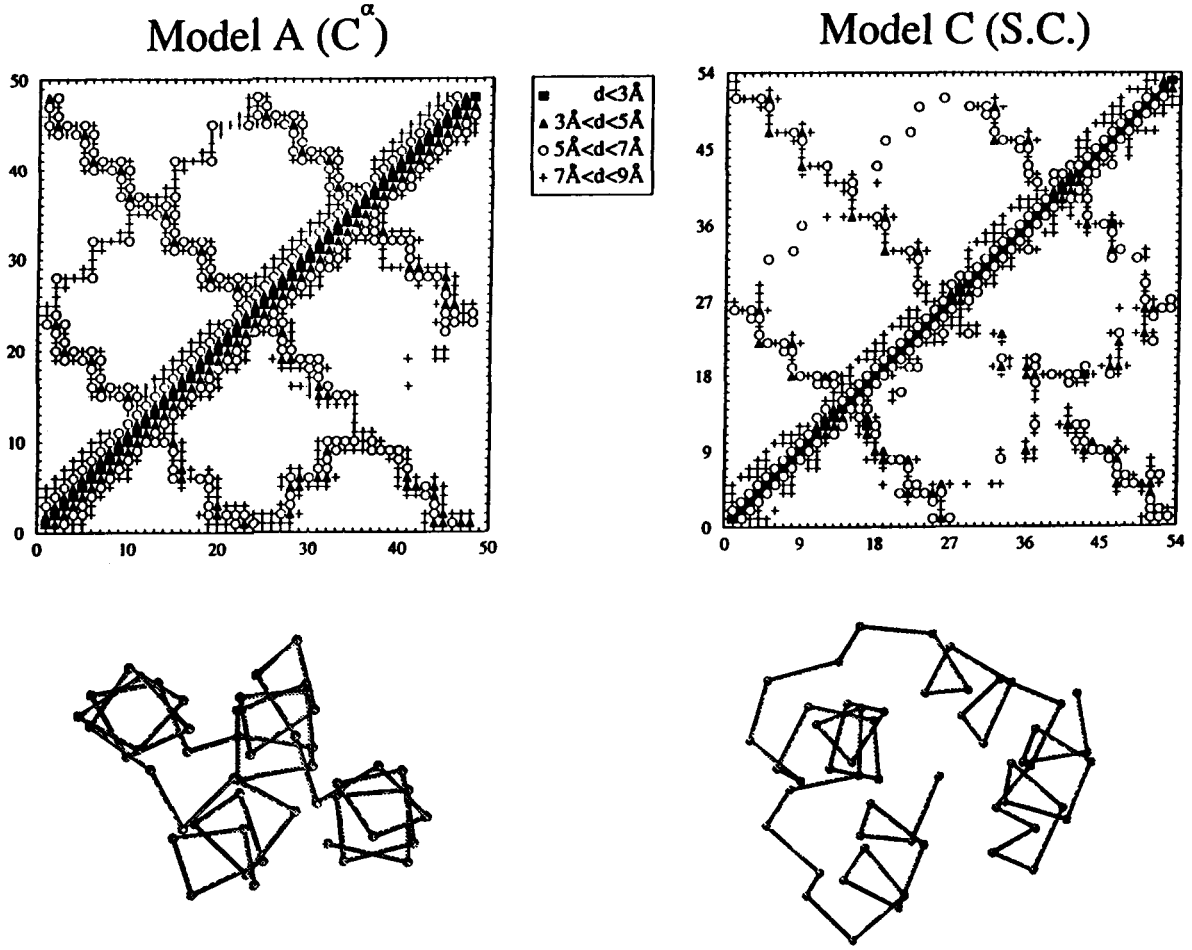

Fig. 5. Contact maps for the models. **Upper left triangle:** contacts in the ideal rigid lattice representation. **Lower right triangle:** contacts in a single conformation taken from the folded state in an off-lattice MC trajectory. Instantaneous backbone conformations resulting from the off-lattice MC trajectories are shown under the corresponding contact maps.

tional isomeric state model[13] for chain molecules, i.e., we only consider the three states *trans*, *gauche*$^+$, and *gauche*$^-$. An α-helix is defined as a series of consecutive *gauche*$^-$ states, while a β-strand would be represented by a series of *trans* states, and a turn by combination of the three possibilities. These torsional states represent the center of a broad region for the torsional values, and not a single value. In fact, the torsional angles can have any value in the range from $-180°$ to $+180°$. As previously,[8] this potential is defined so that it allows one to energetically favor or disfavor any of the three rotational state regions. This is accomplished by fitting a six order polynomial to a three-minima function, the details of which are found elsewhere.[8]

In the helical regions, this torsional term is irrelevant, since it equally weights the three torsional states, with very small barriers between them. The torsional potential is consistent with the turns joining the putative helices in the final folded structure, and therefore there is a small bias built in the model towards the desired folded conformation. It does not prohibit non turn-like conformations in the putative turn-forming regions of the sequence. Rather, it just makes the population of some of the torsional states larger than the others, but the number of transitions between them is still large. Indeed, nonnative torsional states even occur in the folded conformation.

Helical wheel

E_{helix} mimics the hydrogen bond interactions that internally stabilize helical conformations. It has the functional form of a Lennard–Jones 8-6 potential acting between α-carbons i and $i+4$, separated by a distance r, when they are both in the same putative helical fragment existing in the final structure

$$E_{helix} = C\epsilon_h [(\sigma_h/r)^8 - (\sigma_h/r)^6]. \quad (3)$$

Here, ϵ_h is the depth of the potential well. Its value can control the amount of helical secondary structure in the model. σ_h, on the other hand, is the distance at which the potential changes sign. For our simplified models, whose fold is based on diamond lattice structures with square helices, σ_h equals 7.6 Å (2, in the reduced units we are using for the distances). This value creates slightly more expanded helices than those present in real proteins. Finally, C is a constant equal to 9.48 in this equation as well as in Eqs. (4) and (5) presented below. In order to avoid the formation of bulges in the middle of the helices, the potential is only effective when the two virtual torsional angles bracketed by residues i and $i+4$ are in the helical torsional region (the minimum corresponding to the *gauche*$^-$ conformation). This contribution is only operative in the purely helical regions, but it is suppressed in the turn regions.

Tertiary interactions

Finally, E_{long} includes the tertiary interactions between the centers of any pair of side chains corresponding to residues separated by at least three residues along the chain backbone. For glycines, for which the side chain does not exist, the center of the interaction is positioned at the α-carbon. Pairs of hydrophobic (pho) side chains interact favorably through an 8-6 Lennard–Jones potential

$$E_{long} = C\,\epsilon_1\,[(\sigma_1/r)^8 - (\sigma_1/r)^6] \qquad \text{(pho–pho)} \quad (4)$$

while hydrophilic (phi) side-chains repel hydrophobic side chains through the repulsive part (the term corresponding to the exponent 8 in the same potential)

$$E_{long} = C\,\epsilon_i\,(\sigma_1/r)^8 \qquad \text{(pho–phi)}. \quad (5)$$

In both cases, a cut-off distance for the potential is defined, so that the potential is not effective if the distance r between the centers of interaction is larger than 2 reduced units. Pairs of hydrophilic residues are neutral, retaining only the excluded volume condition through a square well potential.

One problem with using the Lennard–Jones attractive–repulsive potentials is that the minimum is very close to the repulsive core. This produces difficulties in side chain packing, especially in the last stages of the folding pathways, which are described in the next section. To avoid the locking of kinetic pathways, the potential is truncated at a minimum distance r_{min}, such that if the distance between interacting centers is less than this value, the potential remains the same. After some testing, we have chosen $r_{min} = 0.90\ \sigma_1$.

In model **A**, the distance between hydrophobic α-carbons in contact in the native four-helix bundle equals the distance between neighbor α-carbons. To have the interaction minimum at this point with an 8-6 Lennard–Jones potential, we must use a value of $\sigma_1 = 0.87$ reduced units. In model **C**, the situation is not so clear, since there are several distances between side chains which can be considered as corresponding to contacts in the native structure. Anyway, we retain the same value, $\sigma_1 = 0.87$, which can be still considered as a reasonable interaction distance between side-chains. The values of ϵ_1 are related to the possibility of folding the model to the designed topology. They will be discussed in detail in following sections.

Model dynamics

Once the potential energy function of the system has been defined, one needs to efficiently sample this energy hypersurface. As previously mentioned, we have chosen a dynamic Monte Carlo algorithm, and in order to keep a certain parallelism between MC moves and a physical time scale, we have used only local moves, which can take place over similar, real time scales.

The most prevalent type of moves are the spike moves (Fig. 6A), in which the α-carbon position of a single residue and the side chains of the adjacent residues are also modified. For the terminii, a move to any point in the spherical surface centered at the neighbor α-carbon and with radius equal to the virtual bond length is attempted (Fig. 6B).

These two types of moves are sufficient to give rather realistic dynamics for expanded conformations. However, they do not allow for small amplitude cooperative motions of several residues, which are necessary to translate or rotate preformed elements of secondary structure. Hence, a model including only spike and end moves could not reproduce folding pathways which exhibit a diffusion–collision assembly mechanism.[14] To solve this problem, we use a new *shifting* move that can affect large portions of the chain, but still retains the local character of the displacement. To perform this *shifting* move (Fig. 6C), an inner residue i is randomly chosen, and a terminal-type move is initially made. Then, the units from $i+1$ to N, with N the total number of residues of the chain, are connected to the new position of i, retaining identical bond vectors as existed prior to the move. This shifts a portion of the chain (the backbone and the corresponding side chains) to a parallel position. This displacement has an amplitude less than the virtual bond length between neighbor α-carbons, and thus, it can be still catalogued as a local move. Nevertheless, since this move affects a number of units considerably larger than the spike or terminal moves (single unit moves), the frequency of attempt is scaled so that only one is attempted every N single unit moves.

This set of moves results in very realistic dynamics, which is similar to BD simulations performed for comparable systems.[8] Thus, one can expect that the folding pathways will be the result only of the physics introduced into the model through the interac-

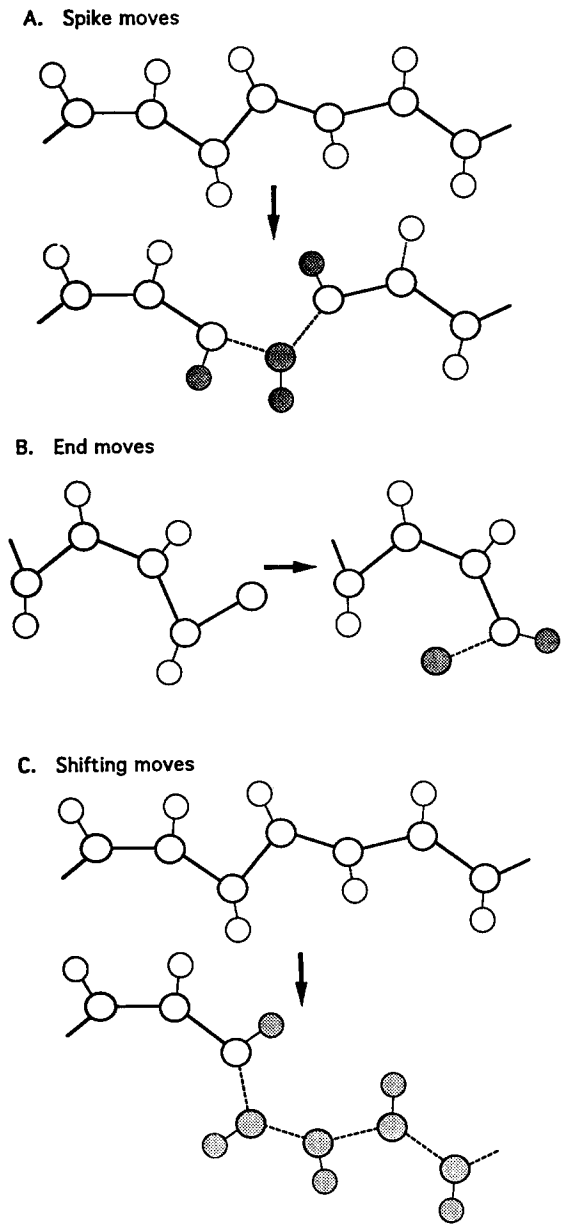

Fig. 6. The move set included in the MC scheme. The dashed virtual bonds and the shaded units are those which have been affected by the move.

tion scheme and will be fully independent of other model considerations.

Any individual move is subject to an acceptance test based on the asymmetric Metropolis criterion. Also, the individual moves are grouped into MC cycles, each one composed by N trials of single unit moves and one trial of the shifting move. The trajectories are usually composed of 2.4×10^6 MC cycles, with the coordinates of the residues recorded for further analysis every 600 MC cycles.

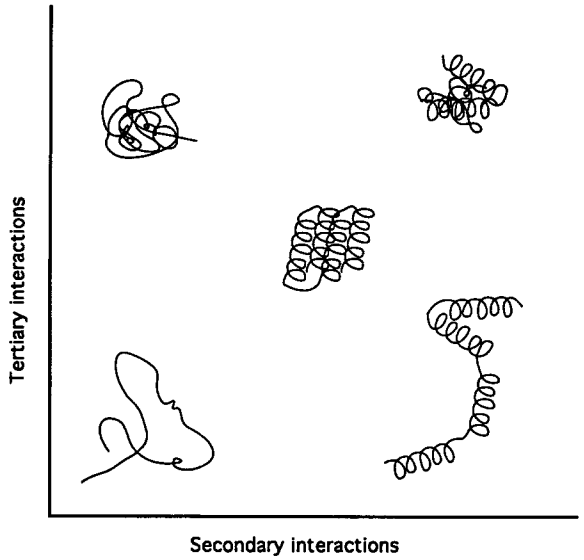

Fig. 7. Qualitative outcome of the MC trajectories, depending on the balance between secondary and tertiary interactions.

RESULTS

Determination of the Folding Conditions: Exploring Parameter Space

Conformational phase diagram

Our aim is to explore a broad set of conditions, and try to determine the behavior of the model in various regimes. Figure 7 presents a schematic summary of the conformations obtained by modifying the relative strengths of the secondary and tertiary preferences over rather extended ranges. This is only a qualitative diagram, which shows the average result of the MC trajectories. Where one is in the conformational phase diagram depends on the balance between the magnitude of the different kinds of interactions, in particular the tertiary interactions and the helical wheel interactions (the latter classified as secondary structure or local interactions). When both interactions are very weak, the chain just behaves as an expanded random coil. If tertiary interactions are too strong, then the chain collapses into a dense globule, as a polymer would in the coil–globule transition, retaining less than 20% of the native secondary structure. As in previous work by Gregoret and Cohen[15] and recent on-lattice calculations including side-chains,[2] the increase in secondary structure resulting simply from compaction is minor. Indeed, if there are no intrinsic helical propensities, a random globule results. On the other hand, if the secondary interactions are the only ones of importance, then the secondary structure freezes, with the associated problems of the packing of rigid helical bodies. It is difficult to interdigitate and adjust side chains if the backbone is frozen and local backbone readjustments accompanying helix sliding are prohibited. Finally, if both secondary and ter-

tiary interactions are simultaneously increased, then the secondary structure develops quickly, but the chain collapses to a globular state with a large helical content, whose exact shape usually depends on the initial conformation at the beginning of the trajectory. In other words, it is kinetically trapped. Only when the local and tertiary interactions are balanced does folding occur to compact states that retain characteristics of proteins.

Effect of the side chains

The limitations in the folding conditions are even more striking for model **C** than for model **A.** In general, we observe that it is much simpler to fold model **A** than model **C,** i.e., the presence of side chains and the restrictions imposed by adequate packing make it more difficult to achieve the final conformation. This seems to be mirrored in real experiments. In our simulations, the reason for this seems to be purely kinetic. While a model with only α-carbons can follow different folding pathways, including in principle the collapse to a relatively compact globular shape which slowly rearranges to reach the folded state, the pathways for the model which includes side chains are much narrower. It is relatively easy for the system to become frozen in compact structures if the parameters controlling the contributions to the potential are not properly balanced.

Balance of the contributions to the potential

To find a set of conditions capable of folding to the same native conformation, independent of the starting unfolded conformation, tuning of the potential parameters has been accomplished. Several points are important in this tuning: the torsional potential enforcing the native conformation of the turn regions, the strength of the potential term acting in helices, which mimics helical hydrogen bonding, and finally, the magnitude and interaction range of the tertiary interactions. Since the zero of the potentials is in a certain sense arbitrary, one has to balance them so that no single contribution becomes frozen at a temperature when the other contributions are "too hot." Otherwise, in extreme cases, this can produce frozen helices or frozen random globules. Also, since the number of hydrophobic contacts in the native structure of model **A** is larger than in model **C,** we should expect the former to fold at a larger temperature than the latter, as in fact happens. If T_C is the temperature at which the folding of model **C** takes place, we have been able to fold model **A,** with similar weights for the potential terms, at a temperature $T_A = 1.4T_C$. Interestingly, and consistent with earlier calculations,[9] this ratio and that between the 21 ideal native contacts for model **A** and the 15 ideal native contacts for model **C** have exactly the same value. Of course, even with fluctuations of about 10% in the temperature, the algorithm will be able to find the folded conformation, although the pathways can be partially modified. Temperatures below the optimum value begin to quench the system. If the temperature is not too low, the ergodicity of the algorithm prevails, and the chain usually attains the folded conformation. If the temperature is too high, the stability of all the intermediate structures is reduced. The folded conformation can still be reached, but its lifetime is shorter than at the optimal temperatures, and most of the conformations spanned by the chain correspond to the unfolded coil. This is merely a reflection of the equilibrium nature of the folding process.

Turn propensities

The native torsional state for the virtual bonds in the native turn regions is biased by means of E_{tors} with respect to the other two torsional states, which are made isoenergetic. In model **A,** just a small energy difference favoring the native torsional state (of the order of $0.75k_BT$ per bond) is enough to fold the correct four-helix bundle. The free energy difference is, of course, much smaller. With this value, we find trajectories in which a Z-bundle transiently appears, but it always dissolves to yield the U-topology (see Fig. 3). In model **C,** due to the larger flexibility of the longer turns, this contribution has to be increased. Since the folded structure for this model **C** is more open than in model **A,** this also makes it easier to find adequate side-chain packing in the Z-bundle topology, with only minor distortions in the secondary structure of individual helices. To avoid this incorrect topology, a difference of energy of $1.5k_BT$ has been found necessary between the native and the two non-native torsional states. Still, the population of native turns in the unfolded state is less than 10%.

Secondary interactions

The helical wheel contribution to the potential, E_{helix}, is controlled by the value of ϵ_h in Eq. (3). We have found the best values to be $\epsilon_h/k_BT = 1.45$ for model **A** and $\epsilon_h/k_BT = 2.0$ for model **C.** These are just the central values in an interval of good sets of ϵ_h. Variations in ϵ_h of about 10% are almost irrelevant for our purposes. Variations up to 15% can still fold the protein, though the pathways begin to be affected. With values out of this range, the folding is practically impossible with our computational resources.

The value of ϵ_h/k_BT employed in our model produces a relatively high helical population for the chains, but this does not mean that we have completely preformed helices which survive unaffected during the whole simulation. Instead, we have a very dynamic picture in which, while maintaining an average helical content of around 50%, the individual helical turns are continuously forming and dissolving. This 50% helix value is rather high and

TABLE I. Summary of Parameters of the Model Potential*

Model	$\Delta\epsilon_{tors}/k_BT$†	σ_h‡	ϵ_h/k_BT	σ_l‡	ϵ_l/k_BT (pho–pho)	ϵ_l/k_BT (pho–phi)
A	0.75 (10)	2.0	1.45 (15)	0.87	1.75 (35)	1.00 (30)
C	1.50 (10)	2.0	2.00 (15)	0.87	2.00 (15)	1.15 (30)

*The energetic parameters ϵ/k_BT correspond to the optimum values, which are centered in a range about which the trajectories yield correctly folded conformations. The average amplitudes of these ranges (expressed as a percentage of the average value) are shown in parentheses. See text for details.
†Energy difference between the native torsional minimum and the other two rotational states, in the putative turn regions ($\Delta\epsilon_{tors}/k_BT=0$ in the putative helical regions).
‡In units of the distance between neighbor α-carbons.

was found to be necessary, at constant temperature, for finding a well characterized folded conformation in our trajectories. Of course, we recognize that this helix content is unrealistically high. Partially, it covers defects in the simplified representation of our model, both at the level of secondary and tertiary interactions, which can be overcome by further improvements in the model; partially, it reflects limitations in computer time. However, even with this large average helical population, a mechanism of assembly based on the diffusion of preformed elements of secondary structure is not observed. Consequently, one can assume that the probability of finding it when a more reduced helical content—closer to physical reality—is obtained, is not large.

Tertiary interactions

Finally, and probably also most importantly, we have to determine the proper set of parameters controlling the tertiary or long range interactions, E_{long}. For ϵ_l/k_BT [Eq. (4)], values around 2.0 for the interactions between hydrophobic groups constitute a good compromise inside a relatively wide range of values. This is true especially for model **A**, in which values even as low as 1.5 for the interaction between hydrophobes can yield the folded structure, though its stability is greatly reduced. In model **C**, one can drop ϵ_l/k_BT by only about 15%. Under these conditions, the model takes a lot of time to fold. The values of ϵ_l/k_BT can also be increased, but especially for model **C**, there is not a wide margin to do so (only about 5–10%), since the collapse of the structure produces kinetically trapped non-native states (see upper right corner of Fig. 7).

The interaction between hydrophobic and hydrophilic residues [Eq. (5)], is much more flexible. In general, values of ϵ_l/k_BT representing between 40 and 70% of the corresponding ϵ_l/k_BT for interactions between hydrophobes yield the correct folded structure. Lower values tend to favor an increase in the population of Z-bundles (and other misfolded compact conformations), while very large repulsive interactions have an effect rather similar to the hard core repulsions, locking kinetic channels that seem to be important along the folding pathways.

Summary of the potential parameters

The optimum values for the parameters controlling the potential contributions are collected in Table I. Only for a set of conditions where secondary and tertiary interactions properly balance, is the desired four-helix bundle the outcome of the trajectory. At first, this might be a cause of concern. However, not every polymer is a protein, and not every amino acid sequence folds to give globular, protein-like, dense states. Indeed, such considerations probably are very important in the experimental design of model proteins.

Folding Pathways

Monitoring the trajectories

Let us begin by considering the folding pathways corresponding to model **A**. To follow the pathway requires some faithful way to represent the events that occur. No single variable is capable of giving a complete description of the dynamics. Therefore, we follow some significant properties, whose combined evaluation allows us to quite accurately reconstruct the ongoing process. In Figure 8, for an isothermal trajectory, we present the variation with time (or number of MC cycles) of (1) the square radius of gyration of the chain, R_g^2, (2) the energy describing the helical interactions, E_{helix} (solid line), and the tertiary interactions, E_{long} (dotted line), and (3) the evolution of the native contacts. Figure 8c tries to represent a three-dimensional picture, in which the distance between the α-carbons corresponding to the residues indicated in the vertical axis has been measured along the trajectory. One point in this plot means that the corresponding α-carbons are at a distance less than 2 reduced units. For every contact, the height of the point in the cloud gives an idea of the distance, which is smaller when the level of the point is lower. This scale goes from about 0.8 to 2.0 reduced units, its width being represented by the black horizontal bar at the top of the plot. More than the values of the distances themselves, we are interested in studying the presence or absence of the contacts and their stability, given by the width of the cloud of points for every contact. A disperse cloud

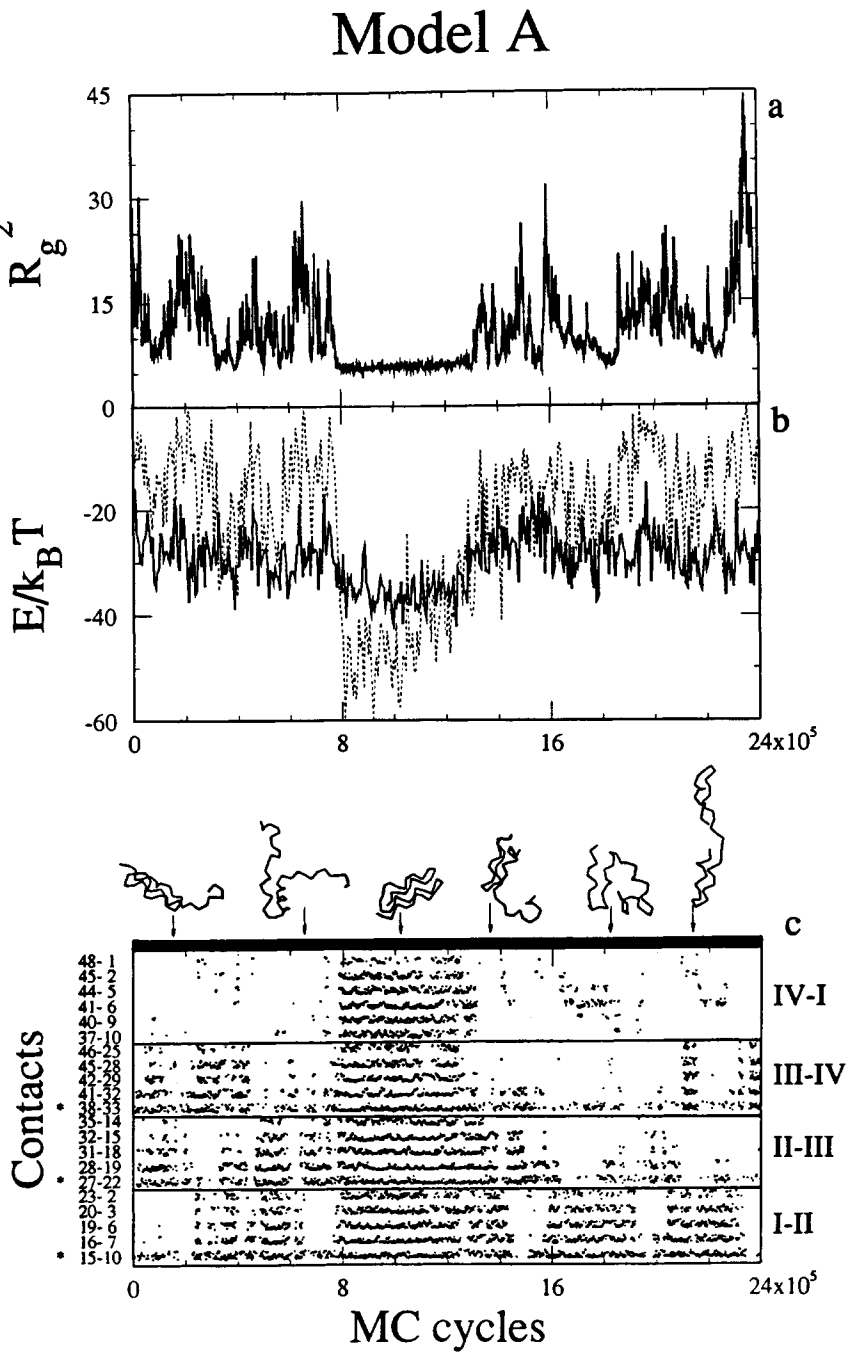

Fig. 8. Folding and unfolding trajectory of model **A.** (**a**) Radius of gyration, R_g, in reduced units. (**b**) Energetic contributions for the helical term, E_{helix} (solid line), and of the tertiary interactions, E_{long} (dotted line) along the trajectory, also in reduced units. (**c**) Evolution of the native contacts between α-carbons (see text for the interpretation of the figure). Asterisks in the vertical axis indicate the contacts closest to the turns between helices. Some snapshots of instantaneous conformations along the trajectory are included. They correspond, successively, to a 3-helix bundle, a central hairpin, the folded 4-helix bundle, a distorted 3-helix bundle, a terminal hairpin, and a double hairpin. The arrows indicate the approximate occurrence of these conformations along the trajectory.

will indicate a very mobile structure, while a narrow cloud is indicative of high stability structures. On top of this last part of the figure, some snapshots of the chain, corresponding to instantaneous conformations along the trajectory, are sketched, in order to further clarify the relationship between the conformation and the different properties being monitored.

Model without side chains

We now proceed to the discussion of Figure 8, which corresponds to a typical folding and unfolding trajectory of the four-helix bundle represented by model **A**. The trajectory begins from a random conformation, not very extended in this particular case, and rapidly develops a substantial amount of secondary structure. The increase in the secondary structure translates to a rapid decrease (since it is a negative magnitude) of E_{helix} (solid line), as reflected in Figure 8b. Actually, the process is so fast that the figure hardly reflects the very steep decrease, but the initial configuration has a value of $E_{\text{helix}}/k_{\text{B}}T$ approximately equal to only -3.

This process is closely followed by some collapse. This collapse does not translate into a random search, but it represents the formation of several distinctive elements of supersecondary structure. The formation of these elements is one of the characteristic features observed in the trajectories (we have computed 30 of them for model **A**), and therefore they could be considered to be *intermediates* in the folding process. This term, however, has to be used with a certain amount of caution, since concepts like this, molten globule[3] and others currently being used when describing the folding pathways, are often employed with different, sometimes non equivalent, meanings. Moreover, since the models lack side chains or have a single rotamer representation of the sidechains, the interplay of side chain and backbone fixation characteristic of the molten globule to native state transition cannot be addressed. Thus, we are studying the kinematics of topology assembly in this series of simulations.

Folding intermediates for model **A**

Our folding intermediates are defined as structures which are easily recognizable during the trajectory, which frequently appear prior to the fully folded structure, and that show a certain stability (though, of course, less than the folded conformation itself). Nevertheless, they do not uniquely define a set of consecutive steps which fully determines the folding pathway, since they are observed in different orders for different trajectories. However, the number of folding trajectories computed is large enough for certain statistical preferences to be apparent, as discussed immediately below.

The intermediates (whose structures are depicted among the conformations sketched in Figs. 8 and 10) appear in the form of terminal hairpins (helices I and II or helices III and IV, with the corresponding turn); the central hairpin (helices II and III, plus two dangling noninteracting tails), the long hairpin, in which helices I and IV appear as a continuation of helices II and III, respectively, the double hairpin, with both terminal hairpins formed, but in a linear or almost linear disposition, and the three-helix bundle (plus a dangling tail, which can equally be helix I or helix IV). Usually, the transformations between these intermediates are very fast for model **A**. The formation of the intermediates from unfolded states takes place quite rapidly, beginning in general from one of the turns that join the helices in the folded structure. They frequently dissolve to give random extended conformations, without retaining any native secondary or tertiary structure. From this point of view, they can hardly be considered as true thermodynamic intermediates in the reaction coordinate on going from the unfolded to the folded state, since their population is very small in comparison with both extremes of the reaction coordinate.

All these structures, together with many structureless random conformations, appear and disappear in the first 50,000 to 800,000 MC cycles. It is interesting to observe how the different quantities respond to their presence. Obviously, the dimensions of the chain, as represented by the radius of gyration, continuously change, from the small values corresponding to the three-helix bundle (or even collapsed states without a clear shape) to the larger values of the long hairpin or more extended unfolded conformations. The tertiary contacts follow a similar behavior, but their enumeration allows a more detailed study of the process. For example, the terminal hairpins are characterized by the contacts between helices I and II or helices III and IV alone. If both types of contacts simultaneously appear, without contacts in the zone corresponding to helices II and III or IV and I, this signifies a double hairpin situation. Contacts in the region of helices II and III alone indicate the central hairpin. If they are accompanied by contacts between helices I and IV, there is a long hairpin. Of course, contacts in two contiguous areas (I–II and II–III or II–III and III–IV) indicate the presence of a three-helix bundle. During this first third of the trajectory, there is a three-helix bundle (with helix I excluded) at the beginning, followed by a double hairpin, an almost folded structure (though without I–IV contacts), a three helix bundle again, excluding in this case helix IV, and finally a central hairpin alone.

Formation of the folded conformation for model **A**

The central hairpin constitutes the seed that yields the folded structure in most of the trajectories. The folded structure is closely preceded by the formation of the central hairpin (frequently, with the three-helix bundle as a subsequent intermediate), though the reverse situation does not always happen. Many central hairpins dissolve or evolve towards structures different from the folded one. In this particular trajectory, we find a central hairpin at 700,000 MC cycles. We see also one contact appearing between helices I and II (the one closest to

the turn joining them) and several sporadic contacts between helices III and IV. And then, very rapidly, at 800,000 cycles, the full set of contacts, clearly reflecting the formation of the folded conformation, appears.

How has this last process happened? The tails emerging from the central hairpin quickly come into position, assembling the fourth helix in the bundle. Although the assembly is almost simultaneous for both tails, helix IV comes into position slightly before helix I, producing for a moment the three-helix bundle, though it cannot be considered as a folding intermediate at this point of the trajectory. It is important to notice that, though the strands which will eventually form helices I and IV have the usual relatively high helical content, they are not by any means frozen helices that diffuse into the correct positions. This is evident from the evolution of the energy in Figure 8b. The formation of the folded structure is accompanied by an important reduction of E_{long}, since the number of hydrophobic contacts tremendously increases. A less pronounced reduction in E_{helix} occurs, due to the formation of the helical turns missing in helices I and IV when they are not in the folded structure. Therefore, we can observe the diffusion of parts of the protein model towards the elements of already assembled folded structure, but they correspond to extended elements of structure and not to preformed, frozen helices.

Once formed, the four-helix bundle remains stable for almost 600,000 MC cycles. There are still a lot of movements taking place in this structure. Most characteristic are the relative small shifts of some helices with respect to others, which slightly modify the hydrophobic core of the folded structure. These add temporal contacts between the central hydrophobic residues of helices I and III, or helices II and IV (i.e., across the diagonals of the bundle; see Fig. 1, model **A**). In addition, there are a myriad of local movements, most involving small distortions of the folded structure. The most frequent are the very fast chain tail movements. We also find movements close to the turns between helices, and even deformations inside the helical turns. All these movements are clearly reflected in the energetic terms, which strongly fluctuate even in the folded conformation of our model, though the chain keeps its topology and its compact conformation, as reflected by the very small oscillations of the radius of gyration.

Fluctuations in the folded conformation of model A

Since the folded structure is rather mobile, one could ask whether there is any difference between the oscillations that take place in the folding intermediates and those appearing in the folded conformation. This is important, since some authors have proposed the existence of a molten globule state, prior to the formation of the final folded structure, in which the secondary structure is essentially formed, but the tertiary contacts are still quite reduced.[3] A similar conclusion could be extracted from our results. We have seen that our intermediates have rather stable secondary structure in the portions of the chain they comprise. Also, we can see, by comparing the width of the point clouds representing native contacts before and after the final folding step, that the vibrations existing in the folded structure, even being considerable, exhibit a more reduced amplitude than those existing in the different intermediates. To further clarify this, we present in Figure 9 the root mean square (rms) deviation between the coordinates of the α-carbons along the trajectory and an ideal set of coordinates for the folded conformation (the rms is never equal to zero due to thermal fluctuations). To distinguish between fluctuations in rms due to unfolded parts of the model and those due to local rearrangements, we also present the rms for two large fragments of the protein: the first without including the residues in the designed helix IV, and the second without including the residues corresponding to helix I in the folded structure. Even in the region where the first three-helix bundle (with helices I–II–III) is formed, the oscillations in rms corresponding to the intermediates are larger than those corresponding to the folded structure. This difference is small since our model without side chains, in its simplified realization, still retains excessive mobility in the folded state. Based on this small difference and the different mean lives of the folded structure and the folding intermediates, we can conclude that the formation of the final folded structure brings an increased stabilization to the protein molecule absent in the folding intermediates. The latter can also be quite compact with a lot of secondary structure, appearing prior to the final conformation.

Model with side chains: intermediates in model C

Let us now consider the folding pathways of model **C**, in which an approximate representation of the side chains has been included. In Figure 10, we present the results of a constant temperature trajectory, with the evolution of the radius of gyration, E_{long} (dotted line) and E_{helix} (solid line), and the evolution of the native contacts, now defined between the side chains of the different hydrophobic residues. There are several features in common with model **A**. Namely, there is a very fast formation of secondary structure in the putative helical regions, as indicated by the fast decrease in E_{helix}, and the collapse of the chain into a partially compact structure, with a small value for the radius of gyration R_g and an increase in the tertiary interactions, with the reduction of E_{long}. However, we do not find the multiplicity of short-lived intermediates which characterized the first part of the trajectory in model **A**. We clearly

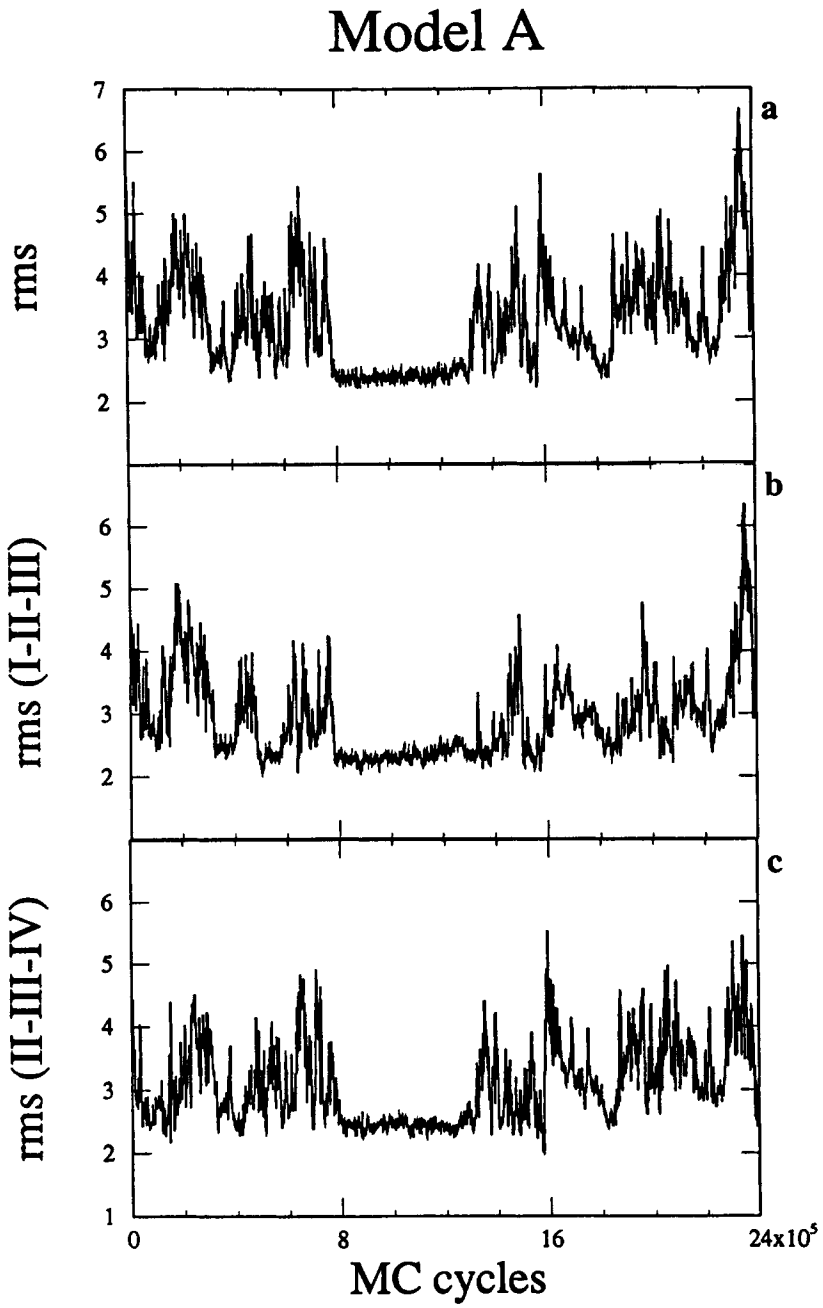

Fig. 9. Root mean square deviation corresponding to the coordinates of the α-carbons of model **A** along a MC trajectory. (**a**) Full chain. (**b**) Residues corresponding to the first three helices (residues 1 to 36). (**c**) Residues corresponding to the last three helices (residues 12 to 48).

see contacts in the regions of helices I–II and II–III, and therefore we are in the presence of a three-helix bundle. It is actually formed from the central hairpin, although the assembly of the N-terminal strand to form the third helix takes place almost simultaneously. A detailed study is presented in Figure 11, where we have magnified the first 5% of the trajectory. The lower part of this figure is identical to those in Figures 8 and 10. The upper part shows the evolution of all the torsional angles in the protein backbone, with the different regions of native secondary structure (helices and turns) separated by the horizontal solid lines. A dot in this plot indicates that a torsional angle is in the range corresponding to the native state. This plot clearly demonstrates the temporal variation of the helical content. The evolution of the strand corresponding to the C-terminus (helix IV in the native conformation) does not

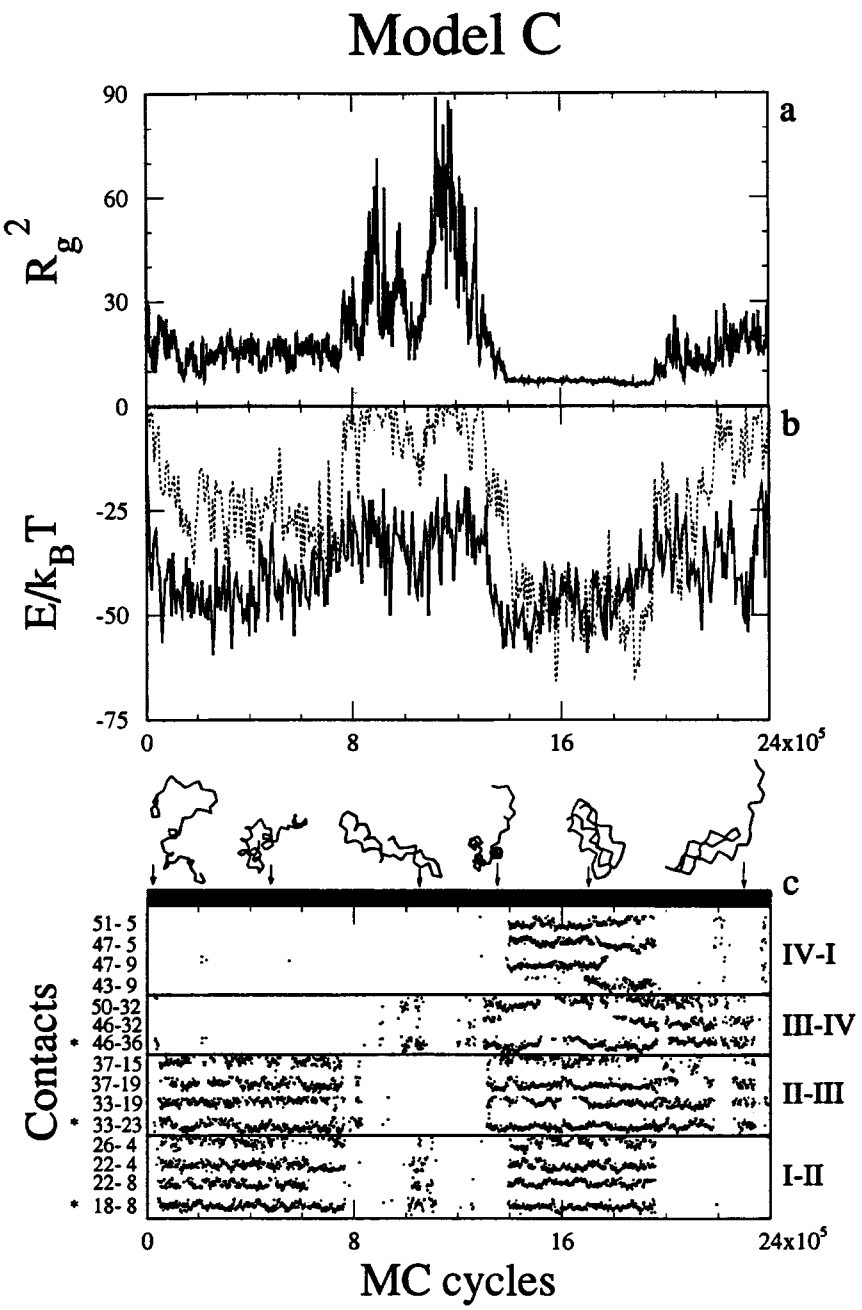

Fig. 10. Folding and unfolding trajectory of model **C**. (**a**) Radius of gyration, R_g, in reduced units. (**b**) Energetic contributions for the helical term, E_{helix} (solid line), and of the tertiary interactions, E_{long} (dotted line) along the trajectory, also in reduced units. (**c**) Evolution of the native contacts between side chains. Asterisks in the vertical axis indicate the contacts closest to the turns between helices. Some snapshots of instantaneous conformations along the trajectory are included. They correspond, successively, to the initial random conformation, a 3-helix bundle (staggered view), a double hairpin, a 3-helix bundle (top view), the folded 4-helix bundle, and a 3-helix bundle (lateral view). The arrows indicate the approximate occurrence of these conformations along the trajectory.

take part in the three-helix bundle which is formed in this portion of the trajectory. Therefore, it shows very fast transitions between the helical and the nonnative torsional states. Thus, the population of native torsional angles is rather large, but one cannot at all say that the helix is preformed.

The same things happen with the other three helical portions of the chain, prior to the formation of the three-helix bundle. The turns between helices are the first to acquire a native conformation, though they are not frozen. The very small number of contacts before 40,000 MC cycles clearly indicates

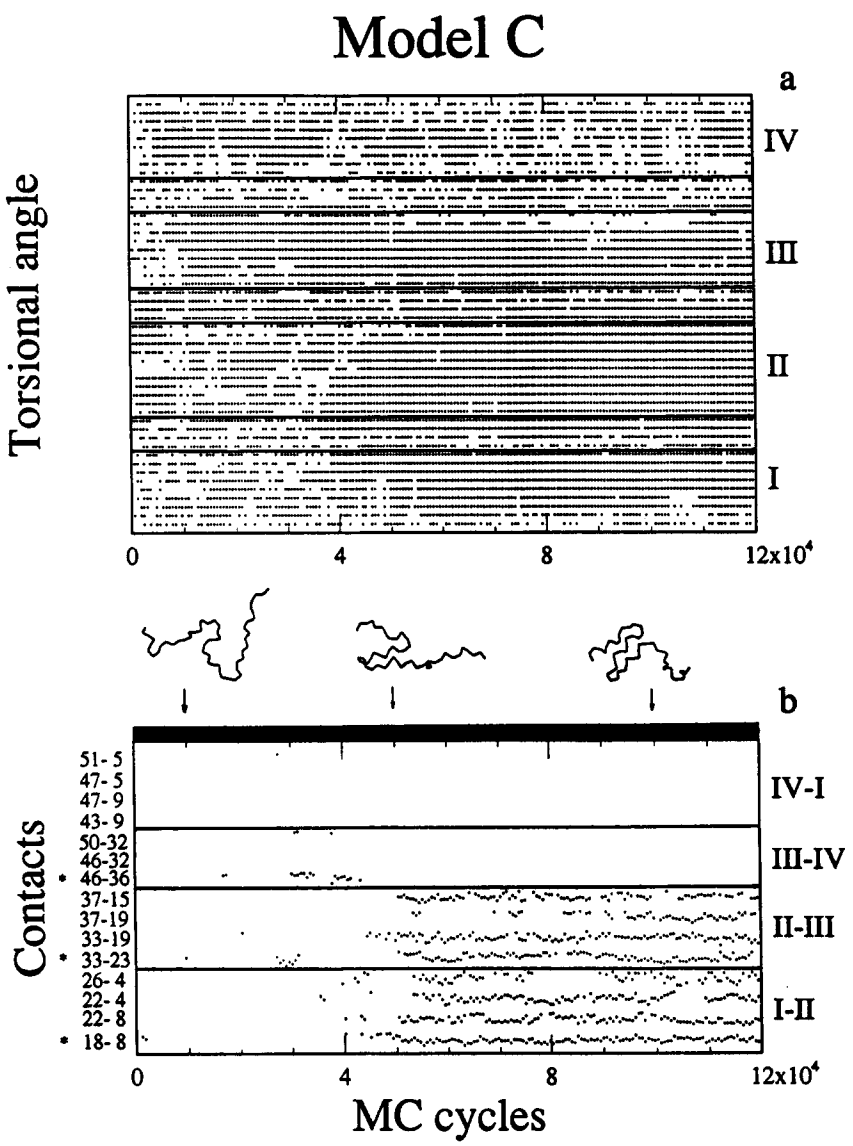

Fig. 11. Initial steps in the folding trajectory of model C. (a) Evolution of the torsional states. One dot in the plot indicates the presence of the torsional angle in the native torsional state. Solid lines separate helical regions from turn regions, according to the folded structure. (b) Evolution of the native contacts between side chains. Snapshots on top of this part of the figure correspond to instantaneous conformations at the time indicated by the arrows.

that the structure is still very expanded. However, strands I, II, and III are beginning to approach one another. When they begin to interact, the attraction between hydrophobic residues and the repulsion between hydrophilic and hydrophobic ones simultaneously creates adequate side-chain packing and correct secondary structure. Observe how the torsional angles of helix II, together with fragments of helices I and III, exhibit substantially reduced fluctuations when the native tertiary contacts appear. Therefore, the formation of the native secondary and tertiary structures is a very cooperative process, which takes place almost simultaneously in the conditions represented by our model, and rather possibly in small globular proteins as well.

The three-helix bundle is rather mobile, as indicated by the width of the contact clouds in Figure 11, but even with local distortions, it is able to survive almost the first third of the trajectory. In essentially all the trajectories computed with this model (about 60 of them), the three helix bundle appears and exhibits a certain intrinsic stability, independent of the fact that it can directly yield the folded structure or dissolve. The terminal helix which first dissolves (Fig. 11), while a distorted central hairpin (represented by the sparse contacts in the region of helices

II–III) survives slightly longer. Nevertheless, all the tertiary interactions disappear a number of MC cycles later, and the chain acquires a very extended structure, with a large radius of gyration. It is not only the energy contribution E_{long} which decreases or increases when the protein model goes from a folding intermediate to an extended conformation or vice versa. Also the value of E_{helix} decreases (has a more negative value) when the intermediate appears and increases when it dissolves. The energy does not go to zero, reflecting the high average helical content, but the observed fluctuations clearly indicate that no frozen elements of secondary structure are diffusing to form or dissolve the tertiary contacts.

Formation of the four-helix bundle: fluctuations in model C

From 825,000 to 1.2 million MC cycles, no recognizable structure appears, with the exception of a double hairpin (both terminal hairpins), which, being unstable, has a rather short lifetime. At about 1.3 million MC cycles, we find again a three-helix bundle, this time including helix IV, with the strand corresponding to helix I staying out of the bundle. This situation does not last too long, and very soon this strand assembles into the desired four helix bundle.

This structure also exhibits oscillations, mainly due to the tails of the terminal helices, and the turns that join the helices, which are very flexible in this model (they are not stabilized by the helical contribution of the potential, nor do they take part in tertiary attractive interactions). But a new important *distortion* appears with respect to model **A**, and it is the angle formed between the terminal helices I and IV and the helices II and III that constitute the central hairpin (see Fig. 5, model **C**). This angle approximately oscillates between 10 and 40° (an accurate measurement is very difficult, since the helical structure is not frozen even in the folded state). It is mainly due to the packing of the side chains, as occurs in real proteins with helical contacts.[16] It is quite encouraging to observe how a very simplified model as ours can even reproduce these details of the physical system.

In this case, as well as in the trajectory for model **A**, the oscillations in the native contacts are less pronounced when the whole folded structure forms. On the other hand, if we examine the rms between the Cartesian coordinates of the backbone α-carbons in the trajectory snapshots and in an ideal reference conformation for the folded state of model **C** (Fig. 12), it is difficult to distinguish between the three- and the four-helix bundles. Therefore, for model **C**, it is not possible to clearly address the main difference between the local structure in the folding intermediates and in the final conformation. It seems that the backbone topology (and therefore the secondary structure) is formed, but the packing of the side-chains does not achieve its perfect matching until the full structure folds. Nevertheless, we are using short side chains which are rigidly attached to the backbone, a fact which couples the two situations, partially blurring the scheme presented above.

Comments on the three-helix bundle

The high population found in this trajectory for the three-helix bundle requires some additional comment, since it poses some fundamental differences with the traditionally accepted thermodynamic theory for the folding of small, single-domain, globular proteins.[4] This theory assumes a two-state transition, in which the completely folded conformation and the unfolded state (the latter represented by a multiplicity of conformations) are the only states with a significant population. In model **A**, we can also observe some intermediates, but we have already seen how fast the transitions are among them, and the short time which they live, considered on an individual basis or in aggregate. In model **C**, on the other hand, the three-helix bundle, at least in this particular trajectory, shows a mean lifetime almost comparable to the folded state. The situation is not so striking when this analysis is extended to all the folding trajectories computed for model **C**. As an average, the three-helix bundle appears during 21% of the trajectory length, while the folded and the unfolded states have values of 43 and 36%, respectively. The population of other intermediates is negligible, since they usually appear as instantaneous conformations in the unfolded state. Probably, the reason for this large population of the intermediate structure is purely kinetic, and does not modify the thermodynamics of the generic folding process. It is clear from the trajectories that, during the life of the three-helix bundle, the fourth strand tries to assemble into the correct conformation, but problems related with side chain packing, the balance between attractive and repulsive tertiary interactions, the flexibility of the turns between helices, and perhaps the average high helical content as well, considerably reduce the chances for this last assembly process to occur. From this point of view, the three-helix bundle might be considered as a kinetically frustrated state.

SUMMARY AND CONCLUSIONS

In this paper, we have studied the folding pathways of protein models representing four-helix bundles, both with and without side chains. First, the question of designing a primary structure, even at a very simplified level, was undertaken. It is clear that, depending on the resolution of the model one is dealing with, the sequence must possess a number of specific features to ensure folding to a unique, native like state. This includes the proper balance be-

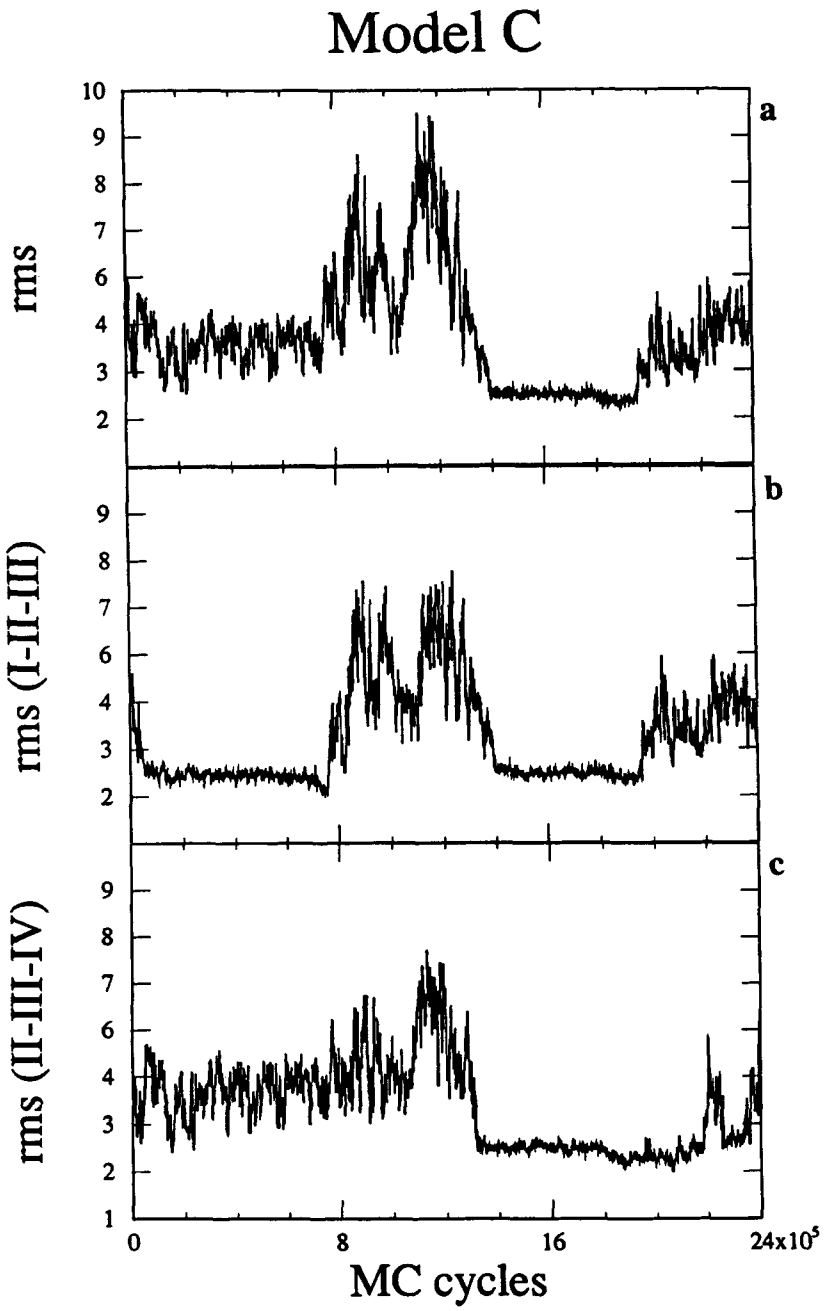

Fig. 12. Root mean square deviation corresponding to the coordinates of the α-carbons of model **C** along a MC trajectory. (**a**) Full chain. (**b**) Residues corresponding to the first three helices (residues 1 to 40). (**c**) Residues corresponding to the last three helices (residues 14 to 53).

tween the different contributions of the local and tertiary interactions defined for the model. The proper packing of the side chains, at least in a simplified spherical representation, also obliges one to carefully choose the range of attractive and repulsive interactions. The former have to be long ranged enough to provide a realistic folding pathway that is accessible in a reasonable amount of time. However, they must also have a rapidly decreasing magnitude to avoid large potentials far from the interaction center, which only leads to a lack of general specificity, and hence to dense collapsed states without secondary structure. The repulsive interactions are partially responsible for the correct geometry by controlling the packing of the side chains. Since these are probably too fat in the spherical representation, this part of the interactions has to be reduced to avoid locking along plausible folding pathways,

and to allow for the folding to occur in a reasonable amount of computer time.

As mentioned above, the proper balance between secondary and tertiary interactions is also essential. We have observed that a relatively high population of secondary structure is necessary in order to find unique folded conformations. While experiments on real proteins are also beginning to show this, the helix content appearing in our trajectories is clearly far too high. There is a reason for this. We need to artificially enforce the native torsional state of the turns between helices because otherwise they are free to change their conformation without any interactions other than those related to the chain connectivity. A better representation of the model in which these turns are stabilized by internal hydrogen bonds or favorable interactions of the side chains would not require this arbitrary component of the potential. Indeed, in some very recent on-lattice work, these arbitrary components can be entirely eliminated, and yet the topological aspects of folding pathways seen here are retained.[17] Furthermore, there are some experiments which indicate that a moderate intrinsic turn population (5–10%) in helical proteins may be physical.[18,19] For helical regions, a more accurate description of the hydrogen bonds would also probably result in less helix content, and more specific tertiary interactions would still allow for a proper packing of the side chains to give a unique folded state. However, even in our model, the relatively large and artificially high helical content in the unfolded state constitutes a very dynamic situation. It is only the formation of the folded conformation that fully determines both secondary and tertiary structures, in a very cooperative process. The adequate packing of the side chains in the last steps of the folding pathways seems to require a certain mobility, which cannot be found in preformed elements of frozen secondary structure.

The formation of the folded conformation in our model does not happen through a random search, especially for the model with side chains, nor it is a simple process of collapse and further rearrangement to grow the secondary structure. A series of intermediates, some of which appear in each and every successful folding trajectory, can be clearly identified. For a model without side chains, there are several possible sequences of intermediates that yield the folded structure, the more frequent being the central hairpin and/or the three-helix bundle. In a model with side chains, on the other hand, the three-helix bundle *always* appears immediately before the final folded conformation is achieved. We have observed some trajectories in which a central hairpin or a long hairpin try to assemble both terminal helices simultaneously. This mechanism, which actually happens (though it is not the most frequent) for model **A**, seems to be precluded when side chains are considered, precisely due to the difficulties in properly packing a large number of side chains at the same time. While this effect might be partially due to the rigid representation of our side-chains with respect to the backbone, it is important to mention that similar folding mechanisms have been observed with different realizations of the model and the MC scheme itself.[7,9,17] Therefore, a certain model independence can be assumed for the general features of our folding pathways, which will then be a natural consequence of the physics of the model.

The above conclusions can be considered as a set of basic recipes for the design of more sophisticated protein models. Indeed, they are currently being employed in the folding of simplified proteins such as the DeGrado four-helix bundles[10] where the only information contained is the amino acid sequence.[17] The qualitative features described here are retained in this more general case. Beyond this, the set of ingredients required to fold these models constitutes a set of very generic observations which may be perfectly valid for the design of real sequences, at least for small globular proteins. Our model is, of course, only a very simplified representation of a real protein, but it has been formulated in the spirit of trying to avoid complicated details, not to create a different (and irrelevant) kind of reality. Further developments of this model are presently under way, and there are substantial indications that these expectations are confirmed.

ACKNOWLEDGMENTS

This work was supported in part by Grant GM-37408 from the Division of General Medical Sciences, National Institutes of Health. A.R. also acknowledges a Postdoctoral M.E.C.-Fulbright Scholarship from the U.S.–Spanish Joint Committee for Cultural and Educational Cooperation. The authors gratefully acknowledge IBM Corp. for providing a portion of the computational resources employed in this work.

REFERENCES

1. Chan, H.S., Dill, K.A. Origins of structure in globular proteins. Proc. Natl. Acad. Sci. U.S.A. 87:6388–6392, 1990.
2. Kolinski, A., Skolnick, J. Discretized model of proteins. I. Monte Carlo Study of Cooperativity in homopolypeptides. J. Chem. Phys., in press.
3. Kuwajima, K. The molten globule state as a clue for understanding the folding and cooperativity of globular protein structure. Proteins 6:87–103, 1989.
4. Creighton, T.E. In "Protein Folding." Gierasch, L.M., King, J., eds. Washington: American Association for the Advancement of Science, 1990:157–170.
5. Skolnick, J., Kolinski, A. Computer simulations of globular protein folding and tertiary structure. Annu. Rev. Phys. Chem. 40:207–235, 1989.
6. Presnell, S.R., Cohen, F.E. Topological distribution of four-α-helix bundles. Proc. Natl. Acad. Sci. U.S.A. 86:6592–6596, 1989.
7. Sikorski, A., Skolnick, J. Monte Carlo studies on equilibrium globular protein folding. III. The four helix bundle. Biopolymers 28:1097–1113, 1989.
8. Rey, A., Skolnick, J. Comparison of lattice Monte Carlo

dynamics and Brownian dynamics folding pathways of α-helical hairpins. Chem. Phys. 158:199–219, 1991.
9. Skolnick, J., Kolinski, A. Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. J. Mol. Biol. 221:499–531, 1991.
10. DeGrado, W.F., Wasserman, Z.R., Lear, J.D. Protein design, a minimalist approach. Science 243:622–628, 1989.
11. Sander, C., Vriend, G., Bazan, F., Horovitz, A., Nakamura, H. Ribas, L., Finkelstein, A.V., Lockhart, A., Merkl, R., Perry, L.J., Emery, S.C., Gaboriaud, C., Marks, C., Moult, J., Verlinde, C., Eberhard, M., Elofsson, A., Hubbard, T.J.P., Regan, L., Banks, J., Jappelli, R., Lesk, A.M., Tramontano, A. Protein design on computers. Five new proteins: Shpilka, Grendel, Fingerclasp, Leather, and Aida. Proteins 12:105–110, 1992.
12. Eisenberg, D., Wilcox, W., Eshita, S.M., Pryciak, P.M., Ho, S.P., DeGrado, W.F. The design, synthesis and crystallization of an α-helical peptide. Proteins 1:16–22, 1986.
13. Flory, P.J. "Statistical Mechanics of Chain Molecules." New York: Wiley, 1969: chapt. 3.
14. Lee, S., Karplus, M. Brownian dynamics simulation of protein folding: A study of the diffusion-collision model. Biopolymers 26:481–506, 1987.
15. Gregoret, L.M., Cohen, F.E., Protein Folding. Effect of packing density on chain conformation. J. Mol. Biol. 219: 109–122, 1991.
16. Branden, C., Tooze, J. "Introduction to Protein Structure." New York: Garland, 1991: chapt. 3.
17. Kolinski, A., Godzik, A., Skolnick, J. Discretized model of proteins. II. Prediction of the three dimensional structure and folding pathway of simple helical proteins. J. Chem. Phys., submitted.
18. Dyson, H.J., Rance, M., Houghten, R.A., Lerner, R.A., Wright, P.E. Folding of immunogenic peptide fragments of proteins in water solution. J. Mol. Biol. 201:161–200, 1988.
19. Dyson, H.J., Merutka, G., Waltho, J.P., Lerner, R.A., Wright, P.E. Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding. I. Myohemerythrin. J. Mol. Biol. 226:795–817, 1992.