Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 2629–2633, April 1992
Biochemistry

Simulations of the folding pathway of triose phosphate isomerase-type α/β barrel proteins

ADAM GODZIK, JEFFREY SKOLNICK*, AND ANDRZEJ KOLINSKI

Department of Molecular Biology, The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, CA 92037

Communicated by Marshall Fixman, December 23, 1991

ABSTRACT Simulations of the folding pathways of two large α/β proteins, the α subunit of tryptophan synthase and triose phosphate isomerase, are reported using the knight's walk lattice model of globular proteins and Monte Carlo dynamics. Starting from randomly generated unfolded states and with no assumptions regarding the nature of the folding intermediates, for the tryptophan synthase subunit these simulations predict, in agreement with experiment, the existence and location of a stable equilibrium intermediate comprised of six β strands on the amino terminus of the molecule. For the case of triose phosphate isomerase, the simulations predict that both amino- and carboxyl-terminal intermediates should be observed. In a significant modification of previous lattice models, this model includes a full heavy atom side chain description and is capable of representing native conformations at the level of 2.5- to 3-Å rms deviation for the Cα positions, as compared to the crystal structure. With a well-balanced compromise between accuracy of the protein description and the computer requirements necessary to perform simulations spanning biologically significant amounts of time, the lattice model described here brings the possibility of studying important biological processes to present-day computers.

In protein molecules, a variety of dynamic processes take place, with characteristic lifetimes from femtoseconds to minutes. So far, no theoretical method has been able to study the dynamic behavior of realistic protein models for times longer than hundreds of picoseconds. Therefore, many interesting and important processes, most notably protein folding, have remained beyond the reach of direct theoretical modeling. Molecular dynamics simulations, although very successful at describing local, short-time processes, would have to span at least an additional four orders of magnitude in time before the process of protein folding could be examined (1).

Here, we describe an advance in the development of a model designed to extend simulation techniques to include long-time, large-scale processes. This model describes the protein as a set of beads embedded in an underlaying cubic lattice. Cα atoms are confined to a specific sublattice, the "210" lattice, and points describing the backbone and amino acid side chains are chosen to give the correct shape and volume of the protein (refs. 2 and 3 and refs. cited in ref. 3). This description of protein structure, when coupled to a dynamic Monte Carlo algorithm, permits study of the long-time, global rearrangements of protein structure on a small work station, and yet the model is sufficiently complex to describe protein structure with a surprising level of accuracy.

Earlier work has shown that even the simplest lattice model correctly reproduces the all-or-none characteristics of the folding process, and substantial effort was put into establishing the conditions for simulating the folding to a unique, native structure (2). The next generation semiquantitatively reproduced the structure and folding of a relatively small protein—plastocyanin (ref. 3 and refs. cited therein). Here, the model is generalized to include a more-realistic, full heavy-atom description of the side chains, and much larger proteins are studied. Taken together, these improvements to the model make it possible to study the long-time behavior of realistic models of large protein systems.

We have attempted to simulate the folding of two large (247 and 267 amino acids) mixed-motif α/β proteins. The major objective of this work is to prove that, without any assumptions concerning the nature of folding intermediates, the model can reproduce the experimental results for one of these molecules; for the other, it makes predictions about the nature of the equilibrium intermediate. Further, it is shown that the present model can closely reproduce the native structure of any protein, both in terms of the rms deviation for the Cα postions and in other similarity measures, as discussed below. This constitutes a rather rigorous test of the entire methodology.

The two proteins studied in this paper, chicken triose phosphate isomerase (TIM) and the α subunit of tryptophan synthase (WSY) from *Salmonella typhimurium*, have the same $(\beta/\alpha)_8$ barrel fold. This motif, called a "TIM barrel," has been found in more than 17 different proteins, with almost no detectable sequence homology between some of them (4). This, along with other experimental data (5), suggests that the rules defining this structural motif may be sufficiently robust to be correctly described by a simplified model (6). Another interesting point is that, according to recent experimental evidence, there are stable, long-lived folding intermediates along the folding pathway of a closely homologous WSY from *Escherichia coli* (refs. 7 and 8 and refs. cited therein). Thus, this protein class seems to be an exception to the typical all-or-nothing transition observed in single-domain, globular proteins (9).

The available experimental evidence on the folding pathway of the WSY is as follows: the amino-terminal fragment, consisting of residues 1–188, as cleaved by trypsin, can fold independently from the rest of the protein. If the fast step in folding the whole protein proceeds in tritiated water and the labeled protein is digested with trypsin, almost 90% (43%) of the hydrogen bonds in the amino- (carboxyl)-terminal fragment are screened from the solvent (refs. 7 and 8 and refs. cited therein). Further mutagenic studies have allowed for a detailed assessment of specific interactions associated with the folding pathway (10). To the best of our knowledge, such studies have not yet been performed for TIM. Thus, our identification of the folding intermediates for TIM is a prediction to be experimentally tested.

THE MODEL

The latest version of the 210 lattice model describes the protein as a collection of beads on a cubic lattice with a lattice

Abbreviations: WSY, α subunit of tryptophan synthase; TIM, triose phosphate isomerase.

*To whom reprint requests should be addressed.

spacing of 1.7 Å. Cα positions are connected by (2,1,0)-type vectors, so that the nearest-neighbor Cα–Cα distance equals 3.8 Å, and the next-nearest-neighbor Cα–Cα distance spans the range observed in real proteins. The six adjacent cubic lattice points around each Cα atom are occupied, providing a backbone of the correct thickness. The side chain conformations depend solely on the backbone conformation and are constructed as a projection onto the cubic lattice of the most probable side chain conformation for a given backbone conformation. In particular, for a given amino acid side chain, whose center of mass lies closest to the mean position for all such side chains in a protein; a structural data base is used (for details, see ref. 2). Two different side chain representations were used in the simulations described below. In the first, the volume of the lattice side chain was close to that of the real side chain, but the number of points was slightly larger than the number of heavy atoms in real side chains, ranging from 17 for tryptophan to 2 for alanine. The second representation has the number of points equal to the number of side chain heavy atoms. The side chains interact by means of a "contact" potential; i.e., only those side chains that lie within the interaction envelope contribute to the nonbonded energy. Simulations have been performed with two different realizations of this energy term, and the interaction envelope thickness was taken to be 1 or 2 lattice units, respectively. Larger side chains were used in conjunction with a smaller interaction range and vice versa. In the latter, the number of interactions in the final structure was ≈50% larger than in the former.

The interaction between amino acid side chains is implemented using a pairwise potential of mean force, based on the Miyazawa–Jernigan statistical hydrophobicity scale (11). All possible tertiary interactions are allowed; i.e., the interaction energy of a particular side chain pair depends on their identity only and not on whether this pair interacts in the native protein. The energy also includes preferences for local distances between Cα atoms that are two, three, or four residues apart, and in addition there is a preference for local chain chirality. These interactions represent the intrinsic propensity for local secondary structure and are based on the structure of the real protein projected on the lattice. The introduction of this latter term is necessary to compensate for oversimplification in describing the local interactions and is explained in detail in ref. 2. Because information about the relative differences in local preferences for native structure is lacking, their magnitude is assumed to be uniform over the entire sequence—this is clearly an oversimplification. Experimental studies indicate that for certain protein fragments it is possible to detect a significant population of native structure in solution, while for other fragments this is not the case (12). *Assuming a uniform preference for local secondary structure consistent with the native fold* allows us to study only those aspects of the folding pathway that result from the folding kinetics due to nonlocal, tertiary interactions or that are sufficiently robust that they show up even in a simplified model.

The Monte Carlo program uses both local moves (2–8 atoms) and moves involving relative motions of arbitrarily chosen sections of the molecule (2). In each Monte Carlo step, at least one attempt is made to move each atom. Because of the possibility of nonlocal rearrangements, the number of cycles is not linearly proportional to time; however, it is monotonic. Using only local rearrangements, one Monte Carlo step would introduce changes equivalent to these occurring in about 1–10 ns (10^{-9} sec). Similarly, larger steps are equivalent to 10–200 ns, depending on the size of the element involved in the rearrangement. Since all movements are confined to a lattice, the majority of calculations are performed using integer arithmetic, with many time-consuming operations being precalculated. This gives the simulation program a three- to four-order of magnitude speed up over conventional, off-lattice molecular or Brownian dynamics simulations, therefore enabling us to study folding and other slow processes.

THE SIMULATION

The simulations consist of three steps: (*i*) obtaining a "native" lattice structure, (*ii*) unfolding it to a random conformation, and (*iii*) refolding it using the Monte Carlo algorithm.

The lattice representations of the two proteins, based on known crystallographic structures deposited with the Protein Data Bank [Brookhaven National Laboratory; entries 1WSY (13) and 1TIM (14), respectively], have been obtained by projecting the native protein structure onto the cubic lattice. The number of contacts in the refined lattice structure is comparable to that in the native structure, where a contact between two amino acids is counted whenever any two side chain atoms lie closer than 3.5 Å. Comparisons of the lattice and crystal structure of the WSY and TIM are presented in Fig. 1 *Upper* for the model with an interaction range of 2 lattice units (3.4 Å) and the smaller side chain representation. The lattice structure for WSY (TIM) has a 3.0- (2.7)-Å rms deviation for the Cα positions from the native structure, and the number of side chain–side chain contacts equals the number of contacts in the real protein with the same interaction range. About 20% (15%) are native; the rest are shifted in register by one or two positions. In agreement with recent theoretical (15) and experimental (16) findings, there appears to be little specificity in the internal packing, and such shifts allow for better packing in the lattice, while keeping the type of contacts virtually unchanged. Fig. 1 *Lower* shows the difference-distance maps between the lattice and the real TIM (WSY) structure. Clearly such differences are distributed uniformly over the whole structure and result from discretization of the lattice "protein." Recovery of the full atom description from the lattice structure is described in ref. 17.

By performing long (500,000 steps) Monte Carlo simulations at high temperature, the folded lattice structures of both proteins have been thermally unfolded; these runs effectively randomize the structure. Moreover, unfolding runs were performed separately prior to each folding simulation.

Starting from such a randomly generated unfolded state, devoid in all cases of any significant amount of secondary structure, isothermal folding simulations were performed either until the protein folded to the native lattice structure or until the simulation length reached 10,000,000 steps (each run requires ≈100 hr of central processing unit time on a Sun Sparc 2+ workstation). The folding temperatures (around 0.55 in internal units) were found by trial and error and on the basis of experience with other protein systems. In the majority of cases, folding occurred within 4,000,000–8,000,000 steps, depending on the starting conformation and the model variant. For WSY, which has a more complicated topology than TIM, four simulations ended in a partially folded structure, where 70–80% of final structure was folded and probably trapped in local minimum. The information for all runs is summarized in Table 1.

The trajectory consists of snapshots taken every 1000 steps, and the easiest way of analyzing such a trajectory is to display it as a movie. Since each folding trajectory displays some distinct characteristics, only general observations common to all trajectories are presented here.

Stage 1. During the first 100,000–200,000 steps, the system rapidly equilibrates. During that time 10–20% of the final number of contacts appear, but most are nonnative. The existence of this equilibration step is due to the high-temperature Monte Carlo runs used to unfold the structure. The conformation after the first 100,000 steps is shown in Fig. 2*A*.

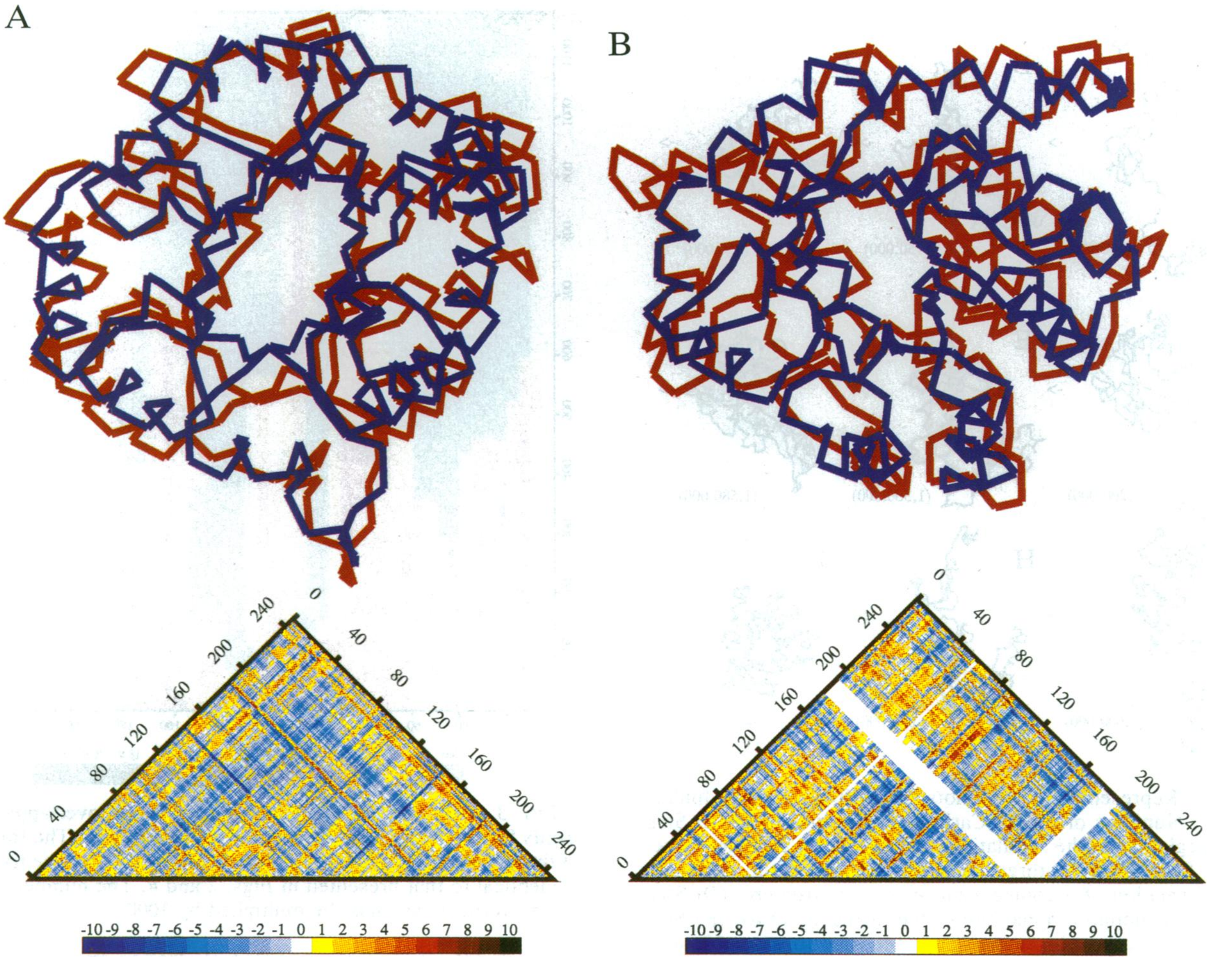

FIG. 1. (*Upper*) Comparison between the lattice (red) and the crystal (blue) structures of TIM (*A*) and WSY (*B*). (*Lower*) Difference-distance maps (scaled in Å) between lattice and crystal structures for the two molecules. The experimental structure of WSY is not defined for residues 56–58 and 178–191; these areas are shown as white stripes on the difference-distance map.

Stage 2. In the first 1,000,000 steps, fragments of correct structure form and dissolve. They usually start as short (5- to 10-residue-long) fragments that grow by an on-site mechanism (3) and then join together by diffusion (18). Several such independent events occur in the whole protein, and two to three finally succeed. (See Fig. 2*B* for a snapshot of the structure with three such domains, at step 750,000, and Fig. 2*C*, where the first attempt to make a compact structure is shown at step 850,000.) This early folding process is studied for the first 1,200,000 steps of simulation in Fig. 3, where the rms deviation from the final structure is displayed for overlapping 15-residue fragments in the form of a "three-dimensional" plot in which the horizontal axis is the number of Monte Carlo steps, the vertical axis represents the position of each fragment in the protein sequence, and the value of the rms deviation is displayed in color, from black (rms deviations >22 Å) to light blue (no deviation). Folding is not unidirectional; examples of fragments of correct structure forming and dissolving are clearly visible as light blue "islands." By the end of the time shown (i.e., 1,200,000 steps), the amino-terminal domain fragment has adopted an almost native structure, with one defect between residues 70 and 85, which is visible in the figure as a white patch.

Stage 3. After 500,000–2,000,000 steps, the first large fragment of the final structure emerges. It always forms near the center of the chain and usually involves 50–70 amino acids from strands 3 and 4, together with adjoining helices. At the end of this stage, there are usually three large fragments of completely assembled structure. They form independently and are separated by 20–30 residues of unfolded chain. In addition to correct side chain interactions, these microdomains are stabilized by some nonnative contacts. This effect is illustrated in Fig. 4, in which both the total number and the number of correct contacts are plotted as a function of Monte Carlo steps. As shown, the total number of contacts

Table 1. Summary of simulations

Protein	Length of simulation, steps	Final rms deviation,* Å	Interaction length	Side chains†	Type of intermediate
WSY	8,000,000	4.8	1	Large	Amino terminal
	8,000,000	4.8	1	Large	Amino terminal
	5,000,000	3.0	2	Small	Amino terminal
	5,000,000‡	3.0	2	Small	Amino terminal
	5,000,000	3.0	2	Small	Amino terminal
	10,000,000	8.2	1	Large	Amino terminal
	10,000,000	13.3	1	Large	Amino terminal
	5,000,000	21.0	2	Small	Carboxyl terminal
	5,000,000	25.4	2	Small	Carboxyl terminal
TIM	10,000,000	4.9	1	Large	Carboxyl terminal
	10,000,000	4.9	1	Large	Carboxyl terminal
	5,000,000	2.7	2	Small	Amino terminal
	5,000,000	2.7	2	Small	Amino terminal

*For the superposition between Cα carbons.
†Within each group the final structure is unique, and rms deviation between folded structures with large and small side chains is equal to 3.4 Å.
‡This simulation is further described in Figs. 2–4.

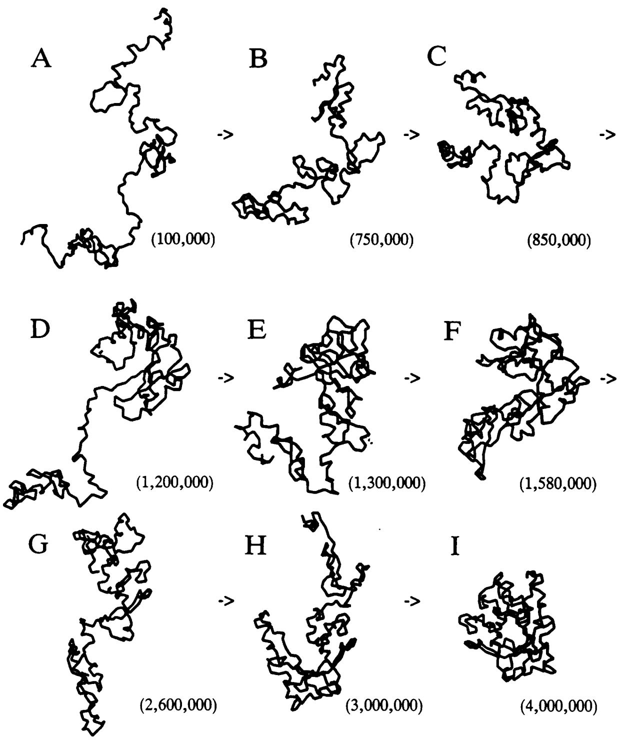

FIG. 2. Representative snapshots taken from a WSY folding trajectory. Numbers of Monte Carlo steps are indicated. (*A*) Step 100,000, beginning of the simulation, the completely unfolded structure. (*B*) Step 750,000, three separate domains form. (*C*) Step 850,000, formation of a compact but incorrect structure. (*D*) Step 1,200,000, the amino-terminal half of the molecule forms. (*E*) Step 1,300,000, the carboxyl-terminal fragment tries to assemble in a nonnative location. (*F*) Step 1,580,000, another attempt to form a compact structure. (*G*) Step 2,600,000, the amino-terminal domain must unfold to accommodate the rest of the molecule. (*H*) Step 3,000,000, the carboxyl-terminal half has formed. (*I*) Step 4,000,000, the native molecule is formed.

is sometimes as much as 30% larger than the number of correct contacts. The breaking of these additional contacts and the simultaneous coalescence of the separate domains is the rate-limiting step in the simulation. The long-lived plateau from 1,200,000 to 3,100,000 steps is associated with the formation of the amino-terminal fragment. The short-lived, slightly higher plateau from 3,100,000 to 3,300,000 steps illustrates the last events in the folding, which in this particular simulation occurred in two steps. In other simulations, this last step is not so pronounced.

Stage 4. After 1,000,000–2,000,000 steps, a larger structure appears, constructed from two of the three domains formed previously. Now, 60–70% of the folded conformation is present. For tryptophan synthase, it always consists of the amino-terminal half of the molecule, while for TIM, it may be either the amino- or the carboxyl-terminal half, depending on the simulation. In the representative simulation, this large intermediate formed earlier (Fig. 2*D* at the step 1,200,000).

Stage 5. This fragment exists for the next 1,000,000–4,000,000 steps, until the rest of the molecule assembles. The assembly of an incorrect compact structure is shown in Fig. 2 *E* and *F*, while the final stages in the assembly of a correct structure, which in this particular simulation involved the complete opening of the amino-terminal intermediate, are shown in Fig. 2 *G* and *H*. For WSY, the helices at both ends are fluctuating, even after the whole molecule is assembled. It is worth noting that for most of the simulation the protein is quite compact and major rearrangements occur during rare but rapid fluctuations.

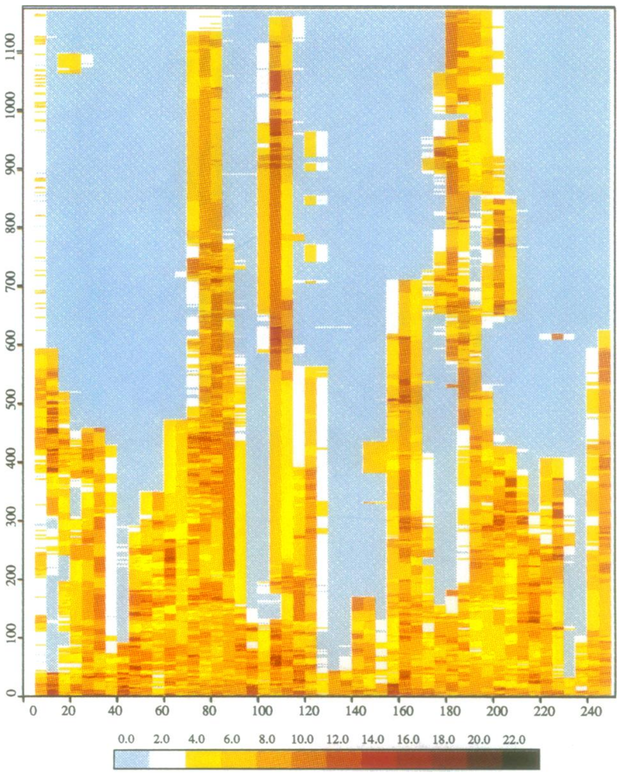

FIG. 3. Plot of rms deviation of 15-residue, overlapping fragments for the first 1,200,000 steps of the simulation. The formation of the amino-terminal fragment can be followed here. The trajectory is identical to that presented in Figs. 2 and 4. The number of steps on the vertical axis must be multiplied by 1000.

Fragments that consist of two β strands and from one to three helices have a significant intrinsic stability. They form and dissolve spontaneously, and it takes a combination of at least two such elements to form a stable structure that usually does not dissolve until the end of simulation. On the other hand, three such fragments tend to adopt a closed, partially nonnative conformation with helices exposed to the solvent and a hydrophobic core built from the β strands. The final step involves the dissolving of nonnative contacts and usually results from a large energy fluctuation. Therefore, the slow step in the assembly of the whole molecule is the opening of the six-membered β-barrel intermediate whose hydrophobic core is partially covered by helices in a nonnative conformation and incorporating the last two strands. As suggested by these simulations, the six-membered β-barrel should form a stable structure. All six-membered β barrels observed in nature are built from two sheets (19), and one may wonder why a symmetrical TIM like structure has not been observed in any existing natural protein.

CONCLUSIONS AND DISCUSSION

The folding behavior of the 210 lattice model of WSY is consistent with experiment. The fact that the overall features of the folding are reproduced by the current model suggests that the difference in tertiary interaction stability between the amino- and the carboxyl-terminal domains may be partly responsible for the experimentally observed folding intermediate. Close examination of the crystal structure of WSY indicates that the carboxyl-terminal fragment has a smaller number of binary contacts between side chains than the amino-terminal one. For TIM, there is no real distinction between the different protein fragments. This is true for both solved TIM structures [from chicken (13) and *Trypanosoma brucei* (20)]. Consistent with this, folding may proceed by

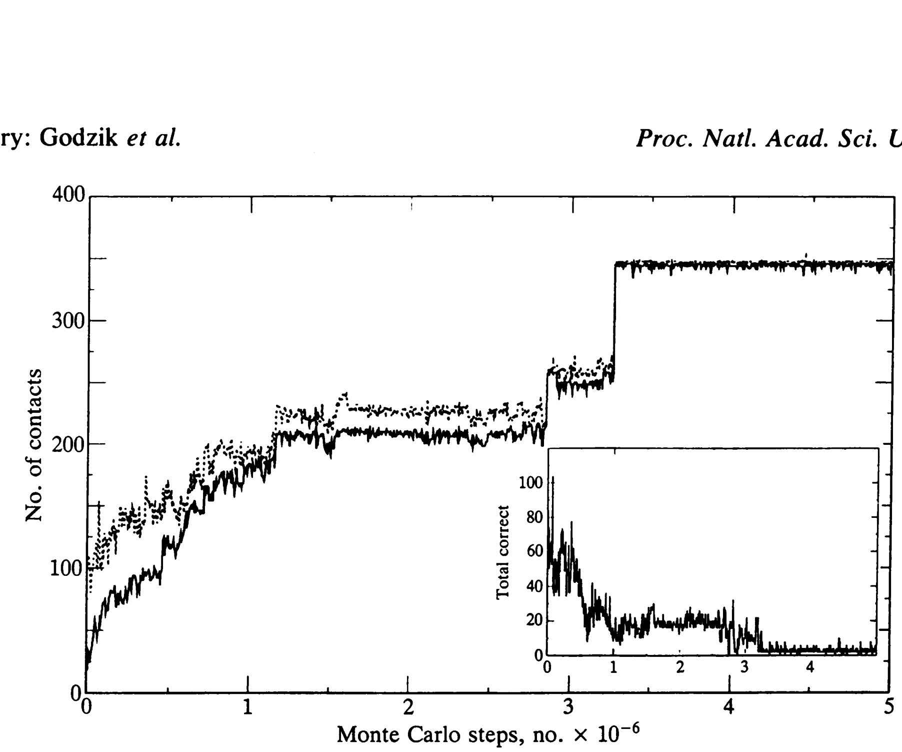

FIG. 4. Plot of the total number (——) and the number of correct (- - -) contacts along the folding pathway WSY. (*Inset*) Difference between these two numbers.

either the amino- or the carboxyl-terminal intermediate. We cannot exclude the possibility that the differential preferences for the secondary structure, which for lack of information are assumed to be uniform in the simulations presented here, may differentiate between these two pathways. While the existence and general character of a stable folding intermediate may be a common feature of TIM barrel proteins, its exact location may depend on the particular protein.

A potentially worrisome problem is the effect of the lattice on the folding pathway. This has been investigated by comparing the dynamics of model protein folding on different lattices [diamond (21, 22) and the 210 lattice (2, 3)] and by performing off-lattice, Brownian dynamics calculations based on a simplified version of the model (23). It turns out that the general characteristics of the folding pathways for simple motifs do not depend on the particular lattice used nor in fact on whether the system is on a lattice at all (23).

The overall picture of the folding process as shown by the present simulation can be described as starting from several independent initiation sites that grow by an "on site" mechanism and are later joined by diffusion. The protein forms a compact structure during most of the folding, and major rearrangements occur during rare but rapid fluctuations.

The simulations presented here show that a lattice model of protein structure makes possible the study of large-scale motions, including folding pathways. The present discretized description of protein structure preserves the most important characteristics of proteins: the tendency of hydrophobic residues to cluster inside of the protein, the packing patterns, and the local folding characteristics of α-helices and β-strands. Clearly, this makes lattice models of proteins the platform of choice for studying long-time processes. With development of the model reported here, we conclude that lattice models of proteins have come of age and are no longer confined to simple, idealized structures.

Stimulating discussions with Drs. H. J. Dyson, W. Beers, R. Lerner, M. Milik, A. Rey, L. Walters, and P. Wright are gratefully acknowledged. We thank the anonymous referees for the very useful comments, which greatly improved the presentation of our results. This research was supported in part by National Institutes of Health Grant GM-37408 and by the Joseph Drown Foundation.

1. Karplus, M. & Petsko, G. A. (1990) *Nature (London)* **347**, 631–638.
2. Skolnick, J. & Kolinski, A. (1991) *J. Mol. Biol.* **221**, 499–531.
3. Skolnick, J. & Kolinski, A. (1990) *Science* **250**, 1121–1125.
4. Farber, G. K. & Petsko, G. A. (1990) *Trends Biochem. Sci.* **15**, 228–234.
5. Luger, K., Hommel, U., Herold, M., Hofsteenge, J. & Kirschner, K. (1989) *Science* **243**, 206–210.
6. Lasters, I., Wodak, J. S. & Pio, F. (1990) *Proteins* **7**, 249–256.
7. Matthews, C. R. (1990) in *Protein Folding*, eds. Gierasch, L. M. & King, J. (Am. Assoc. for the Advancement of Science, Washington), pp. 191–197.
8. Miles, E. W. (1991) in *Conformations and Forces in Protein Folding*, eds. Nall, B. T. & Dill, K. A. (AAAS, Washington), pp. 115–124.
9. Creighton, T. E. (1990) *Biochem. J.* **270**, 1–16.
10. Tweedy, N. B., Hurle, M. R., Chrynuk, B. A. & Matthews, C. R. (1990) *Biochemistry* **29**, 1539–1545.
11. Miyazawa, S. & Jernigan, R. L. (1985) *Macromolecules* **18**, 534–552.
12. Wright, P. E., Dyson, H. J. & Lerner, R. A. (1988) *Biochemistry* **27**, 7167–7175.
13. Hyde, C. C., Ahmed, S. A., Padlan, E. A., Miles, E. W. & Davies, D. R. (1988) *J. Biol. Chem.* **263**, 17857–17871.
14. Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C. & Wilson, I. A. (1976) *Biochem. Biophys. Res. Commun.* **72**, 146–155.
15. Bethe, M. J., Lattman, E. L. & Rose, G. D. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 4195–4199.
16. Lim, W. A. & Sauer, R. T. (1991) *J. Mol. Biol.* **219**, 359–376.
17. Rey, A. & Skolnick, J. (1992) *J. Comp. Chem.*, in press.
18. Karplus, M. & Weaver, W. (1979) *Biopolymers* **18**, 1421–1437.
19. Richardson, J. S. (1981) *Adv. Prot. Chem.* **34**, 167–339.
20. Wienerga, R. K., Kalk, K. H. & Hol, W. G. J. (1987) *J. Mol. Biol.* **198**, 109–121.
21. Skolnick, J. & Kolinski, A. (1990) *J. Mol. Biol.* **212**, 787–816.
22. Sikorski, A. & Skolnick, J. (1990) *J. Mol. Biol.* **212**, 819–836.
23. Rey, A. & Skolnick, J. (1991) *Chem. Phys.* **158**, 199.