Journal of Statistical Physics, Vol. 30, No. 2, 1983

Equilibrium Folding Pathways for Model Proteins

Sanzo Miyazawa[1] and Robert L. Jernigan[1]

Protein conformations have been generated with both a Monte Carlo scheme and a simpler two-state noninteracting globule-coil model. Conformational energies are taken to consist of intraresidue and interresidue terms. Interresidue energies are taken to be proportional to the number of nativelike contacts. To describe probable folding pathways, either energy or the number of native residues are employed as simple one-dimensional folding–unfolding coordinates. By considering only conformations at each point on these coordinates, it is possible to obtain detailed conformational descriptions of relatively rare intermediates on the folding pathway. This technique of "trapping" intermediates and statistically characterizing them is useful for studying conformational transitions. Equilibrium folding–unfolding pathways have been constructed by connecting most probable conformations in order along the folding coordinate. Calculations with the noninteracting globule-coil model have been performed with details chosen to correspond to those in the Monte Carlo calculation for pancreatic trypsin inhibitor. Both pathways are similar. The α helix appears prior to formation of the central beta sheet; beta sheet formation coincides with a large maximum in the free energy because of the attendant loss of conformational entropy. Subsequently the Monte Carlo method indicates two alternative pathways for growth toward either the amino or the carboxyl terminus, followed by completion of the native form. For the globule-coil model, the growth pattern differs somewhat, with the appearance of the single pathway for folding up to the carboxyl terminus prior to completion of folding. This difference may originate in the Monte Carlo sampling procedures or in the simplifications of the globule-coil model.

KEY WORDS: Proteins; conformations; folding.

1. INTRODUCTION

Are proteins similar to collapsed macromolecules in bad solvents? Such a physical situation could be modeled with an intramolecular binding model

Presented at the Symposium on Random Walks, Gaithersburg, MD, June 1982.

[1] Building 10, Room 4B-56, Laboratory of Mathematical Biology, DCBD, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20205.

0022-4715/83/0200-0549$03.00/0

in which compact conformations are obtained by favoring contacts between atoms, up to the limit of volume overlaps. The worse the solvent, then the greater the preference for intramolecular contacts, eventually including even intrinsically unfavorable interactions. For a polymer of homogeneous composition, this would yield numerous possible arrangements, akin to the arrangements available to the strands within a ball of yarn. However, the physical appearance of the amino acid sequence of a protein is not similar to such a strand of yarn because it is adorned with heterogeneous beads, which are different in size, shape, and color. When a protein is rolled into a globular form, its specific amino acid sequence determines its unique tightly packed native conformation, unlike a ball of homogeneous yarn with little order. The long range interactions observed in the native conformation, together with large conformational entropy of denatured proteins, are directly responsible for the "all-or-none" character of the folding-unfolding transition; a denatured protein can progress to its native form, usually without the appearance of substantial intermediate populations. In this paper we consider equilibrium pathways between the unique native conformation and the denatured state.

The large number of intra-molecular degrees of freedom in a denatured protein is a major obstacle to predicting specific native conformations. There has been progress in treating secondary conformations[1-3] by considering short range interactions alone; however, a substantial number of the residues are predicted to be in the wrong conformation, because of approximations in the methods. Schemes for reducing the total number of degrees of freedom by incorporating these results into subsequent treatments of long range interactions encounter further difficulties. The limited number of conformational states usually considered in secondary methods are insufficient to permit most of the correct long range interactions, and achieving some long range interactions probably requires overriding some intrinsic secondary conformational preferences. These effects indicate the importance of long range interactions.

Here we will consider an intramolecular "binding" model of globular protein conformations in which it is assumed that the intramolecular contacts within the native state are intrinsically favorable. The principal assumption in this model is that the native contact pairs are energetically preferable to any other possible contacts for all residues. Nonnative disulfide pairs have been reported as folding intermediates.[4] It appears that some nonnative conformations are stable; however, it is difficult to say whether or not such nonnative intermediates can be transformed easily into native conformations.

There have been numerous experimental and theoretical attempts to elucidate folding pathways. However, attempts to detect intermediates on

folding pathways and to describe their conformational characteristics are rendered arduous by the usual two-state nature of the folding–unfolding transition. Usually, at all points in the transition, the molecules exist in either the native or the denatured forms; populations of partially folded or unfolded conformations are extremely small. In an attempt to overcome this problem, we have devised calculation methods to trap intermediates and to characterize their conformations.

The principal assumption in developing equilibrium folding pathways is that native conformational nuclei grow stepwise until they merge into larger native conformational domains. Similar assumptions have been employed by others in treating protein folding, for example by Tanaka and Scheraga[5] in their three-step mechanism and by Lesk and Rose[6] in their hierarchic organization of compact units within protein molecules. Here as well as in those models, reliance on knowledge of the native conformation is so strong that it appears to be almost impossible to modify these treatments to include folding steps which pass through nonnative intermediate states.

In some recent studies, Monte Carlo methods have been used to simulate the kinetic processes of protein folding and unfolding[7] and also to study the equilibrium[8] of the folding–unfolding transition by generating large numbers of conformations over the range from the native to the denatured conformation. A detailed Monte Carlo study of equilibrium conformations of pancreatic trypsin inhibitor,[8] together with a growth-merge mechanism for the formation of native conformational regions, yields equilibrium folding–unfolding pathways based upon estimated entropies. However, such Monte Carlo methods are difficult and time consuming, and typically provide results with substantial errors; consequently, simpler methods are desirable.

The one-dimensional lattice gas model proposed by Wako and Saito[9,10] has been applied to the problem, discussed in some detail, and termed the "noninteracting local structure model" by Go *et al.*[7,11] We have applied this method but have chosen to term it the *noninteracting globule-coil* model of protein conformations. Simplifications are possible because of the assumptions that each residue takes only native or random coil states and that interresidue interactions can be neglected except within native conformational regions. The requisite formalism is quite similar to that for helix-coil models, but it does permit accounting for long-range interactions within individual native domains. Results were reported[12] for trypsin inhibitor (PTI), ribonuclease A, lysozyme, and myoglobin, with and without the heme group.[13] Previously Abe and Go[14] showed that this globule-coil model can reproduce details of their Monte Carlo kinetic simulations for two-dimensional lattice proteins. Both studies suggest the

appropriateness of the globule-coil model because of the absence of significant numbers of long-range interactions, except within contiguous native regions. Folding pathways obtained for trypsin inhibitor[12] are similar to those obtained in the more detailed Monte Carlo generations.[8] Below we will compare them in detail.

2. CONSTRUCTION OF EQUILIBRIUM FOLDING–UNFOLDING PATHWAYS

Here, the folding and unfolding process is treated at equilibrium. Thermal fluctuations permit a molecule to change its conformation and pass through an activated state, in an all-or-none transition, between native and denatured states. Because conformational changes are intrinsically statistical, such pathways are the most probable ones. In the present equilibrium sense, it must be noted that these most probable pathways for both folding and unfolding are identical except for direction and must pass through the same intermediates.

Abe and Go[14] employed two folding coordinates, the intramolecular interaction energy and the number of random coil residues. They showed that the most probable folding pathways for two-dimensional lattice proteins were in the direction of decreases in both the conformational energy and the number of random coil residues. A close relationship between these two quantities along the most probable pathway was expected because all components of the energy, both the short-range and long-range interaction energies, were taken to favor the native conformational state; therefore, as the conformational energy decreases, the number of random coil residues also decreases. The use of either conformational energy or the number of random coil residues as a folding coordinate is similarly representative although some details may differ. In our Monte Carlo generations we have taken the folding coordinate to be the conformational energy and for the globule-coil model, the number of native residues.

In constructing equilibrium folding–unfolding pathways, first it is necessary to determine the most probable native residues for the conformations at points along the folding coordinate. Even if the most probable conformations along the folding axis can be clearly identified, how to establish connections on a pathway is not so obvious. If complete energy contours in multidimensional conformational space were available, it would be possible to establish most probable folding pathways by locating the lowest-lying energy barriers. In the present cases, we have no information beyond the most probable conformations at points along the folding axis.

In the simple case where most probable conformations change smoothly and continuously from the denatured state to the native state, a most probable pathway could be constructed by directly connecting, in order, the most probable conformations along the folding axis. It is possible that most probable conformations can change suddenly at some points. There, it is not clear whether or not it is appropriate to construct a pathway by simply connecting neighboring conformers along the folding axis. Since this is an equilibrium calculation, it is conceivable that closely related conformations can appear at nonadjacent points on a folding axis.

3. MONTE CARLO CONFORMATION GENERATION

The protein is represented as a chain consisting of hard sphere C^{α} and C^{β} atoms of radius 1.2 Å. The backbone torsional angles (ϕ, ψ) are the only conformational variables and are permitted to take values at every 10°. This fineness of angle space is required to reproduce most of the close contacts on the protein C^{β}–C^{β} contact maps. Also such fine divisions will permit a more realistic estimation of conformational entropy, which is the main purpose of this calculation. Long chains are difficult to generate because of the well-known excluded volume attrition problem. A special sampling method has been employed to reduce the inefficiency of generating conformations. The total 36×36 points for each residue is divided into independent subsets composed of 9 nonadjacent points on the (ϕ, ψ) grid. For sample generations, we have utilized Boltzmann factors of the intraresidue energies with varying bias toward the native conformation. Sampling is in two stages. A small group of conformations is randomly chosen; then a second sampling is performed from among those with permissible excluded volume.

The total energy of a conformation is given by

$$E(\psi_1, \phi_2, \ldots, \phi_N) = \sum_k \left(E_k^{\text{emp}} + E_k^{\text{short}}\right) + \sum_{j>i+1} \sum E_{ij}^{\text{cont}} \qquad (1)$$

where E_k^{emp} is the empirical conformational free energy for this specific conformation of residue k calculated from a statistical compilation of reported (ϕ, ψ) values,[8,12] E_k^{short} is a short-range energy which is favorable if the residue's ϕ and ψ are both within 10° of their native values. Here it is taken as -1 kcal mol^{-1}. The double sum is over all residue pairs with C^{β}–C^{β} distance less than or equal the cutoff distance of 6.5 Å in the crystal structure. If there is no atomic overlap, and the distance is less than

or equal to 6.5 Å, then each such contact is accorded a favorable contact energy of -2 kcal mol^{-1}. A partition function $Z(E)$ for a limited range of energy values is calculated; it is then used to obtain probabilities of each native contact pair being formed. Further details of this method are given in Ref. 8.

4. NONINTERACTING GLOBULE-COIL MODEL

In this model each residue can take only one of two conformational states, the native and the random coil states. The native state of each residue is taken to be the reported crystal conformation. The energy is given by an expression similar to Eq. (1), with only a few differences. The native state energies include the same empirical, short-range and long-range components. The calculation has been performed with a contact map with 94 non-nearest-neighbor contacts corresponding to the lowest-energy conformer obtained in the Monte Carlo calculation. The statistical weight of the random coil state is taken to be the sum of contributions from all states on the (ϕ,ψ) map, except the native point. An additional parameter α has been introduced to represent other contributions to random coil energies; it has been taken to be identical for all residues. Since α is assumed to be independent of residue, averages calculated for fixed n are independent of α. This treatment is the same as in Ref. 12 except that the short-range native energy, namely, -1 kcal mol^{-1}, has now been included to conform to the Monte Carlo energies. Interresidue interactions are included only within each native globule; such interactions within random coil regions or between random coil and native conformational regions are completely neglected.

For the noninteracting globule-coil model, a total partition function is formulated as a sum of partition functions $Z(n)$ which include the statistical weights of all conformations with n native residues. This formulation permits a direct calculation of the most probable native residues at each point along a folding coordinate n. The requisite formalism is given in Ref. 12 in the form of a set of recurrence equations.

5. RESULTS

The atomic coordinates of bovine pancreatic trypsin inhibitor (PTI) were taken from the Brookhaven protein data bank.[15] We have previously[12] examined with the noninteracting globule-coil model the effects of

varying the contact energy and the cutoff distance on the free energies and conformational probabilities. The conclusion was that the folding pathways are not strongly affected by changes in these parameters. Here, we present detailed results at the melting condition, $RT = 0.67$ kcal mol^{-1} for pancreatic trypsin inhibitor.

5.1. Characteristics of the Folding–Unfolding Transition

For both models an "all-or-none" character to the folding–unfolding transition is indicated by the appearance of large free energy barriers which almost completely separate folded states from unfolded states. The maximum free energy barrier is substantially higher in the Monte Carlo calculation. Any comparison of free energies for the two models requires a mapping of one folding coordinate onto the other.

5.2. Most Probable Conformations

Results from the Monte Carlo calculations are given in Fig. 1 in the form of probable long-range contact maps at different points along the folding coordinate. For the globule-coil model, the most probable native residues at each point along the folding coordinate n are examined in terms of the probabilities of each residue being native. A representation of these probabilities and their dependence on residue position i are shown in Fig. 2. Dots represent residues with a larger than average probability of being native at each stage of folding. At small numbers of native residues, intraresidue energies and relatively short-range interresidue energies determine most probable conformations. The alpha helix near the carboxyl terminus, the turn between the two β-strands at 25–26, and the 3_{10} helix near the amino end appear at small numbers of native residues. The long-range beta sheet interactions appear suddenly as manifested by the dots at $i = 18$–36 that appear at $n = 28$. Similar results can be observed in Fig. 1 for the Monte Carlo calculations. For both calculations, the β sheet formation corresponds to the conformation at the free energy maximum.

Generally, it has been found that probable conformations at free energy minima in the range of small numbers of native residues correspond to turns and helices. In contrast, beta sheets require significantly longer-range interactions for stability; consequently, beta sheets usually appear in a later more cooperative step.

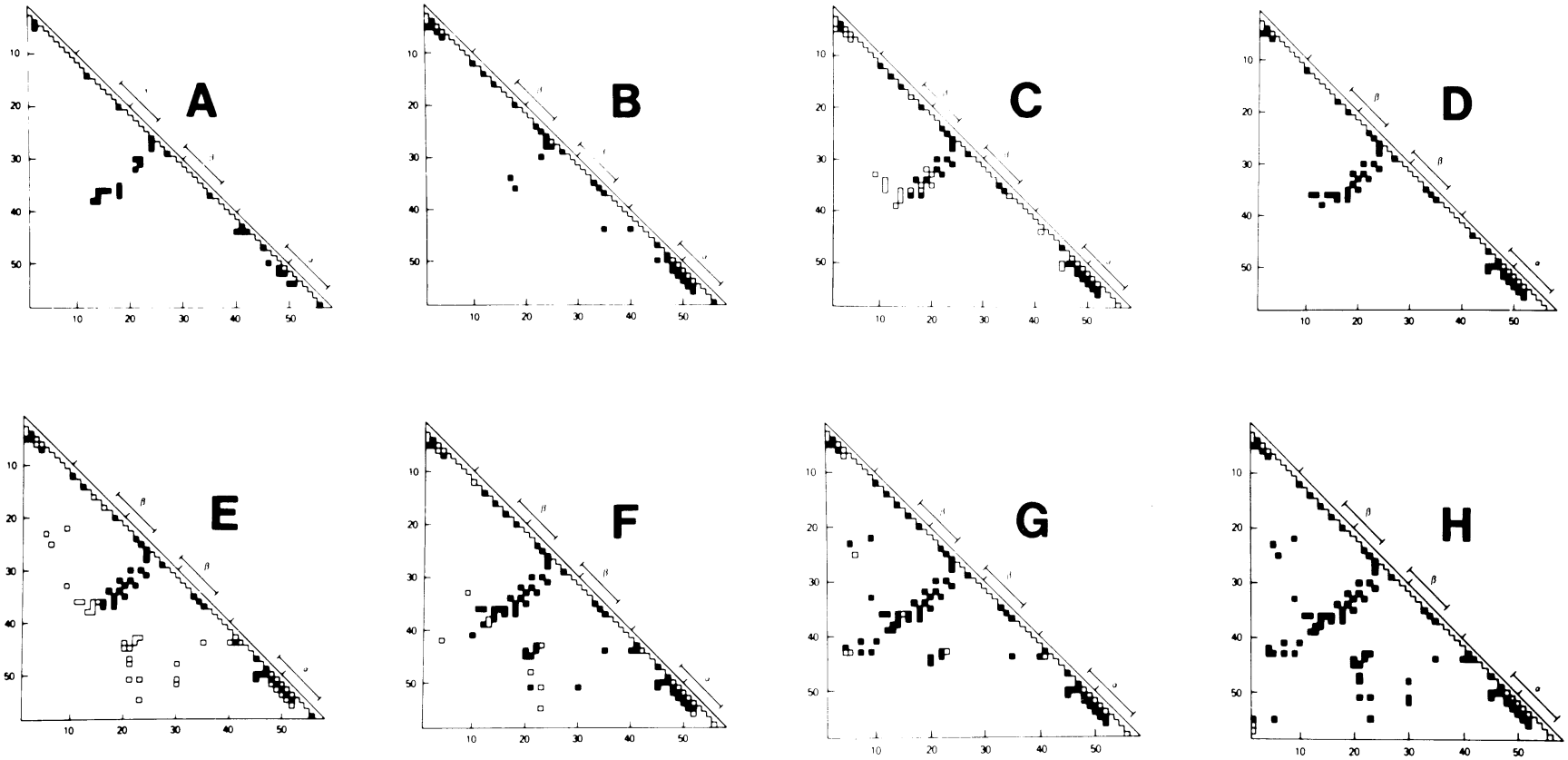

Fig. 1. Contact maps at different stages of folding for PTI from the Monte Carlo calculations. A to H correspond to points on the folding coordinate in order of decreasing energy. A is the most denatured state, and H is the most native state. Numbers represent residue indices. A black square means that the probability for formation of a contact pair is greater than or equal to 3/4; an open square corresponds to a probability less than 3/4 but greater than or equal to 1/4.

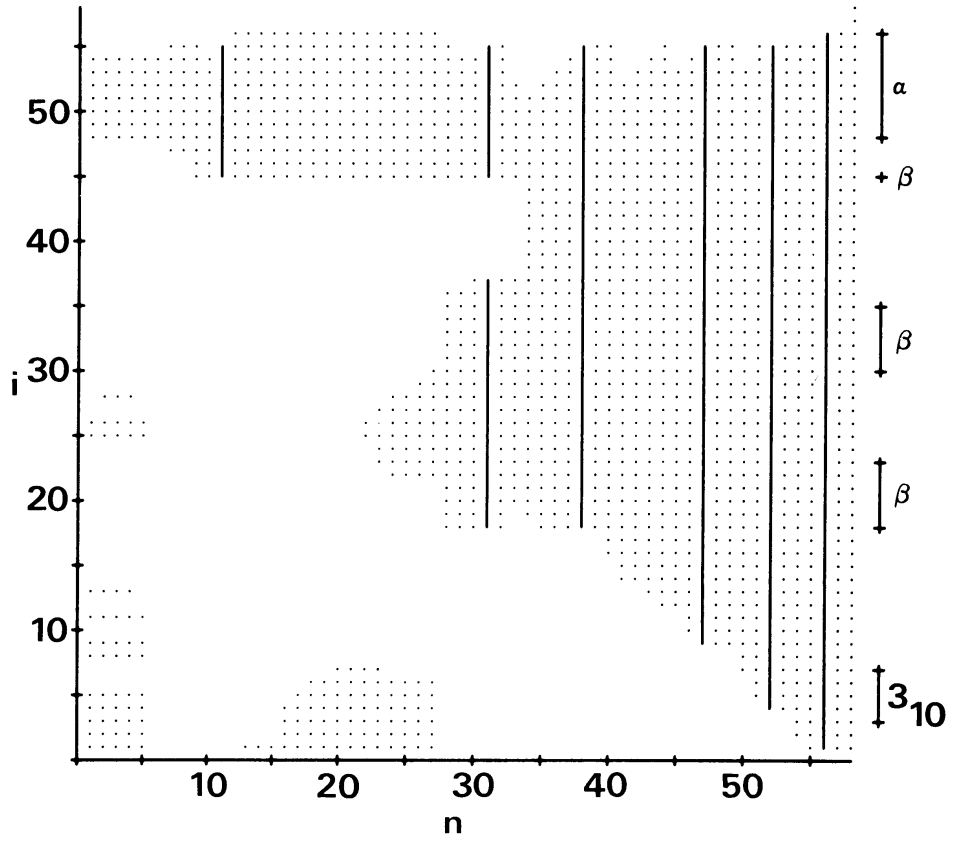

Fig. 2. Locations of the most probable native residues for bovine pancreatic trypsin inhibitor for all possible numbers of native residues, n, with the globule-coil model. The presence of a dot at position (n, i) indicates that the probability of residue i being native is greater than or equal to n/N, where N is the total number of residues. Free energy minima along n are presented as solid lines instead of dots. Secondary regions are shown as bars on the right side of the figure. Results differ somewhat from those in Ref. 12 because different energies and contact maps have been used to correspond more closely to those used in the Monte Carlo calculations. Strong dependences on the details of the native contact map were reported in Ref. 13.

5.3. Folding–Unfolding Pathways

Most probable native conformational fragments at representative points along a folding–unfolding coordinate are shown for both models in Fig. 3 in order of increasing numbers of native residues. For the Monte Carlo calculations, the native fragments were determined by selecting native globules on the basis of the contact maps in Fig. 1. The most probable native residues shown for the globule-coil model usually correspond to conformations at local free energy minima. In the present globule-coil case, folding–unfolding pathways can be constructed simply by con-

necting most probable conformations in order of increasing numbers of native residues. In the Monte Carlo results, there is an alternative folding pathway, shown as dotted lines in Fig. 3. Otherwise, each intermediate state and the order of its appearance on the folding pathways are nearly identical for the two models.

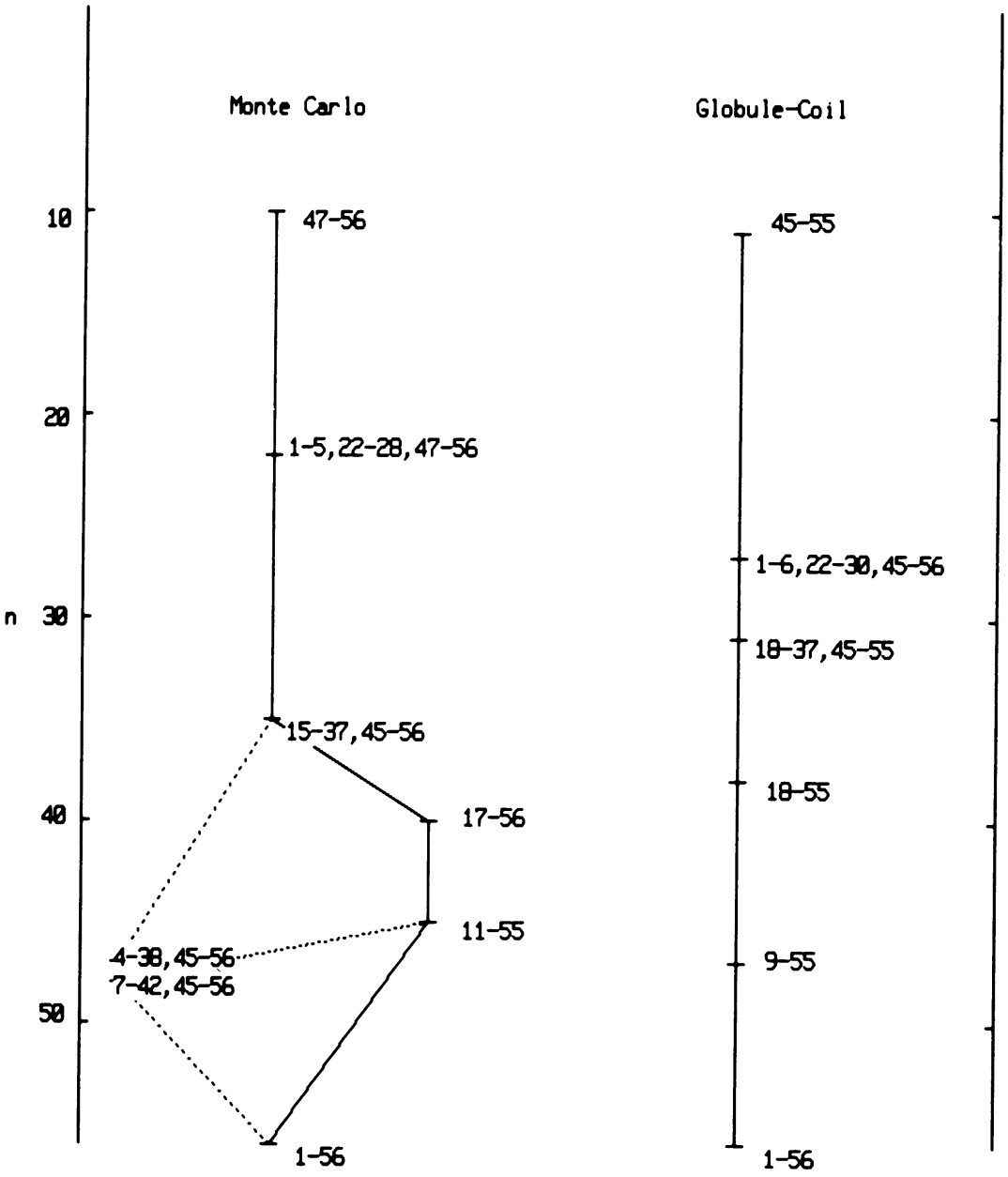

Fig. 3. Probable folding pathways at equilibrium for PTI along the folding coordinate n, the number of native residues. Numbers indicate native residues at different stages of folding, on the left for the Monte Carlo results and on the right for the globule-coil model. Connecting lines indicate folding–unfolding equilibrium pathways.

6. DISCUSSION

The folding intermediates found with the globule-coil model for PTI are almost all consistent with those obtained in the Monte Carlo generation: (1) The α helix forms early and persists. (2) A beta sheet forms near the point of highest free energy. (3) The most probable folding pathway then proceeds with native formation completed toward the *C*-terminal. The other significant pathway observed in the Monte Carlo samples, which corresponds to folding toward the *N*-terminus subsequent to formation of the beta sheet, but prior to completion of the carboxyl half of the molecule, has not been detected with the noninteracting globule-coil model. The reason for the absence of this pathway may reside in errors in the entropy evaluation with the Monte Carlo sampling or in the substantial excluded volume effect which has been accounted for only with the Monte Carlo method.

The present study represents an attempt to fit the results of a computer experiment with a simpler model. The extent of agreement between the folding pathways derived by the two methods as presented in Fig. 3 is gratifying, especially since the computer time required for the Monte Carlo calculation is larger by a factor of 10^3 to 10^4.

REFERENCES

1. T. T. Wu, S. C. Szu, R. L. Jernigan, H. Bilofsky, and E. A. Kabat, *Biopolymers* **17**:555 (1978).
2. S. Bourgeois, R. L. Jernigan, S. C. Szu, E. A. Kabat, and T. T. Wu, *Biopolymers* **18**:2625 (1979).
3. R. L. Jernigan, S. Miyazawa, and S. C. Szu, *Macromolecules* **13**:518 (1980).
4. T. E. Creighton, *Progr. Biophys. Mol. Biol.* **33**:231 (1978).
5. S. Tanaka and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **72**:3802 (1975).
6. A. M. Lesk and G. D. Rose, *Proc. Natl. Acad. Sci. USA* **78**:4304 (1981).
7. N. Go, H. Abe, H. Mizuno, and H. Taketomi, in *Protein Folding*, R. Jaenicke, ed. (Elsevier/North Holland Biomedical Press, Amsterdam, 1980), pp. 167–181.
8. S. Miyazawa and R. L. Jernigan, *Biopolymers* **21**:1333 (1982).
9. H. Wako and N. Saito, *J. Phys. Soc. Jpn.* **44**:1931 (1978).
10. H. Wako and N. Saito, *J. Phys. Soc. Jpn.* **44**:1939 (1978).
11. N. Go and H. Abe, *Biopolymers* **20**:991 (1981).
12. S. Miyazawa and R. L. Jernigan, *Biochemistry*, **21**:5203 (1982).
13. R. L. Jernigan and S. Miyazawa, *Biopolymers*, **22**, to appear (1983).
14. H. Abe and N. Go, *Biopolymers*, **20**:1013 (1981).
15. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.* **112**:535 (1977). The Brookhaven Protein Data Bank. Coordinates used were those of 3PTI.