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

Protein Folding as a Stochastic Process

Nobuhiro Gō[1]

In physiological conditions globular protein molecules assume a specific native conformation uniquely determined by its amino acid sequence. Upon environmental changes the protein molecules undergo reversible unfolding (order losing) and folding (order gaining) transitions, which is similar to the first-order phase transition. Pathways of folding have been intensively studied in the hope of deciphering the code that amino acid sequences carry as to the three-dimensional structure of proteins. A strongly simplified *lattice model of proteins* has been found to be a powerful theoretical tool to simulate the dynamic process of the folding and unfolding transitions. The results of the simulation indicate the existence of stochastic pathways of folding.

KEY WORDS: Protein folding; lattice model; computer simulation; Markov process.

1. INTRODUCTION

Proteins are copolymers of 20 amino acids with genetically determined definite sequences. Many proteins assume specific native more-or-less globular conformations in the physiological state. It is generally believed and has been demonstrated experimentally for some proteins that the specific conformation of a protein molecule is realized as a thermal equilibrium state.[1] This denies involvement of any previous history of the molecule in the determination of the native conformation. The specific native conformation of the protein molecule, even though very complex in general, is uniquely determined by its amino acid sequence and its environment.

When one or more environmental parameters are shifted away from the physiological values, globular proteins generally undergo an unfolding

Presented at the Symposium on Random Walks, Gaithersburg, MD, June 1982.
Work supported by grants in aid from the Ministry of Education, Japan.
[1] Department of Physics, Faculty of Science, Kyushu University, Fukuoka, 812 Japan.

0022-4715/83/0200-0413$03.00/0 © 1983 Plenum Publishing Corporation

transition to assume a disordered unfolded state. This transition is generally reversible upon regeneration of the enviroment, which is the manifestation of the equilibrium nature of the native conformation. From the biological point of view the process of folding is a prototype of morphogenetic phenomena and in this process the information about the three-dimensional structure coded in the amino acid sequence is decoded. Elucidation of this code is a fundamental biological problem. Studies of the detailed process of folding and unfolding transition are expected to lead to the elucidation of the code.

From the physical point of view, a protein molecule is an information-carrying finite system. The folding and unfolding transition is a phenomenon similar to the first-order phase transition. In this paper I will discuss basic statistical-physical problems of the folding and unfolding transition in globular proteins.

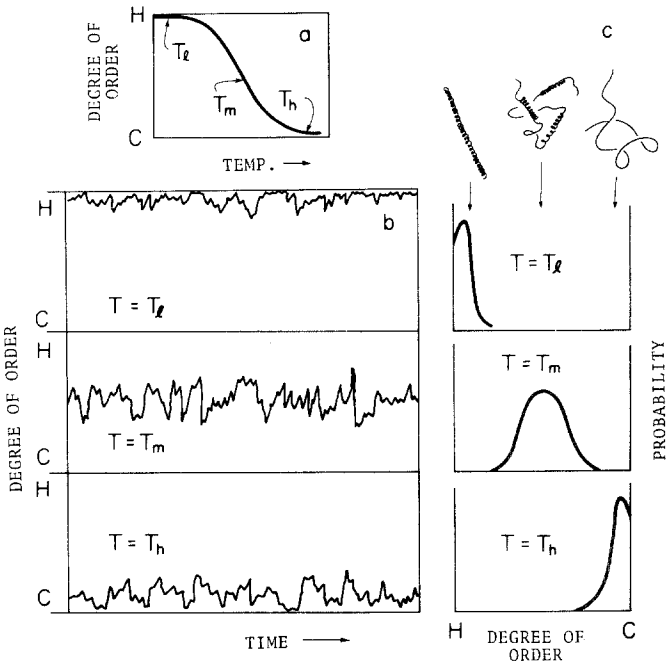

Fig. 1. Schematic illustration of the behavior of homopolypeptide chains in the helix-coil transition. States *H* and *C* are 100% and 0% helical states, respectively. (a) The mean value of a degree of order over many molecules is plotted against temperature. (b) Attention is focused on one molecule. The degree of order of this molecule is plotted against time for three temperatures. (c) Probability of the molecule existing in states with various degrees of order is plotted for the three temperatures.

2. BASIC CHARACTER OF THE TRANSITION

Figures 1 and 2 illustrate the basic statistical-physical character of the folding and unfolding transition in globular proteins as compared with the helix-coil transitions in synthetic polypeptides. The helix-to-coil transition proceeds as the fraction of disordered parts in each molecule increases. In contrast to this the unfolding transition in globular proteins proceeds as the number of molecules existing in the folded state decreases with an accompanying increase of the number of molecules existing in the unfolded state. In this sense the unfolding transition in globular proteins is similar to the first-order phase transition.

The characteristic of the helix-coil transition illustrated in Fig. 1 is a direct consequence of the one-dimensionality of the phenomena. According to the famous theorem due to Landau,[2] macroscopic phases cannot coexist in a one-dimensional system. Thus, at the midpoint of the transition ordered (helical) and disordered (random coil) sections of microscopic lengths alternate within each polypeptide chain. The one-dimensionality of

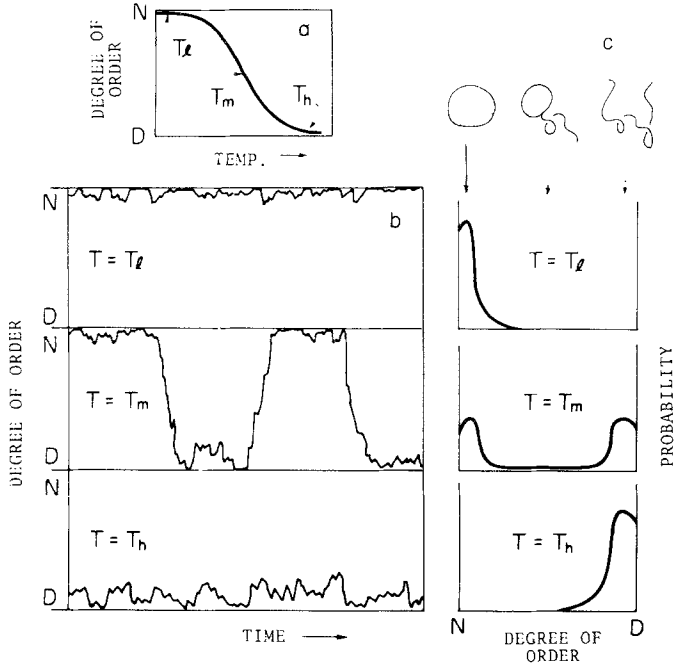

Fig. 2. Schematic illustration of the behavior of protein molecules in the folding and unfolding transition. States N and D are the folded native and unfolded denatured states, respectively. The three parts (a)–(c) show the same aspects as in Fig. 1.

the helix-coil transition is a consequence of the fact that important intramolecular interactions responsible for the phenomenon are those between parts close along the chain. These interactions are called short-range interactions.

The characteristic of the folding and unfolding transition in globular proteins as illustrated in Fig. 2 indicates that intramolecular interactions between parts far along the chain (but close in space) are important. These interactions are called long-range interactions. Thus, *the long-range interactions are important* in the folding and unfolding transitions in globular proteins.

At the same time *importance of the short-range interactions* is also amply evidenced. Most convincing is the fact that local secondary structures within globular proteins such as α-helices, β-strands and turns can be predicted fairly well from the amino acid sequence by various algorithms in which essentially only the short-range interactions are considered.[3]

3. LATTICE MODEL OF PROTEIN

For understanding the folding and unfolding transition in globular proteins, both the long- and short-range interactions must be considered at the same time. It is also essential that the molecule has no simple repeating symmetry. For this very reason a protein molecule carries information and can have complex but specific native conformations.

The extent to which we can understand such systems by analytic treatments is limited. Instead, the method of computer simulation could be a powerful theoretical method. However, the simulation can be carried out only when the model is a strongly simplified one. It takes an impossibly large amount of computer time to carry out simulation for any realistic models of atomic resolution. The conformational changes in one step of computer calculation in realistic models may correspond to those in real proteins taking place in a time interval of the order of 10^{-13} sec (characteristic time for molecular vibrations). Proteins fold roughly in the order of 10^{0} sec. In order to simulate this, 10^{13} steps of calculation are necessary, requiring 10^{14} sec of computer time (10 sec is assumed for one step). If 10^{4} sec is a reasonable upper limit of the computer time for the problem of protein folding, we need a simplification of the model that can cut the computer time by a factor of 10^{10}. This is a drastic simplification. We must attain this simplification without losing the essence of the phenomena. A lattice model of protein folding has been proposed and studied[4-14] as a model which satisfies this criterion.

A lattice protein molecule is a self-avoiding chain polymer on a two-dimensional square lattice[4,5,7-9,11-14] or on a three-dimensional cubic lattice.[6,10] In the present paper I will discuss only results on the two-

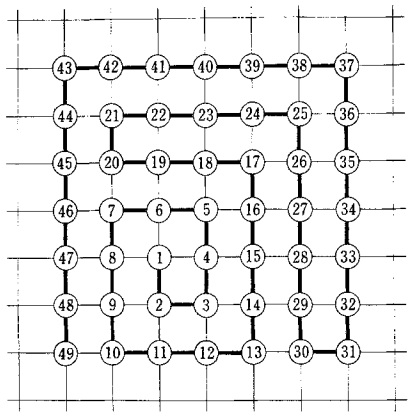

Fig. 3. The native conformation of protein SP in the two-dimensional square lattice.

dimensional model. As a result of specific intramolecular interactions described below, the protein molecule assumes a specific native conformation at low temperatures.

The native conformation of protein SP, which we study in this paper, is shown in Fig. 3. Black squares in Fig. 4 show the nearest-neighbor units

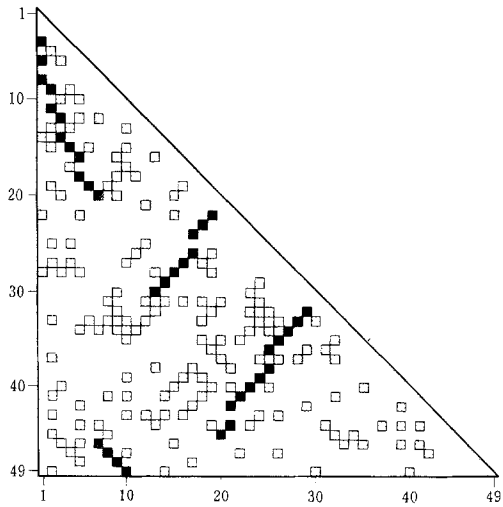

Fig. 4. Specification of the attractive long-range interactions in protein SP in terms of interactable pairs shown by black and shaded squares. Both abscissa and ordinate indicate residue number. Black pairs are those occupying nearest-neighbor lattice points in the native conformation of Fig. 3. When a pair of interactable units occupies nearest-neighbor lattice points, the energy of the system is assumed to decrease by ϵ.

in this native state. In addition to them, 148 randomly selected pairs are shown by shaded squares. We call those pairs shown by black and shaded squares *interactable* units. When a pair of interactable units occupy nearest-neighbor lattice points in an arbitrary conformation, we assume that the energy of the system decreases by ϵ. These interactions are a model of the long-range interactions. When a smaller (larger) number of randomly selected pairs is employed, the specificity of the interactions increases (decreases).

As a model of the short-range interactions, we consider "bond energies," each of which is a function of a "bond angle." When the bond angle at the ith unit ($i = 2, 3, \ldots, 48$) takes the same value as in the native conformation, the bond energy of the ith unit is lower by ϵ' than the other two possible cases. Relative weights of the long- and short-range interactions can be changed by changing the relative values of ϵ and ϵ'.

Simulation is carried out by the Monte Carlo method of Metropolis *et al.*[15] By this method a good sample of the equilibrium population of various conformations at a given temperature T is effectively produced. In this paper we express temperature as a dimensionless quantity T^* defined by kT/ϵ_0, where k is the Boltzmann constant and ϵ_0 is the unit of energy defined by $\epsilon_0 = \epsilon + \epsilon'$. It has been shown[4] that the trial number of the Monte Carlo simulation is approximately proportional to physical time. Therefore, we will analyze records of simulation as dynamical records.

4. RESULTS AND DISCUSSION

Records of long computer simulations carried out at the melting temperatures T_m^* for four different relative weights of the long- and short-range interactions $(\epsilon, \epsilon') = (0.75\epsilon_0, 0.25\epsilon_0)$, $(0.5\epsilon_0, 0.5\epsilon_0)$, $(0.25\epsilon_0, 0.75\epsilon_0)$, and $(0, \epsilon_0)$ are shown in Fig. 5.[8]

Very long runs of 8.0×10^5 trials for the case of $(0.75\epsilon_0, 0.25\epsilon_0)$ are shown in Figs. 5a and 5b, in each of which we observe the unfolding or folding transition only once. This means that the lattice polymer spends most of its time in either the folded or unfolded states. The probability of its being in the intermediate states is very low. This is the behavior expected for systems undergoing a first-order-type transition.

The case of $(0.5\epsilon_0, 0.5\epsilon_0)$ is shown in Fig. 5c. The folding and unfolding transitions are observed to occur more frequently than in the case of $(0.75\epsilon_0, 0.25\epsilon_0)$. Yet the polymer still spends most of its time in either the folded or unfolded states.

In the case of $(0.25\epsilon_0, 0.75\epsilon_0)$ (Fig. 5d) the polymer goes back and forth between the folded and unfolded states frequently. The probability of being in the intermediate states is appreciable.

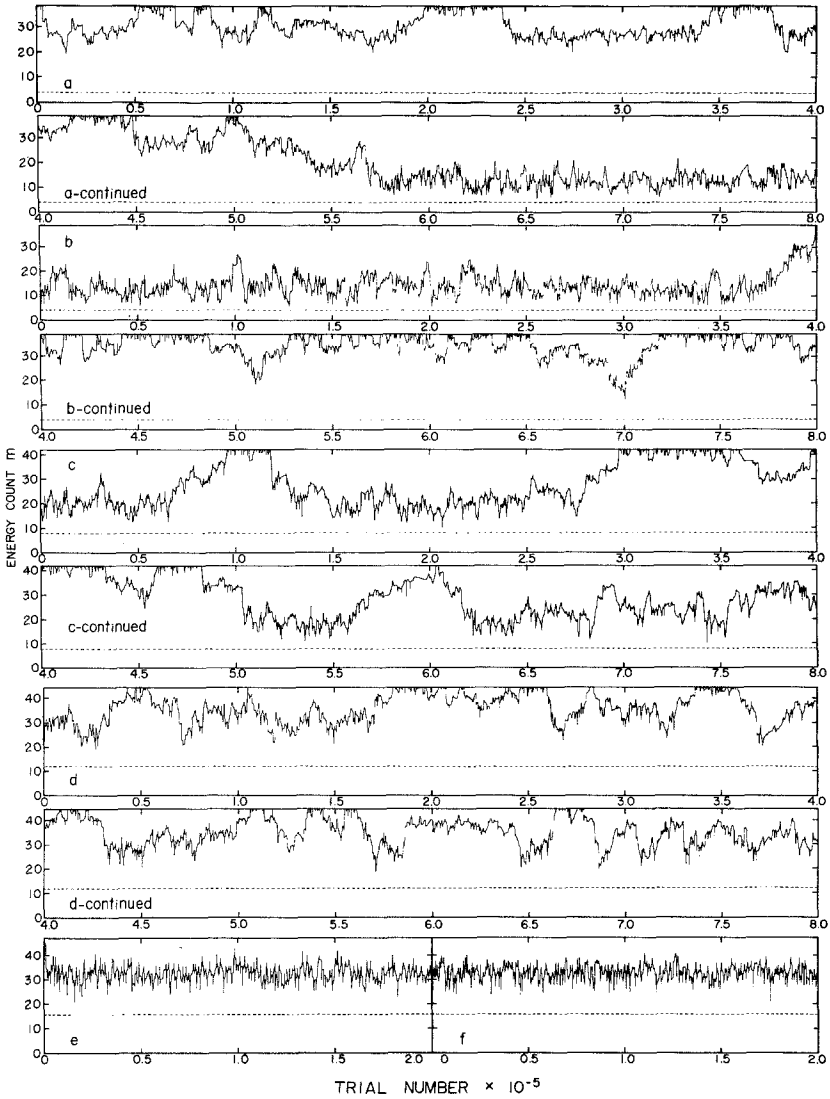

Fig. 5. Records of conformational changes at melting temperatures T_m^* for four different relative weights of (ϵ, ϵ'). (a) and (b): $(0.75\epsilon_0, 0.25\epsilon_0)$, $T_m^* = 0.675$; (c): $(0.5\epsilon_0, 0.5\epsilon_0)$, $T_m^* = 0.6$; (d): $(0.25\epsilon_0, 0.75\epsilon_0)$, $T_m^* = 0.525$; (e) and (f): $(0, \epsilon_0)$, $T_m^* = 0.7$. Ordinate is the conformational energy counted in units of $(-\epsilon_0)$.

The transition in the case of $(0, \epsilon_0)$ (Fig. 5e and 5f) is of the graded type. Conformational fluctuations within a limited range of energy count are observed around the mean value.

In order to quantify the kinetic behavior of the system, normalized time correlation functions defined by

$$\phi_P(t) = \left[\langle P(0)P(t) \rangle - \langle P \rangle^2 \right] / \left(\langle P^2 \rangle - \langle P \rangle^2 \right) \tag{1}$$

are calculated from the record of computer simulation. Here $P(t)$ is the value of a certain physical quantity, P, at time t. As the physical quantity P, we take the long- and short-range interaction energies and the radius of gyration. The computed normalized time correlation functions are shown in Figs. 6a and 6b.[12] They can be expressed as a sum of two simple exponential terms. The slower relaxation time does not depend on the nature of the physical quantities for which the correlation functions are

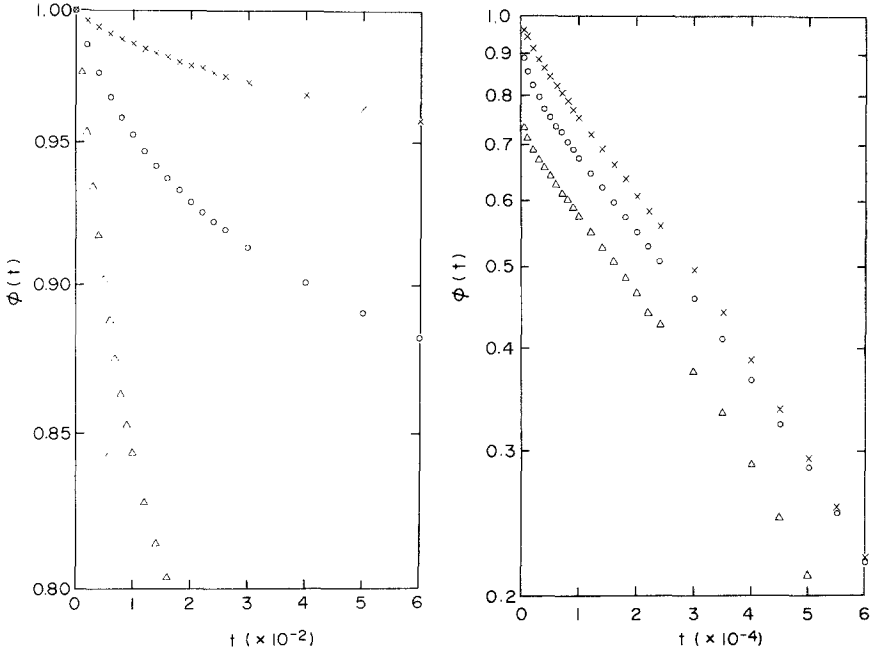

Fig. 6. Normalized time correlation functions for three time-dependent quantities computed from the records of simulation. (Records for another protein AB with a different native conformation[12] are used) $(\epsilon, \epsilon') = (0.75\epsilon_0, 0.25\epsilon_0)$. ×, for the long-range interaction energy; ○, for the short-range interaction energy; △, the radius of gyration. (a) short-time range. (b) long-time range.

obtained. This indicates that the slower relaxation time corresponds to the overall folding and unfolding transition. The faster relaxation time reflects the conformational fluctuations within each of the folded and unfolded states.

From a series of computer simulations for various choices of parameters, the following conclusions have been reached:

(1) Specificity of the long-range interactions as well as the non-one-dimensionality is important for the first-order-type character of folding and unfolding transition.

(2) The nonspecific component of the long-range interactions (a) reduces the cooperativity of the transition and (b) decelerates the kinetic rate of transition.

(3) The short-range interactions (a) reduce the cooperativity of the transition, but (b) accelerate the kinetic rate of transition.

In order to approach the ultimate goal of elucidating the information about the three-dimensional structure coded in the amino acid sequence, we must clarify more details of the process of folding. By inspection of Fig. 5 we see that the series of events in each case of a folding and unfolding transition as monitored by one conformational parameter (here, the conformational energy) are different. This situation does not change, even when monitored by two conformational parameters. Figures 7a and 7b are examples of trajectories of two cases of folding transitions in two-dimensional conformational space.[13] The state point leaves random-walk-like trajectories. It is clear that the folding and unfolding transition is stochastic.

A simplest case is one in which we can define a series of intermediate states between which a protein molecule undergoes transitions as a Markov process. If this is possible, each intermediate state contains the necessary and sufficient information to determine its future destiny in the stochastic sense.

Can the record of conformational changes as monitored by one conformational parameter such as shown in Fig. 5 be regarded as a Markov process? In order to see this, we calculated from the record the conditional probability of transition in one step from one state to another, both characterized by certain values of the conformational parameter. From the conditional probability, we obtained the normalized time correlation function by assuming a Markov process. This normalized time correlation function was found to decay much faster than that obtained directly from the record of simulation.[14] This indicates that there are essential variables hidden behind the record of simulation. For elucidation of the kinetic process of folding, we must define a good set of intermediate states by specifying the hidden variables.

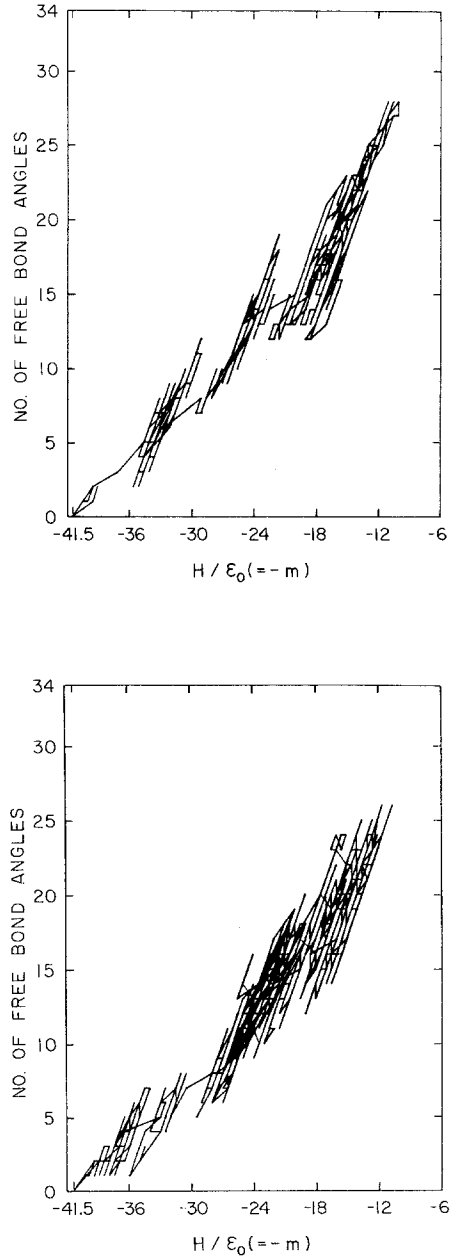

Fig. 7. Trajectories of the state point in two-parameter space in two events of folding in protein SP. Abscissa is the total energy count and ordinate is the number of free bond angles (those that assume non-native values).

ACKNOWLEDGMENTS

Computations were carried out at the Computer Center, Kyushu University and the Computer Center, Institute for Molecular Science.

REFERENCES

1. C. B. Anfinsen and H. A. Scheraga, *Adv. Protein Chem.* **29**:205 (1975).
2. L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Pergamon Press, London, 1958), p. 478.
3. G. Nemethy and H. A. Scheraga, *Quart. Rev. Biophys.* **10**:239 (1977).
4. H. Taketomi, Y. Ueda, and N. Gō, *Int. J. Pept. Prot. Res.* **7**:445 (1975).
5. N. Gō, *Adv. Biophys.* **9**:65 (1976).
6. Y. Ueda, H. Taketomi, and N. Gō, *Biopolymers* **17**:1531 (1978).
7. N. Gō and H. Taketomi, *Proc. Natl. Acad. Sci. USA* **75**:559 (1978).
8. N. Gō and H. Taketomi, *Int. J. Pept. Prot. Res.* **13**:235 (1979).
9. N. Gō and H. Taketomi, *Int. J. Pept. Prot. Res.* **13**:447 (1979).
10. N. Gō, H. Abe, H. Mizuno, and H. Taketomi, *Protein Folding*, R. Jaenicke, ed. (Elsevier, New York, 1980), p. 167.
11. H. Abe and N. Gō, *Biopolymers* **20**:1013 (1981).
12. F. Kanô and N. Gō, *Biopolymers* **21**:565 (1982).
13. N. Gō and H. Abe, submitted.
14. N. Gō and F. Kanô, submitted.
15. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.* **21**:1087 (1953).