UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

H4.SMR/775-5

COLLEGE IN BIOPHYSICS: EXPERIMENTAL AND THEORETICAL ASPECTS OF BIOMOLECULES

26 September - 14 October 1994

Miramare - Trieste, Italy

Protein Folding - An Introduction

Nicholas D. Socci
University of California at San Diego
La Jolla - CA, USA

Properties and origins of protein secondary structure

Nicholas D. Socci,[1,2,*] William S. Bialek,[2] and José Nelson Onuchic[1]
[1]*Department of Physics, University of California at San Diego, La Jolla, California 92093*
[2]*NEC Research Institute, Princeton, New Jersey 08540*
(Received 12 July 1993)

Proteins contain a large fraction of regular, repeating conformations, called secondary structure. A simple, generic definition of secondary structure is presented which consists of measuring local correlations along the protein chain. Using this definition and a simple model for proteins, the forces driving the formation of secondary structure are explored. The relative role of energy and entropy are examined. Recent work has indicated that compaction is sufficient to create secondary structure. We test this hypothesis, using simple nonlattice protein models.

PACS number(s): 87.15.By

Recently, there has been a great deal of interest in the study of proteins from a physical perspective [1–6]. Most of these works have focused on the folding problem; i.e., how does the sequence of amino acids encode the three-dimensional structure of the protein? Although progress has been made in this area, there is still a long way to go before there is a complete understanding of how proteins fold. However, proteins have many other interesting properties. While each protein has a specific structure determined by its sequence, all proteins share several common structural features. They are highly compact, with very little free internal space. More striking is the high degree of order found, which consists of regular periodic arrangements of the main chain into one of a few universal patterns (called *secondary structure*). Roughly 50% of the structure of all proteins is in some form of secondary structure [7]. In this paper we define in a simple, generic way precisely what secondary structure is. This definition will be valid not only for proteins but for simpler polymers and simple proteinlike models. We then use it to investigate what forces are responsible for the formation of secondary structure. Although this is not directly related to the folding problem, a thorough understanding of what factors are responsible for secondary structure may aid in the study of the folding problem.

There has been a great deal of past work attempting to understand the origins of secondary structure. At first it was believed that *local* interactions (local hydrogen bonds or dihedral angle potentials, for example) were responsible. Here, the term local means close with respect to the separation along the polymer chain. For example, a hydrogen bond between monomer i and $i+4$ would be a local interaction, as would an angle potential. Several recent studies indicate that local forces may not be the dominant effect, rather compaction of the chain may be the important factor. By examining exhaustive enumerations of short chains on a lattice, Chan and Dill [8–10] found that as the compactness of the chains increased so did the percentage of secondary structure present. They also found that the maximally compact chains had roughly the same amount of secondary structure as real proteins and the proportions of helices to sheets was also approximately the same. Subsequently, Gregoret and Cohen [11] studied nonlattice models. Their results also suggest that compactness does influence the amount of secondary structure, but they indicate that the effect is most pronounced at densities 30% greater than that of real proteins. In both of these studies, however, local interactions were present. For example, a lattice has a specific set of allowed bond angles, which provides an effective bond angle potential. In the nonlattice work, compact chains were generated using a biased random walk in which the bond angles were chosen not from a uniform distribution but from the distribution observed in real proteins. This also provides an effective angle potential. Therefore, it is not clear from these works whether compaction is sufficient to generate secondary structure. We wish to determine whether compaction, without local interactions, is sufficient.

There are two distinct questions to keep in mind: (1) why do proteins (or other polymers) form regular structures and (2) why do proteins form particular types of secondary structure? Question (1) is equivalent to asking the following: why do proteins form helices and sheets? The second question asks the following: why are these helices α helices and the sheets β sheets? The answer to the second question certainly involves local interactions. It is the specific hydrogen bonding patterns in proteins which favor the formation of α helices. In other polymers, different local interactions would favor other forms. For example, the structures of 179 polymers have been solved and 79 are found to be in one of 22 different types of helices [9,12]. In each polymer the specific types of local interactions determine the preferred type of secondary structure. In this work we are interested in studying the first question: what forces are responsible for formation of regular structures. Specifically we will test the previous suggestions that compaction of the chain is the key driving force. To do so we will be using models without any local interactions. However, without local interactions there is no way of knowing beforehand what types of secondary structure will be formed. Most definitions of secondary structure are specific to a given type of structure (i.e., α helices); consequently one needs

*Present address: University of California, San Diego, La Jolla, CA 92093.
Electronic address: nsocci@ucsd.edu

to know *a priori* what types of secondary structures will occur in order to detect their presence. To overcome this problem we developed a generic method of determining whether secondary structure is present without the need to know *a priori* what its specific form is.

A simple way of defining secondary structure is to realize that it consists of repeating patterns. Consequently the polymer chain should be correlated with itself along the chain. The correlation length should be related to the average size of secondary structures. To detect secondary structure we measure the correlations between different points along the protein chain. Specifically, let θ_j represent the value of the dihedral angle associated with the jth α carbon (see Fig. 1). We then calculate

$$C_\theta(\Delta) = \left\langle e^{i(\theta_j - \theta_{j+\Delta})} \right\rangle_C . \quad (1)$$

The average is over j; that is, over all pairs of angles separated by a distance Δ along the chain. The subscript C indicates that the mean, $\left|\left\langle e^{i\theta_j}\right\rangle\right|$, has been subtracted from $\left\langle e^{i(\theta_j - \theta_{j+\Delta})}\right\rangle$. If secondary structure is present then $C_\theta(\Delta)$ will be nonzero for $\Delta \lesssim l_{\text{avg}}$ where l_{avg} is related to the average length of secondary structure. Note, this definition makes no reference to any particular type of secondary structure; therefore, any form of regular structure will be detected. For example, if helices are present there will be a nonzero correlation length no matter what period the helices have. Equation (1) also has the advantage that it can be calculated analytically in a simple model.

To test our definition we examined the crystal structures from 112 proteins which have been recorded in the protein data bank [13]. The correlation function was calculated for each protein and normalized so $C_\theta(0) = 1$. Then an average correlation function was computed for all proteins. Examining this correlation function (shown in Fig. 2) we see that protein chains are positively correlated up to separations of approximately nine monomers. This is comparable to the average length of secondary structure (roughly ten monomers) measured by others [7]. At distances greater than nine monomers the chains become negatively correlated. This negative correlation may be partly due to *supersecondary* structure, which consists of combinations of secondary structural elements. For example, β sheets are usually followed by

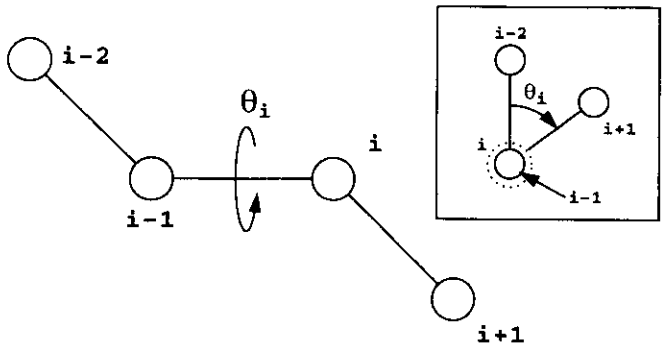

FIG. 1. The dihedral (also called torsion) angle, Θ_i, associated with the ith monomer. The inset shows the view along the bond from monomer $i-1$ to i. The angle shown is defined as positive by our sign convention.

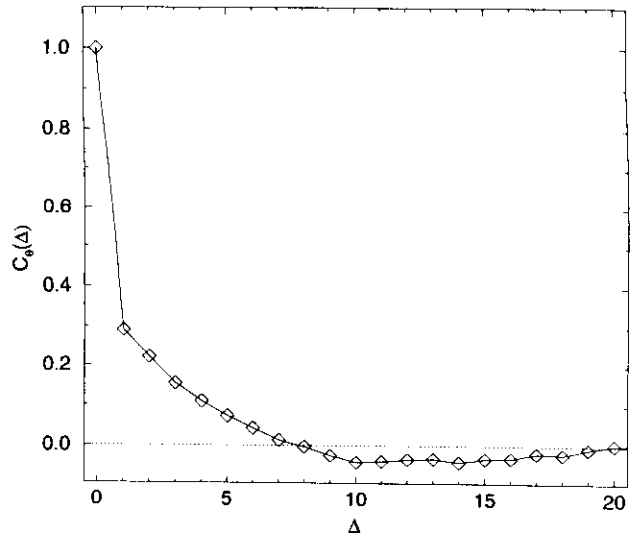

FIG. 2. Real part of the dihedral angle correlation function averaged over 112 proteins from the protein data bank. The distance, Δ, is the number of monomers along the chain. $C_\theta(0)$ has been normalized to one.

reverse turns. There is also the $\beta\xi\beta$ unit where two parallel β sheets are separated by some piece ξ which can be a random coil, an α helix, or another sheet [14]. Eventually the correlations fall off to zero (at around $\Delta = 16$).

We now examine what forces drive the formation of secondary structure, specifically the question of whether the loss of entropy due to compaction is sufficient. To do this we need a model without any local interactions. Lattice models are not acceptable since the restricted degrees of freedom imply local bond angle potentials. An off-lattice model was used instead. As in lattice and other simple models we neglect the internal degrees of freedom of the amino acids and represent each as a single point in space. Monomers that are connected along the chain are constrained to be separated by a fixed distance. The next step is to fold the chains into compact conformations. The following procedure was used. Take a potential energy function whose minima are compact conformations. Then minimize this potential energy to fold the chain. Because the model we are using is a *homopolymer* there are many compact local minima (the number grows exponentially with chain length [10]). We will generate an ensemble of compact conformations, using chains of several different lengths. One can think of this ensemble of different compact structures as representing the collection of native structures of many different sequences of amino acids. We will calculate the average correlation function [Eq. (1)] of the ensemble of compact conformations we generate and look for long range correlations which will indicate the presence of secondary structure. It is important to note that the previous works showing the connection between compaction and secondary structure [8–11] also used a homopolymer model and many homopolymers show secondary structure in their compact states [12]. Therefore, it does not appear necessary to have a heteropolymer and a unique ground state to get secondary structure.

There are several different potentials that have compact minima. The dominant force for the folding of pro-

teins is the *hydrophobic effect* [15]. This is primarily a bulk, entropic effect caused by interactions of the polymer with the surrounding water. The protein collapses to create a hydrophobic core with polar groups on the surface. One could simulate a polymer in a solution of water, however, this is much more complex than necessary. Instead of doing a full water-polymer simulation we simply choose an effective potential which will also cause the polymer to collapse. The particular one used in this work was

$$V(\{\vec{r}_i\}) = \sum_{i=1}^{N-1} \frac{1}{2} k_c \left(|\vec{r}_i - \vec{r}_{i+1}| - l_c\right)^2 + \epsilon \left\{ \sum_{i<j}^{N} \left(\frac{\sigma_{ev}}{r_{ij}}\right)^{12} - \frac{1}{N} \sum_{i=1}^{N} |\vec{r}_i - \vec{r}_{com}|^2 \right\}, \quad (2)$$

where $r_{ij} = |\vec{r}_i - \vec{r}_j|$, $\vec{r}_i$ is the position of the ith monomer, and $\vec{r}_{com} = \frac{1}{N}\sum \vec{r}_i$ is the position of the center of mass. The first term represents the covalent forces that bind the monomers along the chain. The constants k_c and l_c are both set equal to 1, determining the energy and length units. The middle term (which is the repulsive part of a Lennard-Jones potential) is the excluded volume term which prevents the chain from compacting to a single point. The last term is the radius of gyration of the chain. This term provides the compacting force. The two constants, ϵ and σ_{ev}, are determined by examining real proteins. The difference in energy scales between covalent and noncovalent forces determines ϵ. In proteins the typical noncovalent interaction is roughly one-hundredth the energy of a covalent bond, so ϵ is set equal to 0.01 [16]. The compactness of the chains will be controlled by the value σ_{ev}. To determine the value of σ_{ev} and measure compactness we looked at two features of real protein structure: the pair-correlation function (also called the radial distribution function) and the radius of gyration. First, the pair-correlation function was measured for both real proteins and our chains. This function gives the probability that two α carbons are separated by a given distance, indicating how closely the α carbons are packed together. We adjusted σ_{ev} until the position of the nearest neighbor peak for our chains closely matched the one for real proteins [17]. Next, we measured the radius of gyration as a function of chain length for real proteins. Our chains had a slightly smaller radii of gyration as proteins the same length (see Fig. 3). This is not surprising since the potential we used will generate nearly spherical shapes while proteins are ellipsoidal with varying eccentricities. An ellipsoid will have a larger radius of gyration than a sphere of equal volume.

The chains were compacted by minimizing this potential energy [Eq. (2)]. The algorithm used was a conjugate-gradient descent minimizer [18]. At each iteration in this algorithm the energy is decreased, so it is somewhat analogous to a zero temperature Monte-Carlo simulation, in that only energy reducing steps are accepted. There is the possibility that for some potentials this type of algorithm will be trapped in local noncompact minima. However, for the potential used here, this

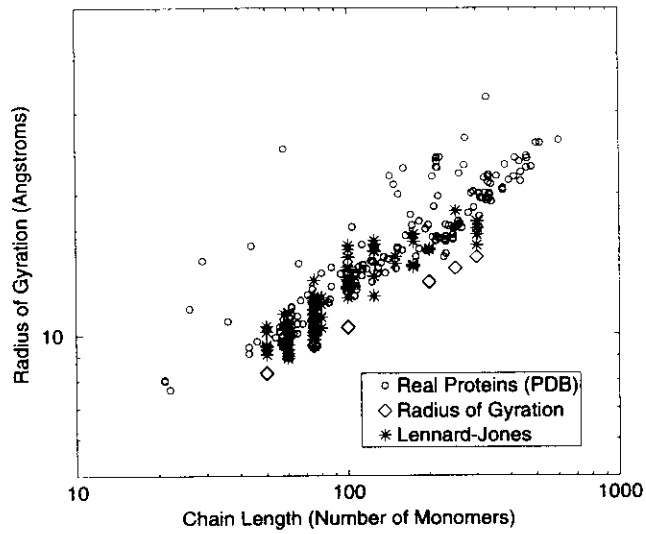

FIG. 3. The radius of gyration versus chain length (plotted on a log-log scale) for real proteins (small circles), chains compacted using the radius of gyration potential (diamonds), and the Lennard-Jones potential (stars). The radius of gyration for the three systems is very similar indicating that they all have the same level of compactness.

was not a problem. All minima that we generated were observed to be compact; i.e., their radius of gyration was roughly the same as those of proteins the same length (see Fig. 3). Starting from a random initial condition (which was taken to be a self-avoiding random walk) 200 chains, ranging in length from 50 to 450 monomers [19], were folded. The average dihedral angle correlation function was then calculated for these chains to determine if any secondary structure was present. Figure 4 shows the average for the compacted chains with the correlation function for real proteins superimposed. The compacted chains show no long range correlations. The plot falls almost immediately to zero, with a slight negative corre-

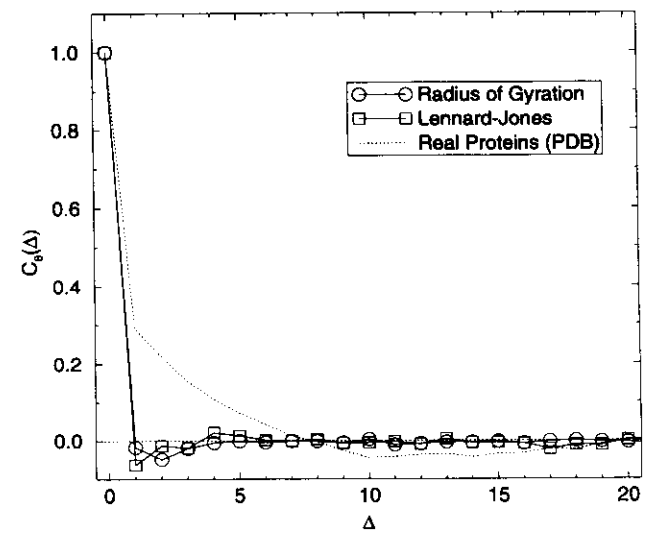

FIG. 4. The two solid lines show the correlation functions for the radius of gyration potential (circles) and Lennard-Jones potential (squares). The dotted line is the real protein correlations (from Fig. 2) for comparison.

lation at separations of roughly two monomers. This lack of any correlations indicates the absence of any secondary structure.

The potential [Eq. (2)] was chosen to have no local interactions other than the one term which bonds a monomer to its two neighbors along the chain. Again, local here means local (close) as measured along the chain, not through space. The excluded volume term is through space local, but in a folded structure any two monomers can interact via the excluded volume term regardless of their separation along the chain. In particular, there is no angle term in the potential (either implicit or explicit). The previous works which did find secondary structure with increasing compactness did have implicit angle potentials. It appears that compacting the chain is not enough to generate secondary structure. It is possible that the particular form of the compacting potential we used destroys secondary structure or was biased in favor of compact conformation without secondary structure.

To test this we tried a different compacting potential, the Lennard-Jones 6-12 potential. We replaced the radius of gyration term in Eq. (2) by a r^{-6} term to give

$$V(\{\vec{r}_i\}) = \sum_{i=1}^{N-1} \frac{1}{2} k_c (|\vec{r}_i - \vec{r}_{i+1}| - l_c)^2 + \epsilon \sum_{i<j}^{N} \left\{ \left(\frac{\sigma_{ev}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ev}}{r_{ij}} \right)^{6} \right\}. \quad (3)$$

By itself the 6-12 potential is too short-ranged to compact an extended chain so we did a two stage minimization. At the first stage we added an additional $1/r$ piece which is long ranged and will collapse an extended chain. Once the chain was semicompact, we finish the minimization without the $1/r$ term. We generated an ensemble of compact chains and measured the average correlation function (see Figs. 3 and 4). Again there were no long range correlations, hence no secondary structure.

To explore the forces responsible for the formation of secondary structure in proteins we have defined a simple, generic method of measuring secondary structure in polymers. This method consists of calculating the angle correlation function along the chain and looking for long range correlations. If secondary structure is present there will be long range correlations with a length comparable to average size of the secondary structure. This method does not depend on the precise details of what type of structure is present and can be used when these details are not known. Real proteins whose structures have been solved were examined and long range correlations were found. This technique was then used to examine whether compaction leads to the formation of secondary structure. Simple models with no local interactions were used and two different compacting potentials were examined. There were no long range correlations indicating the absence of secondary structure. These results indicate that compaction by itself is not sufficient to generate secondary structure. In the previous studies demonstrating a connection between secondary structure and compaction there was always some form of local interactions present. It appears, however, that local interactions are not sufficient since compactness was also necessary to get structure. In proteins the formation of secondary structure appears to result from the combination of both the entropic effect of compaction and local energetic effects. The loss of entropy from compaction is not enough to force the chain into regular conformations. Using our definition of secondary structure further studies can be carried out to determine the relative importance of these two factors.

We acknowledge helpful discussions with S. Skourtis, A. Libchaber, A. Schweitzer, and S. Favarolo. Work at San Diego was funded by the Arnold and Mabel Beckman Foundation and the National Science Foundation (Grant No. MCB-9018768). Portions of this work were done at the University of California, Berkeley and were supported by the NSF, supplemented by funds from Sun Microsystems, Cray Research, and the NEC Research Institute. J. N. O. is in residence at the Instituto de Física e Química de São Carlos, Universidade de São Paulo, São Carlos, SP, Brazil during part of the summers.

[1] H. S. Chan and K. A. Dill, Phys. Today **46** (2), 24 (1993).
[2] J. Bascle, T. Garel, and H. Orland, J. Phys. I (France) **3**, 259 (1993).
[3] G. Iori, E. Marinari, G. Parisi, and M. V. Struglia, Physica A **185**, 98 (1992).
[4] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. USA **89**, 8721 (1992).
[5] E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[6] M. Sasai and P. Wolynes, Phys. Rev. Lett. **65**, 2740 (1990).
[7] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).
[8] H. S. Chan and K. A. Dill, Annu. Rev. Biophys. Biophys. Chem. **20**, 447 (1991).
[9] H. S. Chan and K. A. Dill, Proc. Natl. Acad. Sci. USA **87**, 6388 (1990).
[10] H. S. Chan and K. A. Dill, Macromolecules **22**, 4559 (1989).
[11] L. M. Gregoret and F. E. Cohen, J. Mol. Biol. **219**, 109 (1991).
[12] H. Tadokoro, *Structure of Crystalline Polymers* (Wiley, New York, 1979).
[13] F. C. Bernstein *et al.*, J. Mol. Biol. **112**, 535 (1977).
[14] G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure* (Springer-Verlag, New York, 1979).
[15] K. A. Dill, Biochemistry **29**, 7133 (1990).
[16] The actual value of ϵ is not very critical. It simply must be small enough that the covalent interactions provide rigid constraints between monomers along the chain.
[17] N. D. Socci, Ph.D. thesis, University of California at Berkeley, 1992 (unpublished).
[18] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes* (Cambridge University Press, New York, 1986).
[19] The distribution of lengths was chosen to approximately match the length distribution of real proteins taken from the data bank.

Folding kinetics of proteinlike heteropolymers

Nicholas D. Socci and José Nelson Onuchic
Department of Physics, University of California at San Diego, La Jolla, California 92093-0319

(Received 7 March 1994; accepted 31 March 1994)

Using a simple three-dimensional lattice copolymer model and Monte Carlo dynamics, we study the collapse and folding of proteinlike heteropolymers. The polymers are 27 monomers long and consist of two monomer types. Although these chains are too long for exhaustive enumeration of all conformations, it is possible to enumerate all the maximally compact conformations, which are 3 $\times 3 \times 3$ cubes. This allows us to select sequences that have a unique global minimum. We then explore the kinetics of collapse and folding and examine what features determine the various rates. The folding time has a plateau over a broad range of temperatures and diverges at both high and low temperatures. The folding time depends on sequence and is related to the amount of energetic frustration in the native state. The collapse times of the chains are sequence independent and are a few orders of magnitude faster than the folding times, indicating a two-phase folding process. Below a certain temperature the chains exhibit glasslike behavior, characterized by a slowing down of time scales and loss of self-averaging behavior. We explicitly define the glass transition temperature (T_g), and by comparing it to the folding temperature (T_f), we find two classes of sequences: good folders with $T_f > T_g$ and non-folders with $T_f < T_g$.

I. INTRODUCTION

It has been known for some time that for many proteins the information necessary to specify the native structure is contained within the amino acid sequence. There has been a tremendous amount of research aimed at deciphering this code and determining the final structure from the sequence. Solving this problem is of paramount importance; however, simply knowing how to map sequences to structures would leave many interesting questions unanswered. How do proteins fold to their native structure and, more specifically, how do they manage to fold so quickly? What are the key factors that determine whether or not a given sequence will fold and what the folding time will be? One may argue that it might be necessary to solve these problems before it will be possible to solve the folding problem (i.e., predicting structure from sequence).

A great deal of work (both experimental and theoretical) has been done on the kinetics of protein folding. One extremely useful theoretical technique is to study simple heteropolymer models. The idea is to reduce the complex system of proteins in solution to its bare essentials, leaving only the key features. The advantage of studying these simpler models is that an in-depth analysis (sometimes even an exhaustive one) can be performed, yielding detailed answers and information. This information should, in turn, provide insights into real proteins.

One class of model that is often used in theoretical polymer work is the lattice model, where the monomers are constrained to lie on lattice sites. Excluded volume is included by allowing only one monomer per site. To study dynamics, the Monte Carlo algorithm with a variety of move sets is used. Some of the earliest work using lattice models on proteins was done by Gō and others[1,2] using two- and three-dimensional lattices to examine the folding process. However, the interaction potential they used was somewhat unusual. The native state was explicitly built into the potential. The energy of any given conformation was determined by counting the number of native contacts, i.e., contacts found in the native structure. An attractive contribution to the energy was added for each native contact formed. This potential is somewhat unphysical, depending on an *a priori* knowledge of the native structure. Although much of this early work on lattice models was on simple cubic lattices, Skolnick and others[3–7] have used more complex lattices which are able to more faithfully represent the structure of actual proteins. Using these lattices they are able to model real protein structures (e.g., secondary structure) and study the dynamics of folding and the formation of these structures. However, with increasing complexity it becomes more difficult to study these models in great detail.

Rather than trying to model real proteins exactly, some have opted for simpler models which permit a more thorough analysis. Chan and Dill[8–11] have used a two-dimensional simple cubic lattice model with two monomer types (a polar monomer, P, and a hydrophobic one, H). The potential used models the hydrophobic interaction and is equal to $-\epsilon$ times the number of hydrophobic contacts (HH). They studied short chains, which allowed them to do exhaustive enumeration to measure a variety of properties (both static and dynamic). For dynamics they used both Monte Carlo[9] and transfer matrix methods.[10,11] By using short polymers, they were able to construct the full transfer matrix (this matrix determines the probability of one state transforming to another) and use it to solve exactly for the dynamics of the system. Although their model is simpler than an actual protein, it has yielded a wealth of interesting information and provided valuable insight into proteins and heteropolymers. Their models show a two-phase process similar to that found in proteins. There is a rapid collapse to compact states, followed by slower reconfiguring of the chains to the native structure. Fiebig and Dill[12] show that simple searching strategies, such as the formation of opportunistic hydrophobic contacts, can lead to the globally optimal conformation (na-

 0021-9606/94/101(2)/1519/10/$6.00

tive state), suggesting a possible mechanism for folding. Shakhnovich and others[13,14] have studied the folding of random heteropolymers (the interaction between monomers is picked from a random distribution) on the three-dimensional simple cubic lattice. They examined 27 monomer polymers using Monte Carlo dynamics and also found a two-stage collapse process in folding. They found that by examining an overlap function, which measures how low-energy conformations differ, they could distinguish the difference between foldable and not foldable sequences. From examination of many different sequences, they conclude that the existence of a pronounced energy gap between the native state and the remaining conformations distinguishes good folding sequences.[15] To examine how the specific form of the interaction affects the dynamics of folding, Camacho and Thirumalai[16] looked at two-dimensional lattice systems. They studied the kinetics of three different types of interaction potentials. They found two transition temperatures: A collapse temperature at which the chain forms a compact structure and a folding temperature at which the native structure is formed. They found three stages in the transition from open coil to native structure.

In this work we will continue using the three-dimensional simple cubic lattice model. The polymers will be 27 monomers long and consist of two monomer types. Monte Carlo dynamics will be used to study the collapse and folding kinetics. The chains are too long for exhaustive enumeration of all conformations but are short enough to permit exhaustive enumeration of all maximally compact configurations. This information will be used to determine the minimum energy structure (native state) which will allow us to measure the folding time from extended conformations. We will examine several different sequences and measure collapse and folding time as a function of temperature and sequence. One question to be addressed is which kinetic quantities are sequence dependent and which are sequence independent (*self-averaging*). In addition, we will examine how the glass transition affects the ability of a sequence to fold. A major goal is to define, as precisely as possible, various physically important quantities. Of particular importance will be the determination of the important time scales. One problem with Monte Carlo dynamic simulations is the relation between Monte Carlo steps and physical time. There is no simple connection; in fact, the precise relation may depend on the move set.[10,11] To circumvent this problem, we will relate Monte Carlo steps to physical time by looking for the natural time scales in the problem, such as the collapse and the folding time. Using these time scales, we will then be able to define the glass transition temperature (T_g) of this model. In the past others have speculated that the relation between the folding temperature (T_f) and the glass temperature (T_g) would play an important role in protein folding. Bryngelson and Wolynes[17,18] have proposed that in order for a chain to fold, the folding transition must occur before the glass transition of the system, and the optimal folding temperature would lie between T_f and T_g. Specifically, Wolynes and others state that to optimize folding potentials for structure prediction, one should maximize the ratio of the folding temperature to the glass temperature (T_f/T_g).[19,20] To calculate the glass transition, they used a random energy model-like assumption; i.e., for each given value of the degree of folding, the energies of the different conformation are independent random variables. In our work we will give a direct kinetic definition of the glass temperature that does not rely on this assumption, and show explicitly that the relative values of T_g and T_f will determine the folding properties of a given sequence.

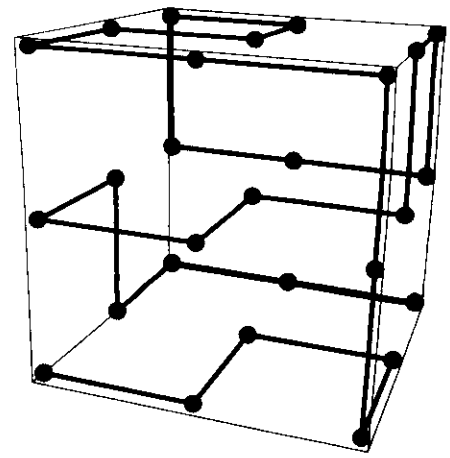

FIG. 1. An example 27 length polymer on a three dimensional simple cubic lattice. The conformation is a maximally compact cube. The light and dark spheres represent the two different types of monomers. The sequence shown here is 013 and the conformation shown is the native (minimum energy) state.

II. MODEL AND METHODS

The model we used in this work is a three-dimensional lattice polymer. Monomers that are connected along the chain are constrained to be nearest-neighbors on the lattice, and only one monomer is allowed per site. (This is the excluded volume condition.) The chain is then a self-avoiding walk on the lattice. The polymers are all 27 monomers long. The maximally compact state is a $3\times3\times3$ cube (see Fig. 1). Although it is not feasible to enumerate all configurations of a 27 monomer chain, it is easy to enumerate all the compact cubes, of which there are 103 346. If we choose a potential that favors the formation of contacts, then the minimum energy conformation will usually be a compact cube. Selecting such a potential enables us to determine the native structure of a given sequence by enumeration of the cubes, since for this simple model the native state is the lowest energy conformation. In addition, the degeneracy of the lowest energy state can be determined. Since we are interested in protein-like polymers which have a "single" native state,[21] we will choose sequences with a nondegenerate ground state, i.e., those with only one lowest energy conformation.

We want a potential that will favor compact states and cause the chain to fold. The dominant force in protein folding is the hydrophobic effect.[22] This force is a many-body interaction between the hydrophobic side chains and the solvent (water). The main effect is to cause the chain to collapse and create a hydrophobic core. In our simulations we model

this effect by using an attractive potential to collapse the chain. This potential favors the formation of contacts between any two monomers. However, we do not want a homopolymer, so the interaction energy is dependent on whether the two monomers in contact are of the same type or not. The potential is given explicitly by

$$E=\sum_{\substack{\langle i,j\rangle \\ |i-j|\neq 1}} H_{t_i,t_j}, \tag{1}$$

where the sum is over all nearest neighbor pairs on the lattice, excluding covalently linked pairs. The type of monomer i is t_i which we will denote with A and B. H_{t_i,t_j} is the interaction matrix given by

$$H_{t_i,t_j}=\begin{array}{c} \\ \mathrm{A} \\ \mathrm{B} \end{array}\begin{array}{c} \begin{array}{cc} \mathrm{A} & \mathrm{B} \end{array} \\ \begin{pmatrix} E_l & E_u \\ E_u & E_l \end{pmatrix} \end{array}. \tag{2}$$

E_l is the energy for a contact between monomers of the same type, and E_u is for contacts between unlike monomers. To collapse the chain, we pick both energies to be negative with $E_l<E_u<0$, favoring contacts between monomers of the same type.

We now need to specify the dynamics of the model. For lattice systems there is no unique way to do so, and it has been shown that different move sets may give very different kinetic behavior. In a study of homopolymer folding kinetics, Chan and Dill[10,11] showed that various kinetic quantities, like the collapse time and the mean first passage time, will depend on the type of moves allowed. Therefore, we wish to choose a set of moves that will give dynamics that are as realistic as possible. Care must be taken in analyzing the results of these simulations. In particular, one must not try to extract too much detail from these types of simulations. The right questions need to be asked. For example, we will look at how folding and collapse time varies with sequence. This is a generic question and the behavior should be universal to all reasonable move sets. A question which would be more difficult (or perhaps not even valid) for this simulation to answer would concern specific details of the folding pathway, for instance, the role of secondary structure formation in folding. One would imagine that depending on the move set used one would find very different answers to this question. To answer such questions, more realistic models with clearly defined dynamics are necessary.

The move set used consists of local moves which preserve the covalent links and keep each lattice site either singly occupied or empty. This set was developed some time ago to study the dynamics of polymers.[23–25] The allowed moves consist of end moves in which the ends of the chain move to an empty adjacent site, corner moves which flip a single monomer and crankshaft moves which move two monomers simultaneously (see Fig. 2). Studies involving this move set have shown that it closely reproduces the relaxation dynamics of the Rouse model.[25,26] More complex moves are possible where more than two monomers are moved simultaneously. They have the advantage of allowing concerted motion of structural elements (like helices). One must take into account the different rates of these more complex

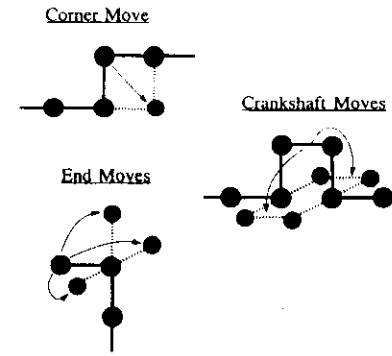

FIG. 2. The three types of moves used in the dynamics simulations. The light circles represent the possible lattice points a given monomer can move to provided that that point is not occupied. In the case of the end and crankshaft moves one of the possible moves is picked at random. Note that the corner and crankshaft moves are exclusive: A nonend monomer can only make one or the other depending on the position of its neighbors along the chain.

moves; i.e., a 5 monomer motion should occur more slowly than the flipping of a single monomer. If this is not taken into account, the time scales will be distorted since different moves, all of which can be performed in one iteration of the algorithm, will have different "physical" times. To correct for this, one can assign a different probability for each of the moves.[3,4] Since our chains are relatively short, we use only the simple one and two monomer moves. We do not believe that including the possibility of concerted motions of large subsections of the polymer will change the answers to the questions asked here.

There is one important comment to be made about this set of moves—they are not ergodic. In particular, it is not possible to reach the configuration in Fig. 3 from an extended chain.[27] The question is whether the nonergodicity of our moves will create a problem. The simple answer to this

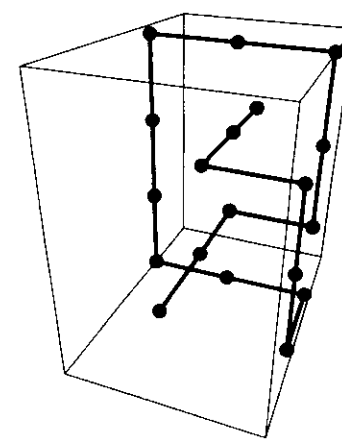

FIG. 3. A knotted conformation (Ref. 27) which can not be unknotted by the moves shown in Fig. 2; consequently, from an unfolded conformation it is unreachable using the same moves. Hence those moves are nonergodic. As long as conformations like this are not the native state they will pose no problem. What is important is the native state is accessible.

TABLE I. The various sequences used in this paper. The last four (005, 006, 007, 013) were generated at random. Sequence 002 was optimized by Shakhnovich (Ref. 15). Sequence 004 is a single monomer mutation of 005 ($B_{13} \rightarrow A$). Both 002 and 004 have the lowest energies possible for the potential used and have native states that are completely unfrustrated. τ_{min} is the fastest folding time for each sequence. T_g is the glass transition temperature (calculated with a $\tau_{max} = 1.08 \times 10^9$). T_f is the folding temperature calculated using the Monte Carlo histogram method. The numbers in parenthesis indicate the uncertainty of the last digit.

Run	Sequence	E_{min}	τ_{min}	T_g	T_f
002	ABABBBBBABBABABAAABBAAAAAAB	−84	2.0×10^7	1.00	1.285(15)
004	AABAABAABBABAAABABBABABABBB	−84	1.6×10^7	0.96	1.26(1)
005	AABAABAABBABBAABABBABABABBB	−82	2.3×10^7	0.98	1.15(2)
006	AABABBABAABBABAAAABABAABBBB	−80	5.2×10^7	1.07	0.95(6)
007	ABBABBABABABAABABABABBBABAA	−80	9.3×10^7	1.09	0.93(5)
013	ABBBABBABAABBBAAABBABAABABA	−76	9.7×10^7	1.01	0.83(5)

problem is that we are interested in the kinetics of folding and as long as the native state is accessible, there should no problems due to the existence of inaccessible states. In particular, we can view these states as being irrelevant in the same way that highly unlikely states are irrelevant for real proteins. The chain in Fig. 3 actually forms a knot. We know that real proteins do not have "tight" knots,[28,29] and it is highly unlikely that, in folding, a protein passes through a knotted state. Strictly speaking, it is not impossible for a protein to become knotted; it is just unlikely. Due to the constraints of the lattice, the knotted state is now inaccessible rather than unlikely; however, its existence will have no effect on the folding properties. One may argue that these inaccessible states may still affect thermodynamic calculations. In particular, what will be the effect of the fact that our moves restrict us to a ergodic subspace of the full phase (conformation) space of the system? That will depend on the relative sizes of the excluded space. In practice, for small chains, the errors introduced by the nonergodicity of this move set is smaller than the statistical error. Comparison with other ergodic algorithms show no change in the results.[25]

A move is made using the Metropolis Monte Carlo algorithm.[30] A monomer is selected at random. If it is an end monomer, then one of the neighboring lattice point is also selected at random. If it is not an end piece, then it can do either a corner move or a crankshaft move, depending on the position of its neighbors along the chain. In the case of the crankshaft, the possible direction is also selected at random. If the move selected would violate the excluded volume constraint by moving the monomer to an occupied site, the old configuration is counted once more in averaging (i.e., a step is considered to have elapsed), and a new monomer is picked. If a move is possible, that is, if the lattice site is empty, then the energy of the new conformation is calculated and compared to the original energy. If the energy decreases, then the move is accepted unconditionally. If the energy increases, then the move is accepted with the usual Boltzmann probability:

$$P = \exp[-(E_{new} - E_{old})/T], \tag{3}$$

where E_{new} and E_{old} are the new and old energies, respectively, and T is the temperature. Note that we have chosen units such that the Boltzmann's constant is unity ($k_B = 1$). Whether a move is accepted or not, one unit of time (a Monte Carlo step) is considered to have elapsed.

There is no simple and direct connection between Monte Carlo steps and the physically relevant times scales of the system. One important result of this work is that we will determine the mapping between the physically important time scales such as the folding time and the computation time scales (Monte Carlo steps). This will be useful in later works in which we will study the thermodynamics of these systems where it will be necessary to know how long it takes for systems to reach equilibrium and explore conformation space. Once again, the precise connection between physical time (like the folding time) and Monte Carlo steps will depend on the details of the move set used. However, we expect that as long as the moves are chosen with care, that is, one attempts to make it as physically realistic as possible, then the qualitative features will remain the same. For example, the behavior of folding time as a function of temperature will look qualitatively the same although the exact folding time (number of steps) will vary.

III. RESULTS AND DISCUSSION

In this work we studied six sequences, all 27 monomers long. Four were selected by the following procedure. First a sequence was generated at random with the appropriate ratio of monomer types. We then enumerated all cubes calculating the number of minimum energy states. Sequences with degenerate minimum energy states were rejected. From the remaining nondegenerate sequences we picked four which had a spread in energy of the native states (-82 to -76). One sequence was obtained by changing a single monomer in one of the original four; i.e., it is a single-site mutation, which lowered the ground state energy, from -82 to -84. The last sequence was taken from a paper by Shakhnovich[15] which gave a method for finding optimal sequences; it also has a ground state energy of -84. Table I shows the various sequences along with some data for each. All six sequences have the same ratio of monomer types, roughly 50:50 (14:13, actually). For these simulations the value for the contact energies are -3 for monomers of the same type and -1 for unlike monomers. Again, the units used are such that $k_B = 1$. The maximally compact conformation of a 27 monomer chain is the $3 \times 3 \times 3$ cube. There are 28 noncovalent or

topological contacts in this cube, so the lowest possible energy is -84. Two of the sequences have this energy for their minimum conformation. This corresponds to a completely "unfrustrated" ground state. By "unfrustrated," we mean that the ground state has no topological contacts between monomers of different types. The other three sequences have ground state energies higher than -84 and, consequently, their ground states have at least one bad topological contact and are frustrated energetically.

Using the Monte Carlo moves described above, we simulated the folding of these sequences and examined how the folding behavior depends on temperature and differs from sequence to sequence. For each sequence we started with a random, completely unfolded, initial conformation. Here, completely unfolded means that there are no contacts between any of the monomers. A temperature was selected, and the sequence was allowed to fold. Once the sequence found its folded state (which we detected by monitoring the energy), we stopped the simulations. The simulations were also stopped if the sequence did not find its folded state within τ_{max} steps. For the simulations in this work, $\tau_{max}=1.08\times10^9$ Monte Carlo steps. This maximum time was picked both so that it was longer than the typical folding time of most sequences and to minimize the actual computer time used in the simulations. It is important to realize that any times longer than τ_{max} are undefined. Ideally, we would want to pick τ_{max} to be longer than any interesting and relevant physical or biological time scale. Since there is no simple connection between Monte Carlo steps and physical time, we cannot directly determine τ_{max}. We chose a first value for τ_{max} which seemed reasonable and then made sure that it was much longer than the folding time for the various sequences.

At each temperature we ran many simulations, each with a different random initial condition (always unfolded). We then calculated an average folding time (τ_f) from these runs. This time is the mean first passage time from the set of unfolded initial states to the folded state. Figure 4 shows τ_f as a function of temperature. Once again, the units of temperature and energy have been chosen so that $k_B=1$. We ran anywhere from 10 to 600 simulations at each temperature and calculated the average folding time. If the folded state was not found within τ_{max} steps, we averaged in τ_{max} for that run. The error bars are the standard deviation of the mean given by $\sigma/\sqrt{N}$, where σ is the standard deviation of the distribution of folding times and N is the number of runs at that temperature. It is important to note that since we average in τ_{max} when the chain does not fold, the error bars are not as meaningful at temperatures where the folding time approaches τ_{max} and may be much larger at these temperatures. In particular, at high and low temperature the points equal τ_{max} with zero error. That is simply due to the fact that at those temperatures the simulations never found the folded state. Figure 5 shows the fraction of times the folded state was found as a function of temperature. It has a maximum plateau over the same temperature range that τ_f has a minimum plateau. At temperatures where the chain folds rapidly, it also finds its native state 100% of the time. At temperatures where the simulations did not find the folded state all the time, the τ_f shown in Fig. 4 is a lower bound to the actual mean first passage time. Figure 6 shows the distribution of folding times for three temperatures. At the temperature of fastest folding ($T=1.58$) the distribution of times is narrowest. For temperatures above and below this the distribution becomes quite broad. All three histograms appear roughly Poissonian. The standard deviations are approximately equal to the means.

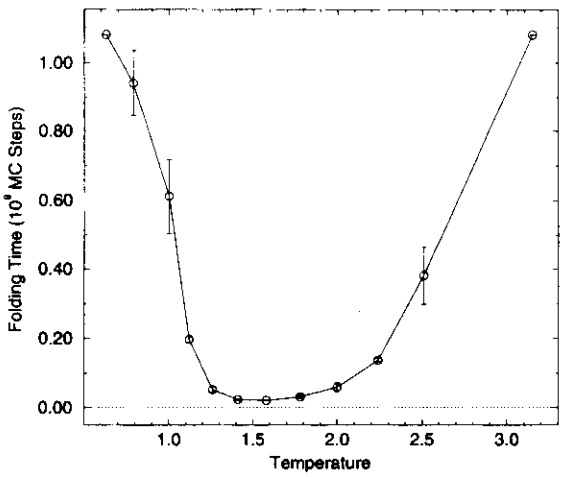

FIG. 4. Mean folding times versus temperature for one sequence (sequence 002). Note both axes are linear scale and the time is in billions of Monte Carlo steps. The error bars are the *standard deviation of the mean*: That is, they are equal to the standard deviation of the folding-time distribution at a given temperature divided by the square root of the number of runs at that temperature.

We observe three different temperature regions, similar to those found in two-dimensional lattice simulations.[9,11] Above a temperature of ~3 and below ~0.65, the chains did not fold within τ_{max} steps. Between these temperatures the folding time drops rapidly to approximately $2\times10^7-5\times10^7$. The fraction of runs that find the folded state increases sharply from 0 to 1 in this temperature range.

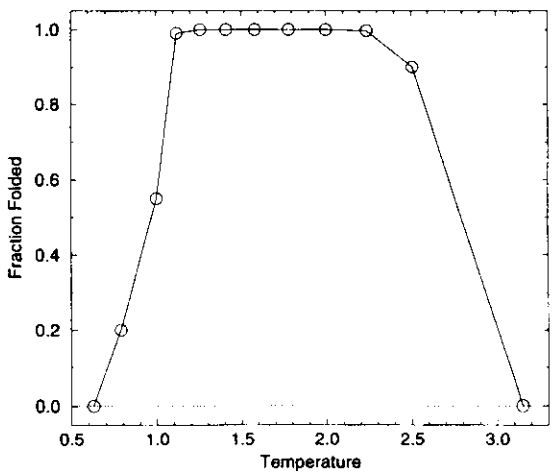

FIG. 5. Fraction of times that sequence 002 folded as a function of temperature. Note that the plateau at which the chains fold 100% of the time corresponds to the minimum folding time plateau in Fig. 4.

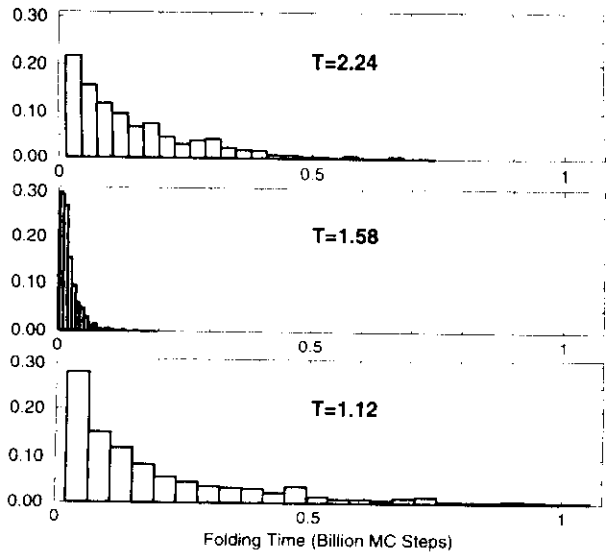

FIG. 6. Histogram of folding times for sequence 002 at three different temperatures. The histogram values have been normalized so the sum over all bins equals one. The minimum folding time temperature (1.58) is shown along with distributions above and below this temperature. The distributions are roughly Poissonian, the standard deviation being approximately equal to the mean.

In the next plot (Fig. 7) the folding time is plotted along with the chain compaction time. The compaction time is simply the number of steps it takes for an unfolded state to reach a maximally compact cube. In addition, we also show a time to reach a nearly compact state, which we define to be a conformation with 25 (out of 28) contacts. The behavior of the compaction time as a function of temperature is similar to that of the folding time, but chains compact much faster than

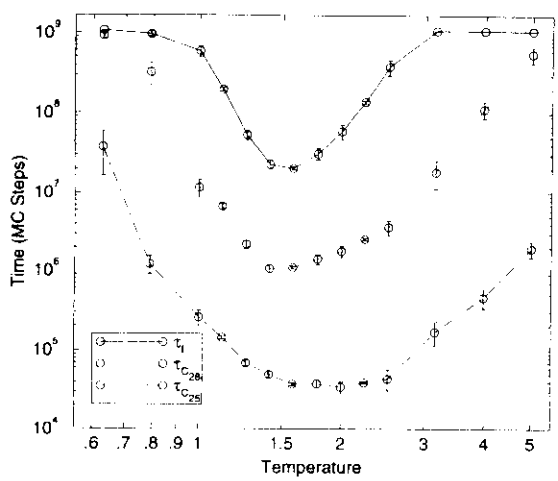

FIG. 7. Folding and Collapse times versus temperature for sequence 002. Note in this figure both axes are log scale and the time is in Monte Carlo steps. The solid line is the mean folding time, τ_f (same as Fig. 4, but now on a log scale). The dotted line, τ_{c28}, is the mean compaction time to any cube. The last line, τ_{c25}, is the mean compaction time to a partially compact conformation with 25 (out of 28) contacts. Error bars are the standard deviation of the mean.

they fold. This behavior is similar to what is believed to occur in real proteins in which the chain first folds rapidly to a compact state and then rearranges itself to the native structure.

Above a temperature of approximately 5, the compaction time approaches $\tau_{\max}$. Above this temperature the free-energy is dominated by the entropic term which favors non compact conformations which are far more numerous than compact ones. In the range from 5–2, we observe the following interesting behavior. The chains compact fairly rapidly, but the folding time is still quite long. In particular, at about $T \approx 3$ the chains compacted easily but never folded within $\tau_{\max}$ steps. We can draw a parallel between this state and the molten globule state of proteins.[31] At this temperature range, the temperature is low enough so that the potential, which favors contacts, drives the chain to a compact conformation, but the temperature is still too high for the potential to drive the chain to the native state. In this range we can imagine the chain is fluctuating about various compact states, "randomly" searching for the native state. This is like the often-discussed Levinthal paradox in which it is argued that a protein could not find its folded structure by random search. If the temperature is high enough there is no strong driving force that favors the native state. When the temperature is low enough, the chain is no longer randomly searching compact conformations but is driven to the folded state. This is, of course, the well-known resolution to the paradox. At the appropriate temperatures proteins do not randomly search for their native state but are directed to it by the shape of the free energy surface.

At still lower temperatures both the folding and compaction time start to increase again. At temperatures slightly less than 1 the folding time reaches $\tau_{\max}$ again and at a temperature of roughly 0.63 the compaction time approaches $\tau_{\max}$. At low temperatures the system is beginning to slow down, kinetically, and is now getting trapped in local *meta*-stable states. Even though we expect, at these low temperatures, the free energy to have a very pronounced minimum at the native state, the system is unable to reach it within a reasonable time. This region is often referred to as the glass phase. We can define a temperature at which the system undergoes a glass transition characterized by the slowing down of various times, such as the folding and compaction times. The autocorrelation time of the system would also increase in this temperature region, indicating that the chain was locally trapped. We define the glass transition temperature (T_g) as the temperature at which the folding time is half way between $\tau_{\max}$ and $\tau_{\min}$ (where $\tau_{\min}$ is the fastest folding time observed). Using this definition we get $T_g \approx 1$. Note that the definition of glass temperature is not the usual thermodynamic definition of temperature. It is not determined by the inverse of the derivative of entropy with respect to energy ($1/T_g = \partial S/\partial E$) at the point where the entropy "vanishes."[32] One difficulty with this definition is its relation to the kinetics of the system. The idea is that as a system is taken out of equilibrium, the time it takes to relax back will increase as the temperature gets closer to T_g. To avoid this kind of assumption, we have given a kinetic definition for T_g in which we will explicitly look for a slowing down of the

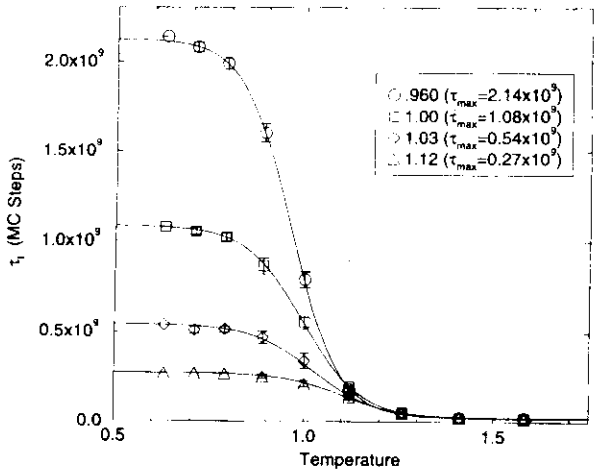

FIG. 8. Folding time as a function of temperature at several different τ_{max}. The glass temperature is the point at which the folding time is halfway between τ_{max} and τ_{min}. The legend shows the glass temperature for each value of τ_{max}.

system. We would expect the precise value of T_g to depend on the moves used and therefore we will not focus on the exact value but on the relative value. In particular later we will compare T_g to another important temperature, the folding temperature (T_f), and we will discover a key relation between the two. Additionally T_g depends on the value of τ_{max}; that is, it will depend on how long we run our simulations. This is a subtle but extremely important fact to remember when studying finite sized systems. When talking about glasslike behavior of a finite system the notion of glasslike depends on the time scale you are looking at. If you wait long enough the chain will always find the native state. To speak of a physically meaningful glass transition one must define the physical time scales of interest. The time scales of importance here are related to the minimum folding time of a good folding sequence. We want to examine our system on a time scale that is reasonable greater then the minimum folding time. For our simple system there is no obvious greater time to pick. For real proteins the life time of the organism would be a reasonable choice, since proteins need to fold on a time scale much shorter than this to be useful. We picked a time that was roughly two orders of magnitude greater than the folding time. Since this is somewhat arbitrary, we investigated how T_g changed as τ_{max} is varied. Figure 8 shows the results of several simulations in which τ_{max} was varied from one fourth the usual value (1.08×10^9) to almost twice this value. We see that although T_g decreases with increasing τ_{max} it does so quite slowly. The difference between the last two is only about 4%; therefore, T_g is not too sensitive to the precise value of τ_{max}.

The glass temperature just defined is related to the folding of the chains. One can also define a glass temperature that has to do with the slowdown of compaction. This would be the temperature at which the time it takes the chain to form a cube (28 contacts) is half way between the maximum and minimum times. Call this temperature $T_g(28)$. Examining Fig. 7 we see that $T_g(28)$ is less than T_g. (It is approximately equal to 0.7.) One could also consider $T_g(25)$ the glass temperature for forming 25 contacts (which is lower still). In general the transition temperature will be a function of some parameter, ρ, which is a measure of the compactness of the chain and/or similarity to the native state. Bryngelson and Wolynes[18] first calculated $T_g(\rho)$ in their random energy model.

The two regions in which the chain fails to fold are qualitatively very different. At high temperatures the energy differences between conformations becomes negligible so all conformations have roughly equal probabilities. The chain is randomly exploring the conformation space. It takes a long time to find the native state by random search due to the vast number of conformations. The free energy is dominated by the entropic term so the native state is no longer the global minimum. At low temperatures the energy differences between states becomes important and the folded state is the global minimum free energy. The problem now is that the barriers between states are too high and at low temperatures there is a very small probability for crossing them. For compact conformations many moves will involve the breaking of contacts which at low temperatures becomes unlikely. In particular, moves that break more than one contact are much less probable that those that break only one. Instead of a random search the chain is now forced in to a very narrow kinetic pathway consisting of those steps with very small free energy barriers. The chain gets trapped in the many local minima.

If we were willing to wait long enough the system would eventually fold. Since our system is finite, the system always has a finite nonzero probability to find the native state. The same is true for the high-temperature case: If we wait long enough, the chain will eventually find the folded state. However, one must remember that at those temperatures the folded state is not the free energy minimum and is therefore not stable. For example, consider the following two temperatures: 2.24 and 1.12. The folding time for these two temperatures is roughly the same. At the higher temperature (as we will see shortly) the chain spends almost zero time in the folded state (less than 0.04%). At the lower temperature the chain spends roughly 77% of the time in the folded state. When we speak of folding time, this is simply the time it takes the chain to find the native conformation. There is another important factor here: Namely, is the folded state stable thermodynamically? We will address this issue at the end of this Paper, where we see that it is not enough that a chain find its native state in a short time, but it must do so at temperatures where the native state is thermodynamically stable.

Let us return to the question of how long is too long to wait for a sequence to fold. Too long is in general determined by other time scales in the system. For proteins, there are a number of biologically relevant time scales, the lifespan of the organism for example. Proteins that do not fold fast enough on this time scale can be considered not to fold at all. Since we are studying a simple artificial system there is no *a priori* time scale to pick, other than limits on the simulation (computer) time. One of the problems with Monte Carlo dynamics is that there is no easy way to "calibrate" them, that

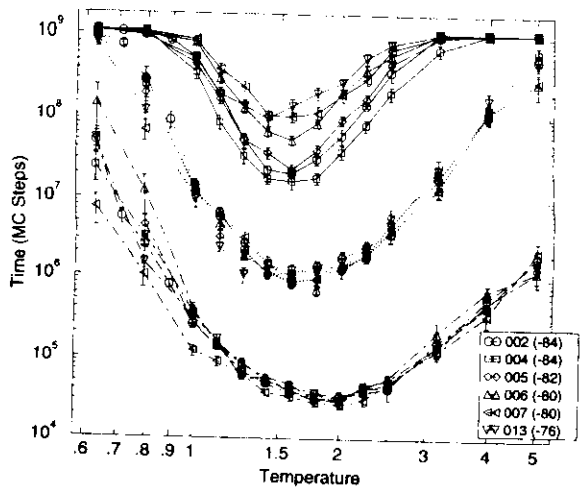

FIG. 9. Folding and collapse times versus temperature for all sequences, plotted on a log–log scale. The top set of solid lines are the folding times, the middle set of dotted lines are the times to compact to a cube and the bottom set are the times to compact to partially compact (25 contacts) conformation. The legend shows the energy of the native states.

is to make a connection between Monte Carlo steps and "real" physical or biological time. What we have done here is to define the relevant time scale as the folding time (or the compaction time) and make sure we ran simulations for long enough that we could see the the variation of folding time as a function of temperature.

We have examined the folding (and compaction) time as a function of temperature for one sequence. We now would like to see how this function varies from sequence to sequence. Figure 9 show a plot of the folding time and compaction time versus temperature for several sequences. From this figure we notice two very interesting features of this model. First the compaction time is sequence independent. All six sequences have roughly the same compaction time for temperatures above T_g. In contrast the folding time is highly sequence dependent. At a temperature of roughly 1.6 there is a difference of nearly an order of magnitude between the fastest and slowest folding sequence. The folding time is also roughly correlated with the energy of the folded state. The lower the energy, the faster the folding time. However, the relation between the energy of the native state and the folding time is not a simple one. For example the two sequences with the lowest energy folded states (sequence 002 and 004, see Table I) have different folding times. The difference is slight but sequence 004 has a consistently faster folding time at all temperatures. Sequence 005, which is a single monomer mutation of 004, has a higher ground state energy (-82), but its folding times are very close to those of the lower energy sequence 002. There also appears to be a fairly large difference between the sequences that have energies below -81 and those that have energies above.

Note that the collapse time is always much faster than the folding time for all sequences. Even sequences that fold slowly collapse as rapidly as the fast-folding ones. This sequence-independent property of the collapse time is often referred to as a *self-averaging* property; it does not depend on the specifics of the sequence but rather on the general character of the ensemble of sequences. It is important to remember here that we are choosing a restricted ensemble of sequences though, namely the subset of sequences with a particular ratio of monomer types (a ratio of 14:13). Sequences that contain a different composition of monomers may have different collapse times than the sequences used here. The folding time is not self-averaging; i.e., it depends on the sequence. So we can view the kinetics of folding as a two-stage process. The first involves a rapid collapse of the polymer. The nature of this collapse is sequence independent. We can picture the polymer in this collapsed state as fluctuating about various compact cube states. This picture has been advanced previously by others.[33] The next step is a medium-to-slow event in which the polymer searches for the minimum energy state among the compact states. The time it takes for the polymer to find its minimum state depends on the specifics of the sequence. The two-phase collapse with two distinct time scales has also been observed for real proteins.[34–36] The first phase is a rapid collapse in which a hydrophobic core is formed. We should expect this collapse to be independent of the specific sequence, depending on the ratio of hydrophobic to hydrophilic monomers. This collapsed state then undergoes rearrangement to the folded structure of the specific sequence. The collapse time below the glass temperature loses its self-averaging property. Examining Fig. 9 we see that below $T_g \approx 1$ the collapse time is no longer sequence independent. In the glass region we would expect the kinetics to depend on the details of the energy surfaces and these details will be sequence dependence. This is expected of a system exhibiting glassy behavior. Note how this contrasts with the high temperature limit, where the collapse times remain sequence independent even as they approach $\tau_{\max}$.

At this point one may be tempted to conclude that we have two types of sequences: fast folders and slow folders. However, this is not the case. In reality what we have are sequences that fold and sequences that do not. In order to see this we need to look at the thermodynamics of these systems. In particular we need to look at the thermodynamic stability of the native state as a function of temperature. To do so, we performed a series of thermodynamic runs using the same Monte Carlo algorithm described above. The system was equilibrated by first running it for 100 million steps, which is on the order of the folding time for most of the sequences. Care must be taken at low temperatures since near the glass transition the system will slow down; i.e., the autocorrelation time will diverge. We looked at temperatures above T_g to avoid this problem. We calculate the following thermodynamic quantity:

$$P_{\text{nat}}(T) = \frac{e^{-E_{\text{nat}}/T}}{Z}, \tag{4}$$

where E_{nat} is the energy of the native state and Z is the partition function. This quantity is the probability that the system is in the native state; that is, it is folded. We define the folding temperature as the temperature at which $P_{\text{nat}}(T_f) = 0.5$; that is when the folded state is half occupied. Note that $P_{\text{nat}}(T) > 0.5$ is a sufficient condition that the na-

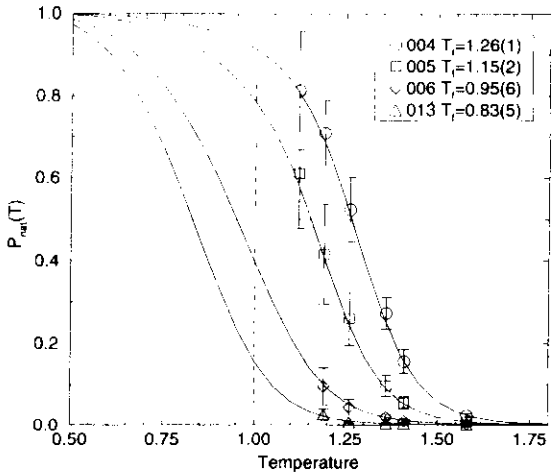

FIG. 10. $P_{nat}(T)$ for several sequences. T_f is defined to be the temperature at which $P_{nat}(T_f)=0.5$. The points were calculated using the standard Monte Carlo procedure. The solid lines are *not* fits to the data. They were calculated using the Monte Carlo Histogram Method (Ref. 37), which enables one to calculate thermodynamic quantities at temperatures other than the simulation temperature. Two of the sequences have folding temperatures above T_g, the others have temperatures below T_g. The vertical line at $T=1$ indicates the glass transition point. The legend shows the folding temperatures, the numbers in the parenthesis indicate the uncertainty in the last digits.

tive state be the global minimum of the free energy. Four sequences were used, each with a different minimum energy ranging from -84 to -76. Several simulations were run at a number of different temperatures. Figure 10 shows the results. In addition to running simulations at different temperatures, we also used the Monte Carlo histogram method[37] to calculate the function $P_{nat}(T)$ at temperatures other then the simulation temperature. Using histograms collected from simulations run at $T=1.58$, we were able to calculate P_{nat} for all temperatures, extrapolating into the glass region. These lines are plotted along with the points calculated from the standard Monte Carlo runs.

The folding temperature, T_f, varies with the value of the minimum-energy state. The lower the minimum energy the higher the folding temperature. More importantly we see that the two lower energy sequences have folding temperatures above the glass temperature, T_g, while the others have $T_f<T_g$. At T_g the lowest energy sequence (which also folds the fastest) is 90% in the native state. The highest energy sequence has a native state population of only 15%. At temperatures at which the folded state of the high-energy sequence is thermodynamically stable this state is not kinetically accessible. Therefore we would say this sequence does not fold. In order for a sequence to be foldable it must meet two conditions. First, it must have a reasonably fast folding time, where by reasonable we mean on the relevant (biological) time scales, and second, the folding temperature must be above the glass temperature. The analogous situation would be a polypeptide which had a folding temperature below the freezing point of the solvent. Such a protein would not be considered foldable.

IV. CONCLUSIONS

Using a simple lattice model and Monte Carlo dynamics we have studied the kinetics (and some thermodynamics) of proteinlike heteropolymer folding. Our results agree with previous works on other simple models and also match some of the properties of the folding real proteins. We find that our models display a two-stage folding behavior. First there is a rapid collapse to a compact state, followed by a slower stage in which the collapsed state rearranges itself to the native structure. We find that the folding time has a minimum plateau at intermediate temperatures and diverges at both high and low temperatures. The same is true for the collapse time. In this work we have examined the folding behavior as a function sequence and have discovered several interesting results. The collapse time and the glass temperature are both sequence independent (self-averaging) quantities. The folding time and temperature are both sequence dependent. The folding time correlates approximately with the energy of the native state: the lower this energy the faster the chain folds. This is consistent with the results found by Shakhnovich[15] that the larger the energy gap of the native state the better the sequence folds. We did not measure the gap, since there is no clear or simple definition of the gap in our system. Another way to view this result is that sequences with unfrustrated native states (native states with no bad contacts) fold best; i.e., we want to minimize energetic frustration of the ground state. However, we expect that this may be a property of these simple systems and that in more complex systems other forms of frustration (geometric or energetic frustration of conformations other than the native state) may play an important role. One would then expect that systems with reduced frustration should give rise to a large number of conformations that are rapidly connected kinetically to the native state (rapid compared to the folding time) or, as first proposed by Leopold and others,[33] a "dominant folding funnel."

An important point we have tried to stress is the issue of time scales, in particular the relevant physical time scales for this system and for protein folding in general. We note that there was no simple way to connect the computation time (Monte Carlo steps) to physical time. Rather than attempt to do so, we simply ran our simulations for a reasonable number of steps and then observed the folding time for the system. It is this folding time that now becomes the key time scale. For example, when we say a sequence does not fold what we mean is that is does not fold within a time that is over an order of magnitude greater than the folding time for the fast sequences. Since we are looking at finite systems we know that they will all fold given enough time. What is important is whether they fold in a reasonable time where reasonable is the folding time for the faster sequences. For real proteins, this time scale would be some suitable biological time.

By examining the behavior of folding time versus temperature we defined the glass transition temperature of this system. Below this temperature the kinetics slow down, causing the folding time to increase rapidly. Also the collapse times lose their self-averaging property and are now dependent on sequence. Most importantly we observed that for the

slow-folding sequences the folding temperature (the temperature at which the native state is half populated) is below the glass temperature. This indicates that these sequences will never fold since at temperatures where the native state is thermodynamically stable it is kinetically inaccessible. Good folding sequences have T_f greater than T_g. It has been suggested by others[19,20] that a good design principle for optimizing folding would be to maximize the ratio T_f/T_g. We observe this result explicitly in our simulations.

Perhaps the most interesting observation is that even simple systems such as these display a wide variety of complex and intriguing properties, many of which are shared by real proteins. This is particularly compelling in that one can much more easily study these simple systems and understand their behavior in great detail. By examining slightly more complex models we hope to understand how much of protein behavior is unique to proteins and how much is shared by the general class of heteropolymer systems. Hopefully, much of the apparent complexity of proteins will be understandable in the context of simpler model systems.

ACKNOWLEDGMENTS

We would like to gratefully acknowledge the computational assistance of A. Schweitzer. We also thank S. Skourtis, P. G. Wolynes, and K. Dill for helpful discussions. J. N. O. is a Beckman Young Investigator. This work was funded by the Arnold and Mabel Beckman Foundation and by the National Science Foundation (Grant No. MCB-9018768). J. N. O. is in residence at the Instituto de Física e Química de São Carlos, Universidade de São Paulo, São Carlos, SP, Brazil during part of the summers.

[1] Y. Ueda, H. Taketomi, and N. Gō, Biopolymers **17**, 1531 (1978).
[2] N. Gō, J. Stat. Phys. **30**, 413 (1983).
[3] J. Skolnick and A. Kolinski, J. Mol. Biol. **212**, 787 (1990).
[4] A. Sikorski and J. Skolnick, J. Mol. Biol. **212**, 819 (1990).
[5] J. Skolnick and A. Kolinski, J. Mol. Biol. **221**, 499 (1991).
[6] A. Kolinski, M. Milik, and J. Skolnick, J. Chem. Phys. **94**, 3978 (1991).
[7] A. Kolinski and J. Skolnick, J. Chem. Phys. **97**, 9412 (1992).
[8] H. S. Chan and K. A. Dill, Annu. Rev. Biophys. Biophys. Chem. **20**, 447 (1991).
[9] R. Miller, C. A. Danko, M. J. Fasolka, A. C. Balazs, H. S. Chan, and K. A. Dill, J. Chem. Phys. **96**, 768 (1992).
[10] H. S. Chan and K. A. Dill, J. Chem. Phys. **99**, 2116 (1993).
[11] H. S. Chan and K. A. Dill, J. Chem. Phys. (in press).
[12] K. M. Fiebig and K. A. Dill, J. Chem. Phys. **98**, 3475 (1993).
[13] E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[14] E. I. Shakhnovich and A. M. Gutin, Nature **346**, 773 (1990).
[15] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. USA **90**, 7195 (1993).
[16] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. USA **90**, 6369 (1993).
[17] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[18] J. D. Bryngelson and P. G. Wolynes, J. Phys. C **93**, 6902 (1989).
[19] M. Sasai and P. Wolynes, Phys. Rev. Lett. **65**, 2740 (1990).
[20] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **89**, 4918 (1992).
[21] When we say a protein has a single native state we are ignoring the different conformations real proteins may have. Our model is too coarse to represent the slight differences between conformations of proteins.
[22] K. A. Dill, Biochemistry **29**, 7133 (1990).
[23] P. H. Verdier and W. H. Stockmayer, J. Chem. Phys. **36**, 227 (1962).
[24] H. J. Hilhorst and J. M. Deutch, J. Chem. Phys. **63**, 5153 (1975).
[25] K. Kremer and K. Binder, Comp. Phys. Rept. **7**, 259 (1988).
[26] M. T. Gurler, C. C. Crabb, D. M. Dahlin, and J. Kovac, Macromolecules **16**, 398 (1983).
[27] N. Madras and A. D. Sokal, J. Stat. Phys. **47**, 573 (1987).
[28] G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure* (Springer-Verlag, New York, 1979).
[29] We are using the word knot here in a descriptive sense meaning what is commonly meant when one puts a knot in a string. Whether proteins have knots in them in the mathematical sense is rather subtle since proteins are open loops and mathematical knots are defined on closed loops.
[30] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. N. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[31] O. B. Ptitsyn, in *Protein Folding*, edited by T. E. Creighton (W. H. Freeman and Company, New York, 1992), Chap. 6, pp. 243–300.
[32] B. Derrida, Phys. Rev. B **24**, 2613 (1981).
[33] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. USA **89**, 8721 (1992).
[34] K. Kuwajima, Proteins **6**, 87 (1989).
[35] K. Kuwajima, Current Opinions in Biotechnology **3**, 462 (1992).
[36] R. L. Baldwin, Current Opinions in Structural Biology **3**, 84 (1993).
[37] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **61**, 2635 (1988).