On the transition coordinate for protein folding

Rose Du[a)]
Department of Physics and Center for Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Vijay S. Pande
Department of Physics and Center for Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge,Massachusetts 02139 and Department of Physics, University of California at Berkeley, Berkeley, California 94720

Alexander Yu. Grosberg[b)] and Toyoichi Tanaka
Department of Physics and Center for Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Eugene S. Shakhnovich
Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138

(Received 16 June 1997; accepted 24 September 1997)

To understand the kinetics of protein folding, we introduce the concept of a ''transition coordinate'' which is defined to be the coordinate along which the system progresses most slowly. As a practical implementation of this concept, we define the transmission coefficient for any conformation to be the probability for a chain with the given conformation to fold before it unfolds. Since the transmission coefficient can serve as the best possible measure of kinetic distance for a system, we present two methods by which we can determine how closely any parameter of the system approximates the transmission coefficient. As we determine that the transmission coefficient for a short-chain heteropolymer system is dominated by entropic factors, we have chosen to illustrate the methods mentioned by applying them to geometrical properties of the system such as the number of native contacts and the looplength distribution. We find that these coordinates are not good approximations of the transmission coefficient and therefore, cannot adequately describe the kinetics of protein folding. © *1998 American Institute of Physics*. [S0021-9606(98)50701-5]

I. INTRODUCTION

In recent years, the thermodynamics of protein folding has been widely studied and is now much better understood.[1–8] The random energy model (REM) has been employed to describe heteropolymer freezing with much success.[3,6,8–10] However, there are some systems that do not satisfy REM assumptions.[11] Because these systems move from a phase in which the conformations have little or no overlap (unfolded state) to a phase dominated by one conformation, violation of REM assumptions does not lead to significant changes in the thermodynamics of the system. However, the kinetic aspects of the problem depend very much on how the system moves through conformational space, making the statistical dependence of states a major issue. The dependence of the kinetics on various aspects of the conformational space remains unsolved. In this article, we will examine the methods which we can use to gain further insight into what these aspects are and how they relate to the kinetics.

From the thermodynamics, we know that under appropriate conditions, the heteropolymer system has two distinct, thermodynamically stable states: the folded state and the unfolded state.[6,12–14] The heteropolymer system therefore bears kinetic resemblance to two very different systems, namely, chemical reactions and systems with first order phase transitions (as in vapor-liquid condensation). While protein folding, chemical reactions, and vapor-liquid phase transitions all involve two thermodynamically stable states, they occur on vastly different length scales. Chemical reactions occur on the molecular level, involving molecules which are small relative to proteins and vapor-liquid phase transitions occur on a macroscopic level, involving visible vapor droplets. Thus the heteropolymer system is intermediate between these two extreme length scales.

Chemical kinetics is often described using transition state theory.[15] According to this theory, in going from one valley to the other, the system will closely follow the minimum energy path. The trajectory of this minimum energy path is known as the reaction coordinate. Thus in a chemical reaction, the system will progress along the reaction coordinate. Because all motions orthogonal to the reaction coordinate lead to higher energy states, the system typically oscillates rapidly in these directions as it proceeds along the reaction coordinate. A typical free energy profile consists of two minima corresponding to the two stable states and a maximum which corresponds to a state known as the transition state (TS). The transition state is not only the maximum energy state along the minimum energy path, but also the minimum energy state with respect to coordinates orthogonal to the reaction coordinate. Thus the transition state resides on

[a)]Electronic mail: rdu@mit.edu
[b)]On leave from the Institute of Chemical Physics, Russian Academy of Sciences, Moscow 117977, Russia.

 0021-9606/98/108(1)/334/17/$15.00

a saddle point and from there, the system progresses along the steepest descent paths.

On the other hand, in systems involving first order transitions, the dynamics are often dominated by a few relevant variables. In such theories, these variables play the role of order parameters, while the other variables, which are less dominant, contribute to the nonlinearities, dissipative effects, and noise in the equations for the dominant variables. In nucleation theory, for example, the relevant variable is the nucleation radius.[16,17] The free energy of a spherical droplet is determined by the balance between the decrease in bulk energy due to the increase in volume and the increase in surface tension due to the increase in surface area. The critical radius is the radius at which these two effects cancel each other. When the nucleation radius reaches the critical size, the nucleus rapidly increases in size and the system condenses. Rapid nonspherical fluctuations of the nucleus occur as its nucleation radius changes and are similar to the movements orthogonal to the reaction coordinate mentioned above.

Analogous to the reaction coordinate in chemical kinetics and the ''order parameter'' in the dynamics of phase transitions, we define the transition coordinate to be the coordinate along which the system progresses most slowly. By this, we mean the coordinate with the largest relaxation time, τ_{slowest}. A good transition coordinate is one along which the system relaxes much more slowly than any other coordinate, that is, $\tau_{\text{slowest}} \gg \tau_{\text{other}}$. When such a coordinate exists, motions along all other coordinates (orthogonal to the transition coordinate) progress much more quickly. In other words, the system equilibrates much more rapidly with respect to these coordinates than it does with respect to the transition coordinate. So in the time it takes for the system to move slightly along the transition coordinate, it would have equilibrated in all other coordinates. The time it takes for the system to proceed from one stable state to the other is therefore on the order of τ_{slowest}. This implies that distance along the transition coordinate determines how far one state is from another kinetically, whereas distances along the other coordinates do not. Therefore, if we let p be the transition coordinate, the kinetically relevant landscape is the free energy as a function of the transition coordinate

$$F(p) = -T \ln \int e^{-E(p,x_1,\ldots,x_n)} dx_1 \ldots dx_n , \tag{1}$$

where x_i is a coordinate of the system and $E(p,x_1,\ldots x_n)$ is the energy of the system. In $F(p)$, all other degrees of freedom have been integrated over or annealed, leaving p, the transition coordinate, as the only quenched parameter. This results in a reduction of the dimensionality of the free energy phase space and hence a great simplification of the problem.

Because the heteropolymer system consists of two stable states, the free energy profile will typically consist of at least two minima, of which the two lowest correspond to the two stable states, and at least one maximum. At this point, the transition coordinate may seem to be very similar to the reaction coordinate. However, while the reaction coordinate as used in chemical kinetics is also a transition coordinate, the transition coordinate is a much more general concept. No assumption is made concerning the existence of a path in the multidimensional phase space of the system. In particular, the transition coordinate is not necessarily associated with any path in the phase space (Fig. 1). Indeed, there can be many paths going from one state to the other while the kinetics itself is described by one parameter.

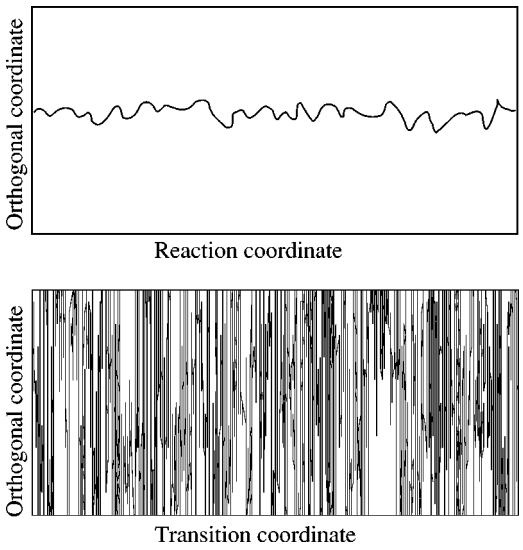

FIG. 1. Schematic of the transition coordinate vs the reaction coordinate. (a) The reaction coordinate as defined in chemical kinetics is the trajectory of the path along which the system progresses. Small fluctuations about the orthogonal coordinates can occur while the system travels down the reaction coordinate. (b) The transition coordinate generalizes the concept of a reaction coordinate in that it is not necessarily associated with a path in phase space. The main defining characteristic of a transition coordinate is that it is the slowest coordinate of the system. Thus the system rapidly equilibrates in the orthogonal direction as it moves slowly along the transition coordinate. Note that this movement is not necessarily small as in the case of the reaction coordinate.

To translate the abstract concept as given above to a specific computable quantity, we define the transmission coefficient, p, for any given microstate in a system dominated by two stable macrostates A and B, to be the probability of arriving at A before arriving at B.[18] For any system with two sufficiently stable states, that is, one whose equilibrium distribution of states is sufficiently bimodal, we can, in principle, always obtain p for any microstate, either computationally or experimentally. For the heteropolymer system, we let A be the folded state and B be the unfolded state. Thus p is the probability of folding before unfolding. From here on, we will discuss p in the context of the heteropolymer system.

To determine p computationally for any given conformation in the heteropolymer system, we can perform a large number of folding simulations starting at that conformation and determine the probability of folding before unfolding. This procedure will be described in detail in the following sections. By varying the number of folding simulations used to determine p, we can compute p to the desired degree of accuracy.

Aside from being a computable and measurable quantity, p is useful because it has many of the fundamental properties which we attribute to the transition coordinate. First of all, p

has two distinct values for the folded and unfolded states, namely, $p=1$ and $p=0$. More important, p is a measure of the kinetic distance between a given conformation and the folded (or unfolded) state. This follows from the definition of p. Clearly, any state with $p>0.5$ is more likely to fold first than to unfold first and is therefore kinetically closer to the folded state. A similar argument holds for $p<0.5$.

The case $p=0.5$ merits some special attention. Such a state is as likely to fold as it is to unfold. It is therefore reasonable to define the ensemble of conformations with $p=0.5$ as the transition state ensemble. This ability to measure kinetic distance makes p an attractive quantity independent of the transition coordinate. In addition to elucidating the nature of protein folding kinetics through measuring kinetic distance, the transmission coefficient allows us to easily identify the transition state. This would aid in protein folding simulations as folding from the transition state is very rapid and could therefore be simulated in present large scale molecular dynamics (MD) simulations.

Although we can, in principle, use p as the transition coordinate, it has a number of drawbacks. p contains within it geometrical and energetic information about the system which is not easily discernible. Thus it is difficult to use it to characterize the system in a physically transparent manner. In addition, it is possible that the transmission coefficient can change significantly from conformation to conformation connected by a single Monte Carlo move, making it difficult to follow the details of the kinetics through Monte Carlo simulations. Furthermore, p is computationally time consuming to calculate and technically difficult, if not impossible, to measure experimentally.

In general, there may or may not be one coordinate that is much slower than all other coordinates. If there is no such coordinate, then the system does not have a good single transition coordinate. Instead, the multiple transition coordinates which are slow relative to other coordinates constitute a multidimensional space whose topological complexity increases rapidly with the number of dimensions involved. In such cases, specification of all the components of the transition coordinate of the system is necessary for specifying its kinetic distance from the (un)folded state. In addition, the transmission coefficient no longer depends on the transition coordinates in a simple way. Even so, the transmission coefficient will still serve as the best possible measure of kinetic distance.

In this article, we will concentrate on the existence of a single transition coordinate. If there is a single transition coordinate for the system, then there is a bijective mapping between p and the transition coordinate. To see this, first recall that the folded and unfolded states have $p=1$ and $p=0$, respectively. Now consider two states along the transition coordinate. The state which is closer to the folded state along the transition coordinate is also kinetically closer to the folded state. Since p also measures kinetic distance, its transmission coefficient is closer to 1. Thus p increases monotonically as we move along the transition coordinate and for every point on the transition coordinate, we can associate a p. Furthermore, any two states with different p's must belong to different points along the transition coordinate because their kinetic distances from the folded state are different. We can therefore conclude that there is a bijection between the transmission coefficient and the transition coordinate. Thus if the transition coordinate exists, it is well-represented by p. This means that a heteropolymer which progresses from $p=0$ to $p=1$ will pass through all intermediate values of p just as it would pass through all values of the transition coordinate. In addition, any dependence of the transition coordinate on physical parameters such as temperature will be reflected in p as well.

Now consider the case in which the free energy profile along the (single) transition coordinate is dominated by a single peak with two minima corresponding to the stable states. If we have an ensemble of folded and unfolded states that are in equilibrium, then at the peak there is an equal number of states going to the folded state as to the unfolded state. Thus the probability of forming the folded state must be $p=0.5$. The transition state ensemble therefore occurs at the peak of the free energy profile. Thus our definition of the transition state in terms of p is consistent with the definition of the transition state as given in chemical kinetics with the reaction coordinate taken to be the transition coordinate. Unlike chemical kinetics, however, we have an ensemble of transition states as opposed to one transition state since many different conformations can have the same p.

We must keep in mind that the concept of the transition state residing at the peak of the free energy profile only holds for this specific scenario of a single dominant free energy barrier. Indeed, in chemical kinetics, a more complicated free energy profile (e.g. with multiple saddle points or trajectories) renders the concept of a transition state inapplicable.[19] However, because of the greater generality of the transition coordinate, the transition states remain defined ($p=0.5$) regardless of the complexity of the free energy landscape. In such cases, it is no longer necessary for the transition state to coincide with the maximum of the free energy profile.

Because of p's computability for short-chain polymers on a lattice, and its resemblance of the transition coordinate, we will use p as a guideline, and proceed to examine the physical properties of p and determine whether various physical parameters can closely approximate p. As we will utilize computer simulations to do so, our results are experimental in nature and reflect the actual properties of the system.

II. COMPUTATION OF THE TRANSMISSION COEFFICIENT

We model the heteropolymer by a self-avoiding chain on a cubic lattice. The energy of a conformation is given by $E=\Sigma_{I<J}B_{S_IS_J}\Delta(r_I-r_J)$ where B_{ij} is the interaction energy between species i and j, $\Delta(r_I-r_J)=1$ for nearest neighbor contacts (but not neighbors along a chain) on the lattice and $\Delta(r_I-r_J)=0$ otherwise. We have used several interaction energy models including the Ising model, the hydrophobic model, the independent interactions model, and the Go model (Table I).

TABLE I. Different models of interactions used in determining the kinetic behavior of polymer systems.

Model name	Matrix
Hydrophobic model (HP)[a]	$\begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}$
''Ferromagnetic'' Ising model[b]	$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$
Black and white model ($BW_{\theta=0.13}$)[c]	$\begin{pmatrix} 1.27 & 0.13 \\ 0.13 & 1.27 \end{pmatrix}$
Independent interactions model (IIM)[d]	independent random energies
Go model[e]	$B_{ij}=\begin{cases} -1 & \text{if } i \text{ and } j \text{ are nearest neighbors in native state} \\ 0 & \text{otherwise} \end{cases}$
Miyazawa and Jernigan model (MJ)[e]	''realistic'' protein interactions

[a]Reference 20.
[b]Reference 21.
[c]Reference 22.
[d]Reference 23.
[e]Reference 24.
[f]Reference 25.

The folding is done by Monte Carlo simulations and includes corner flips, end flips, crankshafts, and null moves.[26] Before calculating the transmission coefficient, we first perform a very long (''equilibrium'') Monte Carlo simulation with 10^9 time steps to determine the temperature range in which the equilibrium distribution as a function of the number of native contacts, Q, is sufficiently bimodal. For these cases, there are two maxima in the distribution, one at maximum Q and the other at Q_u. It is therefore reasonable to define the ensemble of unfolded states to be all conformations with $Q \leqslant Q_u$. In practice, the transmission coefficient does not change significantly with respect to small deviations in the definition of the unfolded state so that we can increase the upper limit, Q_u, slightly in order to speed up the calculations. To computationally calculate the transmission coefficient for any given conformation, we proceed as follows. We start with the given conformation and allow it to fold (or unfold). When a conformation either folds or unfolds, we record its status, start over from the same initial conformation and repeat the process. This procedure is similar to that used in determining first passage time to fold or unfold. The transmission coefficient is computed from the series of independent runs by dividing the number of runs in which the polymer folds first by the total number of runs.

To determine the error in our computation of the transmission coefficient, we calculated p for some conformations by varying the parameters (i.e. the seeds). We found that the variance in the transmission coefficient varies approximately as $1/\sqrt{N}$, where N is the number of runs, as it should be.[27] Thus to obtain an error of within 5%, we did at least 400 folding simulations for each conformation to obtain p.

III. ENERGETIC VERSUS ENTROPIC PROPERTIES OF THE TRANSMISSION COEFFICIENT

We will begin by examining the various physical properties of the transmission coefficient for short chain heteropolymers on a lattice. We can divide these properties into two major categories: energetic and entropic. By examining the correlation between p for systems with sufficiently different interactions, we can determine whether the dominant contribution to p is energetic or entropic. If p is determined primarily by entropic factors, then there should be a strong correlation between the transmission coefficients of a given conformation under two different systems even if various energetic factors, such as the sequence or interaction matrix, are considerably different in the two systems. To address this question computationally, we have calculated the transmission coefficient for a large number of conformations. In particular, we have exhaustively enumerated all 18-mer conformations with $Q>8$ where Q is the number of native contacts and have calculated p for each conformation. For an 18-mer, Q ranges from 0 to 16 but we only examine those with $Q>8$ because (1) it is impractical to exhaustively enumerate all conformations of an 18-mer and (2) there is a sufficient number of conformations with $Q>8$ to give us a good idea of the properties we are examining.

To determine the entropic contribution, we compared the transmission coefficient for 18-mers under the Go model and a well-designed independent interactions model (IIM).[23,24] In the independent interactions model, interaction between each pair of monomers is chosen from a Gaussian distribution. Thus the number of monomer species is equal to the number of monomers involved. For a given target state, the polymer is designed by pulling down the target state energy through the process of simulated sequence annealing[28] so that the target conformation is the unique minimum energy conformation. A sequence designed in this manner folds much faster and more reliably to the native conformation than a random sequence. In the Go model, the only good interactions are those between nearest neighbors in the target conformation. All other monomers do not interact. In particular, the interaction between two monomers I and J is $B_{IJ}=-\Delta(r_I-r_J)$ where $\Delta(r_I-r_J)=1$ if I and J are nearest neighbors (but not along the chain) in the target conformation and $\Delta(r_I-r_J)=0$ otherwise. Thus the Go model is well-designed by definition.

The degree to which the Go model and the IIM are correlated can be determined by the connected correlation function[29]

$$C(B,B')=\frac{\Sigma_{ij}p_i\delta B'_{ij}\delta B_{ij}p_j}{\sqrt{(\Sigma_{ij}p_i\delta B^2_{ij}p_j)(\Sigma_{ij}p_i\delta B'^2_{ij}p_j)}}, \tag{2}$$

where p_i is the probability of occurrence of species i, B_{ij} is the interaction energy between species i and j, $\overline{B}=\Sigma_{ij}p_iB_{ij}p_j$, and $\delta B_{ij}=B_{ij}-\overline{B}$. $C(B,B')=1,0,-1$ when B and B' are completely correlated, completely uncorrelated, and completely anticorrelated, respectively. The mean correlation between an interaction matrix and a set of random matrices is thus 0. The correlation between the Go model and

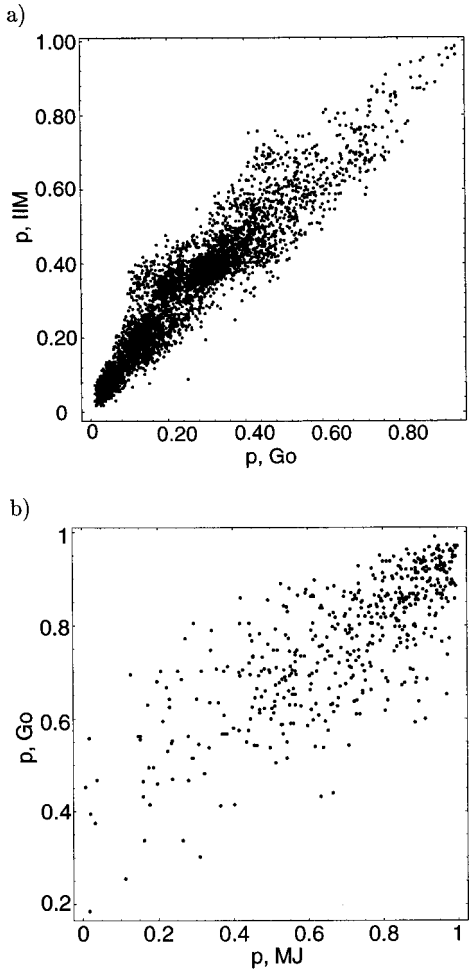

FIG. 2. Correlation of the transmission coefficient, p, between different interactions. (a) Comparison between the Go model ($T=0.6$) and a well-designed IIM ($T=0.75$) for an 18-mer. The transmission coefficient is calculated for 5142 conformations at $Q=9$. The strong correlation suggests the importance of entropic factors in the transmission coefficient. (b) Comparison between the Go model ($T=0.8$) and the MJ model ($T=0.16$) for a well-designed 48-mer. The transmission coefficient is calculated for 500 conformations with $Q=19$ and $K=37$ where Q is the number of native contacts and K is the total number of contacts. There is still a correlation between the transmission coefficients under the different interactions for the 48-mers suggesting that entropic factors still dominate. However, the scattering of points suggests that energetic factors also play a role.

the well-designed IIM that we are considering is 0.456, therefore they are weakly correlated. Note that they must be somewhat correlated since they both lead to the same native conformation.

When we compare p of 5142 conformations with $Q=9$ for the well-designed IIM with that of the Go model, we see a strong correlation with correlation coefficient=0.94, which suggests that interactions may be less important than entropic factors (Fig. 2). Due to the considerable difference between the IIM and Go models, if entropic factors did not contribute significantly to the transmission coefficient, we would not expect much correlation between the p's in these two systems.

However, we cannot truly compare the magnitudes of the correlation between the interaction matrices and the correlation between the p's since there is no clear relationship between them. To support the above conclusions, we examine the correlations of other interactions. In particular, we examine the HP, Ising, and $BW_{\theta=0.13}$ models (see Table I). For each interaction, the design was done for the same native conformation as for the Go model and the IIM described above. Because the design results in different sequences for each of the two-letter interaction matrices, the correlation between the matrices is obtained by expanding the 2×2 matrices into 18×18 matrices such that each monomer is treated as a different species. The 18×18 matrices can then be compared.

TABLE II. Correlations between different interaction matrices (see Table I) and the corresponding transmission coefficients. The correlations between the matrices are given in the upper triangle and the correlations between the transmission coefficients are given in the lower triangle. The correlations between the interaction matrices given here are calculated for expanded matrices (see the text).

	AF	$BW_{\theta=0.13}$	HP
AF	1	0.04	0.18
BW	0.74	1	0.88
HP	0.81	0.98	1

The correlations between the transmission coefficients for 91, $Q=12$, 18-mer conformations under each of the interactions are higher than the correlations between the interaction matrices themselves (see Table II). While we cannot draw any definite conclusion when the correlations between the interaction matrices themselves are fairly high, a very strong correlation between the transmission coefficients coupled with a very weak correlation between the interaction matrices strongly suggests that entropic factors are indeed dominant.

To confirm the generality of the above results, we compared p for 48-mer conformations under the Go and MJ models. The MJ model is an approximation of realistic protein interactions obtained by considering only interactions between amino acids. There are thus 20 species in this model corresponding to the 20 amino acids. The correlation between the Go and the well-designed MJ model as defined by Eq. (2) is 0.256, thus they are weakly correlated.

For the purpose of this comparison, 500 48-mers with $Q=19$ and $K=37$, where Q is the number of native contacts and K is the total number of contacts, including nonnative ones, were chosen as follows. We start with a polymer in the native conformation and allow it to unfold at a high temperature ($T=10^9$) under homopolymeric interactions. When the homopolymer goes through a conformation with the desired Q and K ($Q=19$ and $K=37$), we save that conformation in our database and repeat the process from the beginning with the native conformation. Thus 500 unfolding runs were done to obtain the 500 $Q=19$ and $K=37$ conformations. While this procedure is efficient for obtaining 48-mer conformations, it has a slight drawback. Conformations generated using this procedure may not be completely random as there are many conformations that cannot be reached by homopolymer unfolding. In particular, such conformations may

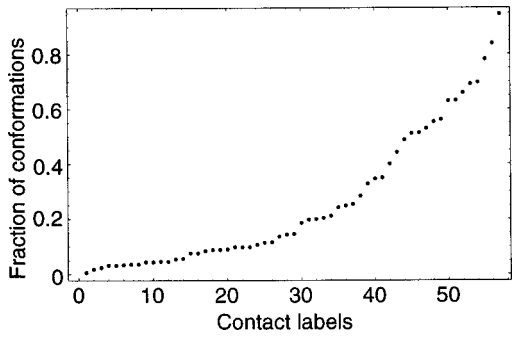

FIG. 3. Frequency of occurrence of contacts in 500 48-mers with $Q=15$ and $K=18$ that are generated through a homopolymer unfolding procedure. Note that 7 out of the 57 possible contacts occur in >0.6 of the conformations with 2 contacts occurring in >0.8 of the conformations. The high frequency of occurrence of these contacts is due to the bias inherent in the unfolding procedure used to generate them.

not have diverged sufficiently from the native conformation. Indeed, certain contacts appear in a very high fraction of conformations even when only approximately half the native contacts are left (Fig. 3) which is indicative of the bias present. However, we will see below that the possible nonrandomness of the 48-mer conformations does not affect our conclusions significantly.

As in the case of 18-mers, we find that there is a stronger correlation (correlation coefficient=0.77) between the transmission coefficients under different interactions than between the interaction matrices themselves (Fig. 2). The increased scattering and hence decrease in correlation in this case indicates that while entropic factors dominate, energetic factors also play a role. Note that the inclusion of conformations that cannot be reached by homopolymer unfolding will increase the dominance of entropic factors since these conformations are likely to have low transmission coefficient irrespective of the interactions involved. Our conclusion regarding the dominance of the entropic factors will thus remain unchanged.

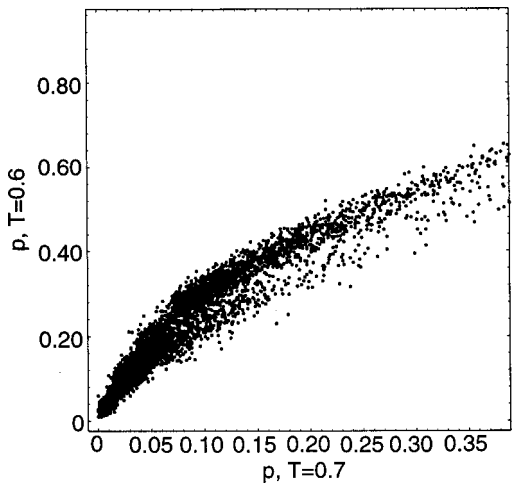

FIG. 4. The correlation between the transmission coefficients at different temperatures. The transmission coefficients of 5142 18-mers with $Q=9$ under the Go interactions at $T=0.6$ and $T=0.7$ have a correlation coefficient of 0.94. The states appear to shift in kinetic distance from the folded state in response to environmental factors.

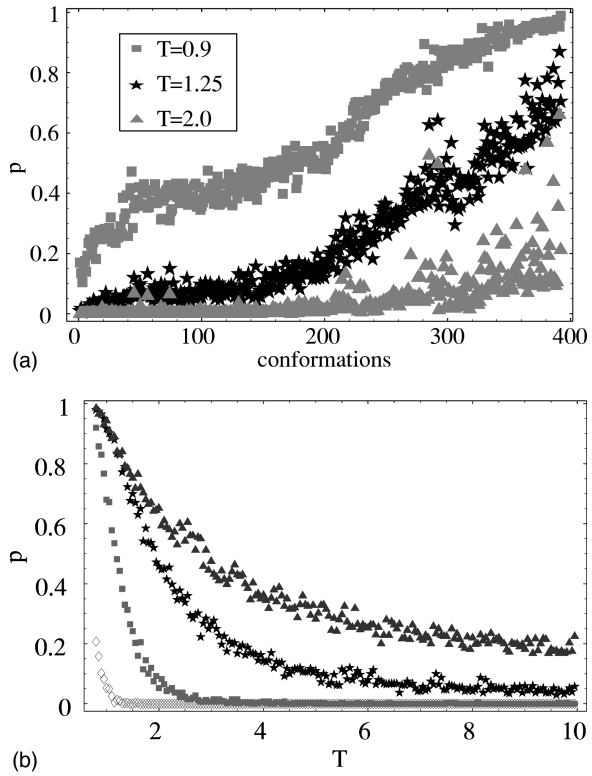

FIG. 5. The dependence of the transmission coefficient, p, on temperature. (a) The transmission coefficients for 392 18-mer conformations with $Q=11$ under the IIM at $T=0.9$, 1.25, and 2.0 tend to decrease with temperature. While the rate of decrease is different for each conformation, the relative magnitudes of the transmission coefficients remain generally the same as the temperature varies. (b) The transmission coefficients for 4 18-mer conformations with $Q=11$ under the IIM is shown to decrease as a function of T but at different rates.

In Fig. 4, we compare the transmission coefficient for 5142 18-mers at $Q=9$ using the Go model at two different simulation temperatures. We see that the transmission coefficients generally shift over as the temperature increases. In addition, we obtained the transmission coefficients for 392 18-mers at $Q=11$ using the IIM at various temperatures (see Fig. 5). We see that p decreases with increasing temperature. The correlation coefficients between the transmission coefficients at different temperatures is 0.94 in the Go case, and given in Table III in the IIM case. The decrease in correlation between transmission coefficients with an increase in temperature difference is due to the different rates of change in p as a function of temperature for different conformations. To understand what the changes in p due to changes in temperature means, note that changing the temperature is equivalent to changing the variance of the interaction matrix given that the mean interaction is 0 which is the case here. This is because the behavior of the system is determined by the factor E/T where E is the energy and T is the temperature. Thus rescaling the temperature is equivalent to rescaling the energy, which is, in turn, equivalent to rescaling the variance of the interaction matrix when the mean interaction

TABLE III. Correlation coefficients between the transmission coefficients of 392 18-mers with $Q=11$ under the IIM (see Table I) at different temperatures. The correlations decrease as the temperature difference increases because of the different rates of decrease of the transmission coefficients as a function of temperature.

T	0.90	1.00	1.25	1.50	1.75	2.00
0.90	1	0.989	0.942	0.850	0.735	0.627
1.00		1	0.965	0.885	0.775	0.668
1.25			1	0.965	0.889	0.796
1.50				1	0.968	0.909
1.75					1	0.975
2.00						1

is 0. Thus the shifting of states due to temperature is equivalent to that due to a change in the energy scale. Figs. 4 and 5 thus imply that while the magnitudes of the transmission coefficients are affected by the variance of the interaction matrix, their relative magnitudes (in terms of ordering) generally remain the same, emphasizing again the importance of entropy. An interesting observation is that, while we would expect environmental factors such as temperature to shift the states which are in the transition state ensemble,[14] it is unusual that the states shift over as if the transition state ensemble is a window which moves depending on various parameters.

In addition to the correlation in p, we can examine the mean and variance of the energies of conformations with a given p value (Fig. 6). We find that, except for the rise in energy for very low p, there is no correlation between the energy of a conformation and the transmission coefficient of the same conformation.

Thus, we see that the transmission coefficient and the nature of the transition state do not depend very strongly on the energy. As there are only two aspects to a system, energy and entropy, we conclude that p and the nature of the transition state depend on entropic aspects much more than on how one simulates the energy. This suggests, for example, that a simulation which models the entropic aspects well but uses a simplified potential for interactions (i.e. even the very simplified Go potential) should yield insight into the nature of the transmission coefficient. Note that for a polymer system, entropy arises from the different geometries of the various conformations, hence entropy and geometry are equivalent concepts. Thus in studying the entropic aspects of the system, we can equivalently study the geometrical aspects.

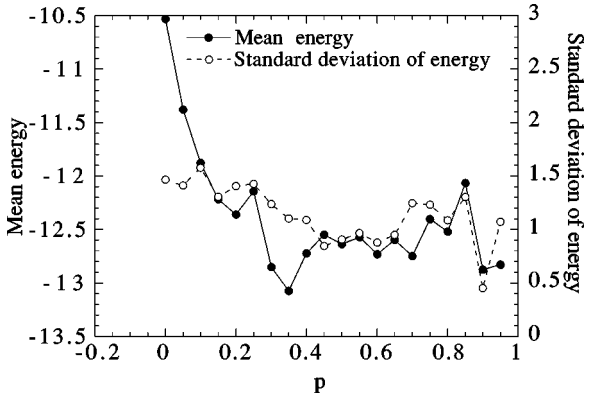

FIG. 6. Mean and standard deviations of the energy vs the transmission coefficient for 18-mers under the IIM ($T=0.75$). There is very little, if any, correlation between p and the mean energy indicating that energetic factors are not dominant factors in determining the transmission coefficient. Note that the width of the distribution at each p is approximately the same as the range of the mean energy.

IV. GEOMETRICAL ORDER PARAMETERS AS APPROXIMATIONS OF THE TRANSMISSION COEFFICIENT

Since we have shown in Section III that the major contribution to the transmission coefficient comes from entropic factors, it is natural to consider various geometrical factors as approximations of the transmission coefficient. To this end, we introduce two methods by which we can determine the degree to which any parameter approximates the transmission coefficient. The first method involves exhaustive enumeration of polymer conformations and the second method involves trajectories of the system in phase space. The conformations which are examined and for which the transmission coefficients are calculated are determined by the method used. In the first case, the conformations are chosen randomly at each Q, while in the second case, the conformations are determined by the Monte Carlo trajectory.

To illustrate these two methods, we will consider the two most basic geometrical properties of a polymer on a lattice, namely, the number of native contacts (nearest neighbor contacts in common with the target conformation), Q, and the distribution of looplengths. By looplength, we mean the distance along the polymer between monomers that are in contact.

A. Method 1. Exhaustive enumeration

We first consider the number of native contacts, Q, as a putative approximation of the transmission coefficient. We exhaustively enumerated all 18-mer conformations with $Q>9$. In addition, we chose 250 random conformations at each $Q\leqslant 9$ since they are too numerous to completely enumerate (see Table IV).

For each of the conformations, we did 1000 Monte Carlo runs to calculate the transmission coefficient using the ferromagnetic Ising model (see Table I). For $0<Q\leqslant 9$, we have extrapolated the data from the 250 conformations at each Q to all conformations with the corresponding Q by assuming

TABLE IV. Number of conformations for each Q for two different target conformations. The total number of conformations is 9 872 284 420. Target I was used with the Ising model and target II was used with the IIM model. A short description of each model is given in Table I.

Q	Total No. of conformations for target I	Total No. of conformations for target II
16	1	1
15	0	0
14	4	6
13	14	5
12	47	91
11	269	392
10	1042	2059
9	7232	15 488
8	48 799	81 790
7	353 551	546 771
6	2 471 410	2 798 567
5	15 684 249	13 942 023
4	88 326 150	67 282 698
3	416 746 831	332 803 192
2	1 489 412 159	1 208 553 781
1	3 561 699 968	3 182 091 765
0	4 297 532 694	5 064 165 791

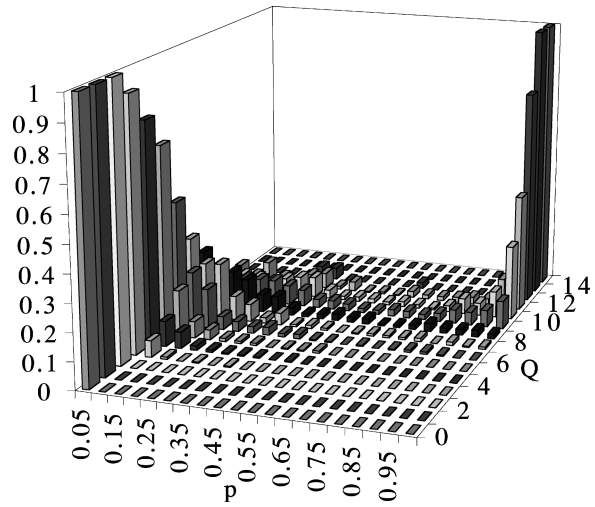

FIG. 7. Distribution of 18-mer conformations under the Ising interactions at $T=0.7$ (see Table I) over p and Q. The distribution gives the probability for a state with a given Q to have a particular transmission coefficient. It is normalized so that the sum of the probabilities over p at each Q is equal to 1. The broadness of the distribution at intermediate Q's indicates that Q is not adequate in describing the kinetics of the system. Note that this broadness is an intrinsic characteristic of the system which does not decrease with an increase in accuracy in the calculation of p. Here, we used 1000 Monte Carlo runs to calculate p for each conformation which gives an error of approximately 3% for p of any particular conformation.

that the 250 random conformations are representative of all other conformations at that Q.

In addition to 18-mers, we also examined 48-mer conformations which were chosen as described in Section III. Recall that these conformations are slightly biased because they are generated through a homopolymer unfolding algorithm. However, we will see below that this bias does not affect our conclusions. The sequence used was designed to be fast folding under the MJ interactions at $T=0.16$.[30] The transmission coefficient for each of these conformations was calculated using 1000 Monte Carlo runs and the MJ model (see Table I).

We first consider the distribution of 18-mer conformations over p and Q which is normalized in p (so that the sum over the probability of obtaining p for all p at a given Q equals 1). If Q were a good approximation of the transmission coefficient, the distribution of p for any particular Q should closely approximate a Dirac delta function and p should be strongly correlated with Q. However, we find that, even though there is a general increase in p as Q is increased, the transmission coefficient distribution is broad and multimodal for intermediate Q's (Fig. 7). In addition, the correlation between p and Q as defined by

$$\langle pQ\rangle_c=\frac{(\langle pQ\rangle-\langle p\rangle\langle Q\rangle)}{\sqrt{(\langle p^2\rangle-\langle p\rangle^2)(\langle Q^2\rangle-\langle Q\rangle^2)}} \tag{3}$$

is extremely low with $\langle pQ\rangle_c=0.22$.

Similar results were obtained for 48-mers under the MJ interactions (Fig. 8). The mean and standard deviations of the p distribution were calculated for various Q and K. For each pair of Q and K, we have chosen 200 conformations as described earlier. As before, the mean transmission coefficient increases with Q. While the width of the p distribution decreases with Q, it remains extremely broad at low and intermediate Q's (close to 0.5).

The broadness of the p distributions for both the 18-mers and 48-mers is not due to the error in the calculation of p for any particular conformation which decreases as $1/\sqrt{N}$, where N is the number of runs used to calculate p. Here the width of the distribution is a fundamental characteristic of the system which does not change with the number of Monte Carlo runs used to calculate each p. This broadness and multimodality are clearly seen in the transmission coefficient distribution for the 18-mer IIM (see Table I) for which p was

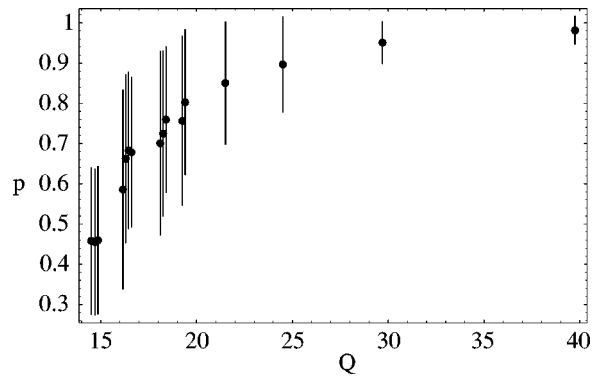

FIG. 8. Transmission coefficient for a 48-mer under the MJ model at $T=0.16$ (see Table I) as a function of Q. Each point represents the mean p for conformations with the same number of native contacts, Q, and the total number of contacts, K. The mean p for conformations with the same Q but different K's is plotted at a slightly different Q so that it can be differentiated. The error bars on each point give the standard deviation of each distribution.

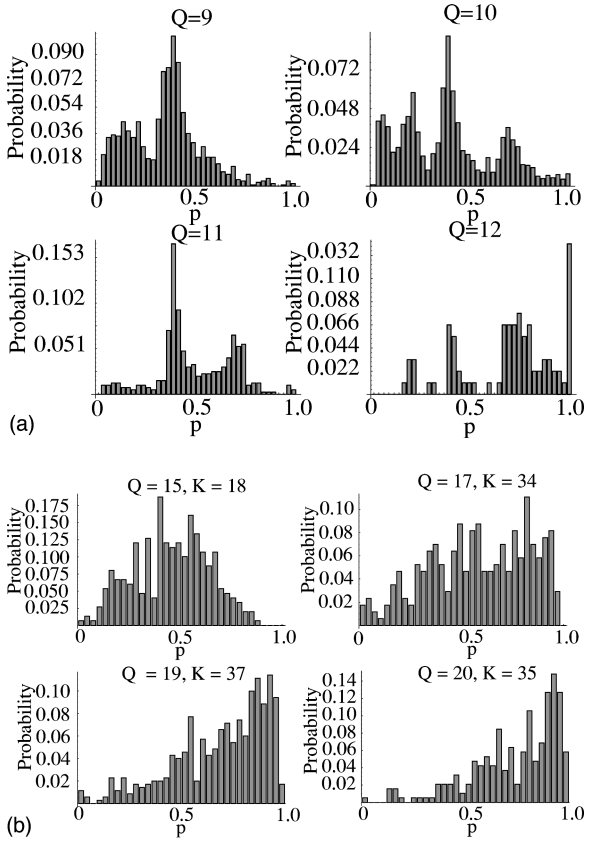

FIG. 9. Distribution of p values for a given Q (indicated above each graph) for 18-mers under the IIM at $T=0.75$ (see Table I) and for 48-mers under the MJ model (see Table I) at $T=0.16$. The transmission coefficient is calculated for conformations with and without nonnative contacts. (a) The number of 18-mer conformations under the IIM for which p was calculated is 1114, 2059, 392, and 91 for $Q=9$, 10, 11, and 12, respectively. These conformations were chosen randomly from all conformations with those Q's. (b) p was calculated for 200 48-mer conformations under the MJ model for each of the following pairs of Q and K: $Q=15$, $K=18$, $Q=17$, $K=34$, $Q=18$, $K=37$, and $Q=20$, $K=35$. For both 18-mers and 48-mers, each p was calculated using 1000 Monte Carlo runs which gives an error of approximately 3% for p of any conformation. The multimodality and/or broadness of the distributions is intrinsic to the system and does not decrease by increasing the number of runs. The broadness of the distributions indicates that Q is not a good approximation of the transmission coefficient.

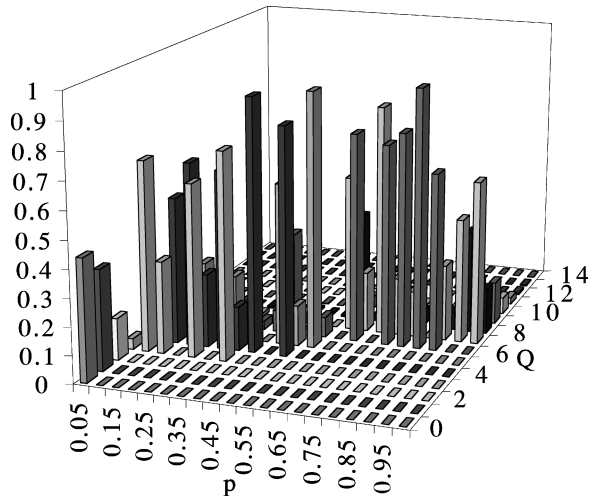

FIG. 10. Distribution of 18-mer conformations under the Ising interactions at $T=0.7$ (see Table I) over p and Q. The distribution gives the probability for a state with a given p to have a particular Q. It is normalized so that the sum of the probabilities over Q at each p is equal to 1. The fairly narrow distribution over Q for a given p implies that we can infer the geometrical properties of a conformation by knowing its kinetic properties.

computed for conformations with $Q>8$ and for the 48-mer MJ model (see Table I) for which p was calculated for various Q's and K's (Fig. 9). A possible explanation for the multimodality is that certain conformations are trapped at low p so that in order for these conformations to fold, they must first unfold to conformations with low Q and refold again into those conformations with the same Q that fold more easily. Note that the possible bias in the 48-mer conformations due to the method by which they were generated (see Section III) can only decrease the broadness of the p distribution. This is because these conformations are at worst biased towards increased similarity with the native conformation and hence with each other. Any attempt to increase the randomness of the 48-mer conformations used will thus further emphasize the broadness in the p distribution that we observed.

Now if we normalize the $p-Q$ distribution for the 18-mers under the Ising interactions over Q instead of over p (Fig. 10), we find that the trend of increasing Q with increasing p is more apparent and that the distribution of p for a given Q is more monomodal but still broad. However, the distribution of conformations with a given p normalized over Q is much narrower. This suggests that, while it is difficult to predict the transmission coefficient of a particular conformation from its Q, we can infer geometrical properties of a conformation (its Q) by knowing how well it folds.

Thus, we conclude that the number of native contacts cannot adequately describe the kinetics of a system. We may suspect that this is due to different roles and the relative importance of various contacts. Indeed, if we consider only those conformations without nonnative contacts, we find that Q is a slightly better approximation of p; the distributions are somewhat more monomodal (Fig. 11). This effect is seen in both 18-mers and 48-mers and is therefore quite general. It is reasonable since nonnative contacts must be eventually broken and thus slow down the kinetics. Even so, the distribution remains broad further emphasizing the inadequacy of Q as a kinetic descriptor. The small peaks which remain in the distribution suggests that some native contacts are more important than others. Moreover, some native contacts may restrict the movement of the polymer such that it is more difficult to form other native contacts. For example, nonlocal contacts (i.e. far along the sequence) may greatly restrict the nature of the polymer conformations as compared to local contacts. We will address this issue when we discuss loop-lengths.

The inadequacy of the number of native contacts, which is a global parameter, in approximating the transmission coefficient, and hence in describing the folding kinetics, suggests that the folding process is highly dependent upon local parameters. To understand the nature of this dependence, we

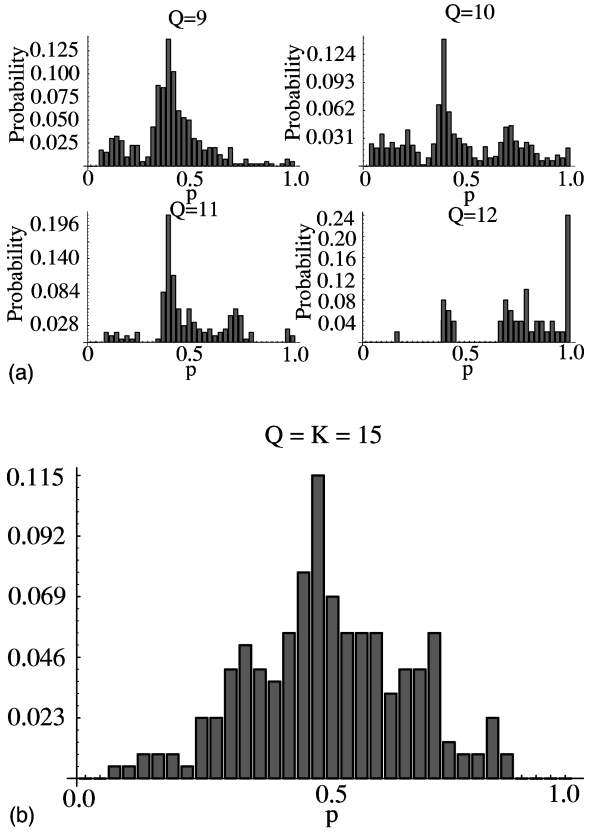

FIG. 11. Distribution of p values for a given Q (indicated above each graph) for 18-mers under the IIM (see Table I) at $T=0.75$ and for 48-mers under the MJ model (see Table I) at $T=0.16$ for conformations without nonnative bonds. (a) The number of 18-mer conformations for which p was calculated is 403, 655, 176, and 50 for $Q=9$, 10, 11, and 12, respectively. These conformations were randomly chosen. (b) p was calculated for 213 48-mer conformations at $Q=15$ generated by homopolymer unfolding. For both 18-mers and 48-mers, each p was calculated using 1000 Monte Carlo runs which gives an error of approximately 3% for p of any conformation. This distribution is somewhat narrower than the distribution for conformations with and without nonnative bonds (see Fig. 9) because nonnative bonds play a role in hindering the kinetics of a polymer. However, the distribution remains broad thereby stressing the inadequacy of Q as a kinetic descriptor.

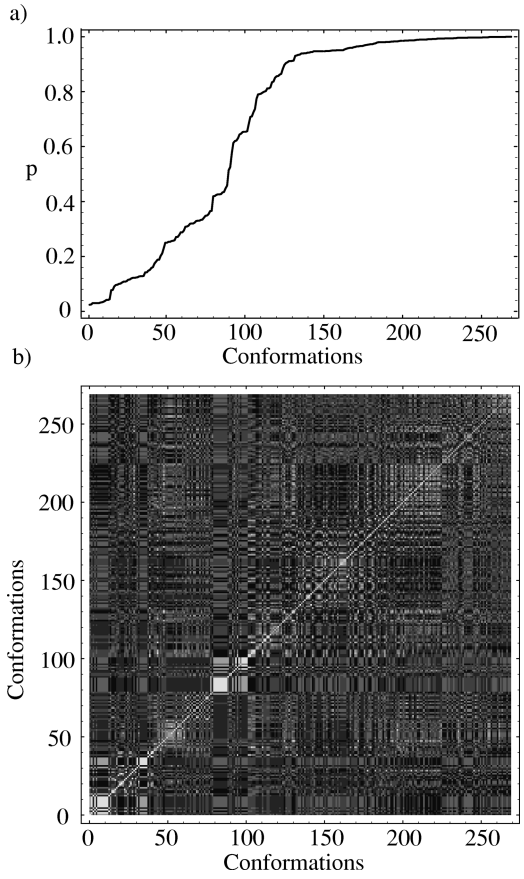

FIG. 12. Comparison between the overlap and the transmission coefficient among all 18-mer conformations with $Q=11$ under the Ising model at $T=0.7$ (see Table I). (a) The transmission coefficient was calculated for all 18-mer conformations with $Q=11$ under the Ising model. These conformations were subsequently sorted and indexed in order of increasing p. Note that the steepness of the slope is related to the number of conformations with a particular p. When the slope is very large, the number of states with that p is very low. Similarly, when the slope is very small, the number of states with that p is very high. (b) The conformations are indexed in order of increasing p as described in (a). The overlap between two conformations is indicated by the gray level such that white indicates $Q_{\alpha\beta}=16$ and black indicates $Q_{\alpha\beta}=0$. If particular contacts are important in determining p, then we would expect the areas of the graph which correspond to conformations with similar p's to be lighter than other areas. For example, we would expect conformations with indices >150 ($p>0.9$) to overlap with each other more than with other conformations which have low p. However, we can see that the overlaps between all pairs of conformations are almost uniform, indicating that particular contacts are not strongly correlated with the transmission coefficient.

utilize our database of exhaustively enumerated conformations and their transmission coefficients to examine the relationship between various local characteristics of the polymer and the transmission coefficient.

To see the existence of any kind of contact similarity among conformations with similar p's, we consider the pair overlap, $Q_{\alpha\beta}$, between all conformations α and β,

$$Q_{\alpha\beta}=\sum_{I\neq J}\Delta(r_I^{\alpha}-r_J^{\alpha})\Delta(r_I^{\beta}-r_J^{\beta}), \tag{4}$$

where $\Delta(r_I-r_J)=1$ if I and J are nearest neighbor contacts, and $\Delta(r_I-r_J)=0$ otherwise. If we look at the overlaps among all conformations at a given Q, we do not see any increase in the overlaps among conformations with similar p's (Fig. 12). This suggests that contacts, even when considered as local parameters between pairs of conformations, cannot adequately describe the folding kinetics either.

At the same level of locality, we consider looplengths as geometrical descriptors of a polymer on a lattice. Recall that looplength is the distance along the polymer between two monomers that are in contact. In general, we might expect monomers that are connected by long looplengths to play a larger role in determining the kinetics of the system. First of all, it may be more difficult to bring these monomers together than it is to bring monomers that are nearby along the polymer together. Second, the wrong contacts formed by these monomers may be more restrictive in terms of preventing correct contacts to form. In the distribution of looplengths over p for conformations with $Q=11$ for the Ising

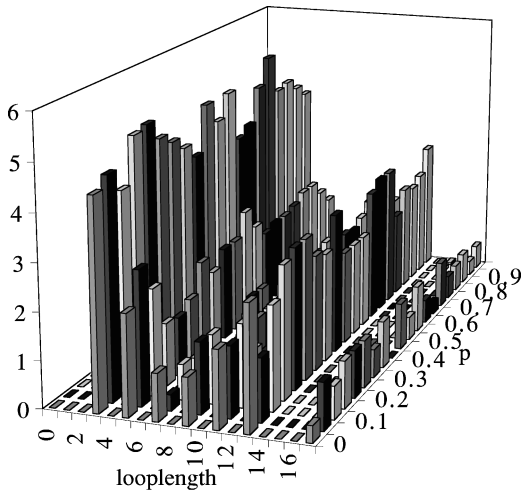

FIG. 13. Distribution of looplengths over p for conformations with $Q=11$ under the Ising interactions (see Table I) at $T=0.7$. The looplength distribution is averaged over all conformations with $Q=11$ at a given p. There is no significant difference between the looplength distributions of conformations with high p vs those with low p implying that looplengths are also poor indicators of the kinetics.

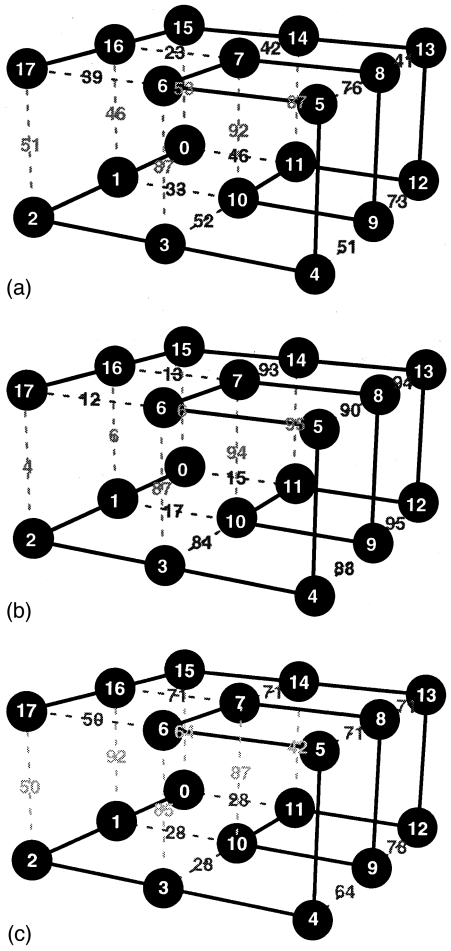

FIG. 14. Probability of finding particular contacts for an 18-mer using the IIM at $T=0.75$ (see Table I). 403 18-mer conformations under the IIM with $Q=9$, no nonnative contacts, and $p=0.05\pm0.05$, 0.50 ± 0.05, and 0.95 ± 0.05 were examined for the frequency of occurrence of each native contact. Solid lines indicate polymeric bonds along the polymer while dashed lines indicate contacts between nearest neighbors in the target conformation. Numbers next to the contacts indicate the percent of conformations examined with the indicated p which has that contact. Conformations with the following transmission coefficients were examined: (a) $p=0.05$, (b) $p=0.50$, (c) $p=0.95$.

model (see Table I), we averaged the looplengths of all conformations with a particular p. We find that there is no significant difference in the looplength distribution of conformations with different p's (Fig. 13). This implies that looplengths are also insufficient as kinetic descriptors.

To see if there exists any similarity at all among conformations with similar transmission coefficients, we now examine the structures of various conformations at an even more microscopic level. In particular, instead of considering the number of common contacts, we now consider the frequency of occurrence of particular contacts at various p's. For 18-mer conformations under the IIM, we find that there is no striking difference among conformations with $p=0.05$, 0.50, and 0.95 (Fig. 14). For 48-mer conformations under the MJ model, however, we do see that certain contacts appear more frequently in conformations with high p ($p>0.8$) than low p ($p<0.2$) and vice versa (Fig. 15). In particular, we find that in 48-mers, conformations with high p's tend to exhibit a set of common bonds near the center of the conformation which has the appearance of a ''nucleus'' (Fig. 16). However, such ''nuclei'' may exist or be increased because the 48-mers were generated using a homopolymer unfolding procedure. Indeed, for each pair of high p conformations, a few of the contacts considered as part of the ''nucleus'' oc-

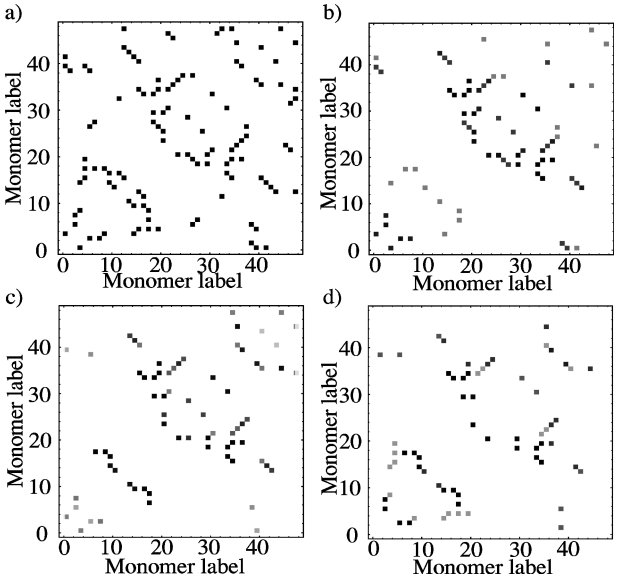

FIG. 15. Probability of occurrence of particular contacts for a 48-mer with $Q=15$ and $K=18$ using the MJ model (see Table I) at $T=0.16$. The probabilities are calculated using a sample of 500 48-mer conformations generated using a homopolymer unfolding procedure. The contact matrix for the native conformation is given in (a). (b) $p<0.2$, (c) $0.45<p<0.55$, and (d) $p>0.8$ were examined for the probability of occurrence of all possible contacts. Each represents a contact matrix. Monomers are labeled according to their positions along the polymer. The darkness of each square represents the probability of occurrence of that contact such that black indicates a probability of 1 and white indicates a probability of 0. The contacts between monomers numbered less than 20 occur much more frequently in high p conformations than low p conformations, while the contacts between monomers 31 and 34 occur much more frequently in low p conformations. There are thus local similarities among conformations with similar p's. However, the existence of particular contacts in a conformation does not always imply that it will have a particular p.

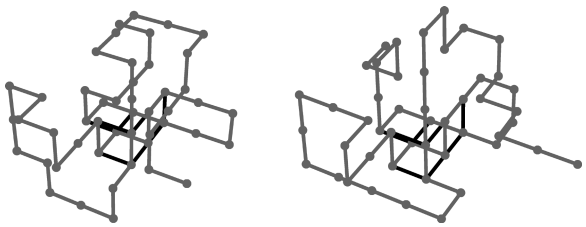

FIG. 16. Common contacts (indicated by dark lines) between conformations with high transmission coefficients tend to be clustered in the center of the conformation. The existence of such contacts usually indicates that the conformation has a high transmission coefficient. However, not all conformations with high transmission coefficients contain such contacts. In addition, a few of the contacts occur in almost all conformations considered, regardless of their transmission coefficient, due to the homopolymer unfolding procedure used to generate them.

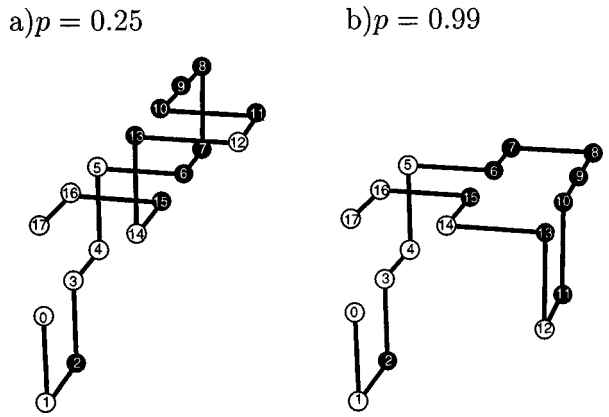

FIG. 17. Two conformations with the same contacts but very different p's ((a) $p=0.25$ and (b) $p=0.99$, Ising model, $T=0.7$). Note that these conformations have two domains that are rotated relative to one another along the axis formed by monomers 7, 6, 15 and 14. The two domains in the conformation with high p are in their native positions while they are rotated relative to one another in the conformation with low p. The difference in p is partially due to the restricted set of Monte Carlo moves which does not allow rotation of entire domains as a whole.

cur not only in high p conformations, but in almost all conformations considered. On the other hand, a majority of the contacts forming the ''nucleus'' rarely occur in conformations with low p. These contacts are the ones between monomers with labels less than 20 (Fig. 15). Furthermore, assuming that the existence of such a ''nucleus'' is not an artifact of the homopolymer unfolding procedure, while the existence of such a ''nucleus'' usually implies that the conformation has high p, not all conformations with high p contain a ''nucleus.'' Thus there are two problems that arise. First, because the existence of a ''nucleus'' is not universal in all conformations with high p, it is not a robust characterization of such conformations. Second, even if all conformations with high p contain such a nucleus, it is an extremely local characteristic that is difficult to characterize by any parameter or transition coordinate.

It may be thought that the inadequacy of contacts as kinetic descriptors is specific to our short-chain lattice model and Monte Carlo move set. In particular, there are some conformations with exactly the same contacts but very different p's (Fig. 17). This situation occurs particularly when the polymer is divided into multiple domains, each of which has good contacts that do not change when we move the domains relative to one another.

The problem of domains is fundamental to the Monte Carlo method. In particular, in order for a conformation with multiple domains to fold using elementary moves from the Monte Carlo move set, the domains must first unfold then refold in order to move together. If there are N monomers in a domain, the time it takes to unfold and refold the domain grows exponentially with N.[31] In reality, however, the amount of time it takes for two domains to move together in a viscous fluid grows as a power of the radius of the domain.[33] Since the radius of a domain increases as $N^{1/3}$, the time for two domains to move towards one another increases as N^{α}, where α depends upon the properties of the fluid. It may be possible to resolve this discrepancy by changing the Monte Carlo move set or by considering off-lattice models. However, pairs of conformations with the same contacts but significantly different p's ($\Delta p > 0.5$) such as the ones in Fig. 17 form only a small fraction of all conformations (see Table V). This suggests that our results are not due to issues specific to our short-chain lattice model and Monte Carlo move set but have much greater generality.

By resorting to exhaustive enumeration, we have shown that the number of native contacts and the looplength distribution are both inadequate in describing the kinetics of the polymer system. For such a polymer system, there is no other (independent) global geometrical descriptor for the system aside from the number of native contacts and the looplength distribution. This implies that small scale characteristics of the system play an important role in determining the

TABLE V. Statistics of 18-mer conformations under the Ising model (see Tables I and IV) at $T=0.7$ with the same contacts and $\Delta p > 0.5$.

Q	Total No. of pairs	No. of pairs with the same contacts	No. of pairs with the same contacts and $\Delta p > 0.5$	Ratio of pairs with the same contacts to all pairs	Ratio of pairs with $\Delta p > 0.5$ to pairs with the same contacts	Ratio of pairs with $\Delta p > 0.5$ to all pairs
14	6	2	0	0.333	0	0
13	91	10	0	0.110	0	0
12	1081	14	2	0.013	0.143	0.002
11	36 046	300	39	0.008	0.130	0.001
10	542 361	1487	269	0.003	0.181	0.000

kinetics of the system. While our results may have been affected by the specific properties of our model such as short chain length, cubic lattice, and particular Monte Carlo move set, we have seen that the same conclusions hold for both 18-mer and 48-mer systems implying that our results are not significantly affected by such issues.

B. Method 2. Trajectories in phase space

The second method by which we can determine whether a parameter closely approximates p involves examining the trajectory of the system during a Monte Carlo run of 10^6 steps as it traverses the phase space with p as one of the coordinates. We will again consider the number of native contacts, Q, as an approximation of the transmission coefficient. We will thus examine the trajectory of the system as it travels across the $p-Q$ plane.

To implement this method, we begin the simulation with a random conformation. At every five Monte Carlo steps, we determine Q of the conformation as well as its p. By doing this, we can see how Q correlates with p during an actual folding or unfolding. In Fig. 18, we have the results for a 27-mer under the Go interactions. The Go model was employed because folding (or unfolding) occurs very rapidly in this model thereby making the calculation of p for a large number of conformations more feasible. The Monte Carlo for the 27-mer was done at $T=2.9$ and $T=2.6$, slightly below the first order transition temperature. These two temperatures were chosen since very long Monte Carlo runs (10^8 steps) at these temperatures lead to sufficiently bimodal distributions (in terms of Q). $T=2.6$ was examined since the folded state is very stable at this temperature so that an unfolded polymer will fold very quickly to its native state allowing fast folding runs. $T=2.9$ was chosen since the distribution is more bimodal at this temperature so that the system will fold and unfold sufficiently many times during a long Monte Carlo run (10^6 steps). We do not consider the 48-mers in detail here since we have seen in the previous sections that the kinetic results are not significantly affected by changes in chain sizes or monomer interactions.

In Fig. 18, we see that Q follows the large scale behavior of p very closely, that is, very high (low) p corresponds to very high (low) Q. This is in agreement with Fig. 7 where the distribution of p over Q is fairly narrow at very high or very low Q. Thus this method is consistent with the method involving exhaustive enumeration described above. While Q follows the large scale behavior of p very well, it does not accurately follow the changes which occur in p on the smaller scale (Fig. 19). In particular, Q does not accurately follow p as it goes through the transition state (Fig. 20). That is, as p changes from a low value ($p<0.5$) to a high value ($p>0.5$), Q remains approximately the same. This behavior is also true for 48-mers under the Go model (see Table I).

The nature of the dependence of Q on p can be quantified by obtaining the correlation between p and Q as defined by (3). If we obtain the correlation over the entire trajectory ($T=2.9$), then the correlation is very high, $\langle pQ\rangle_c=0.97$, which reflects the large scale similarity between the p and Q. However, if we restrict ourselves to the transition regions

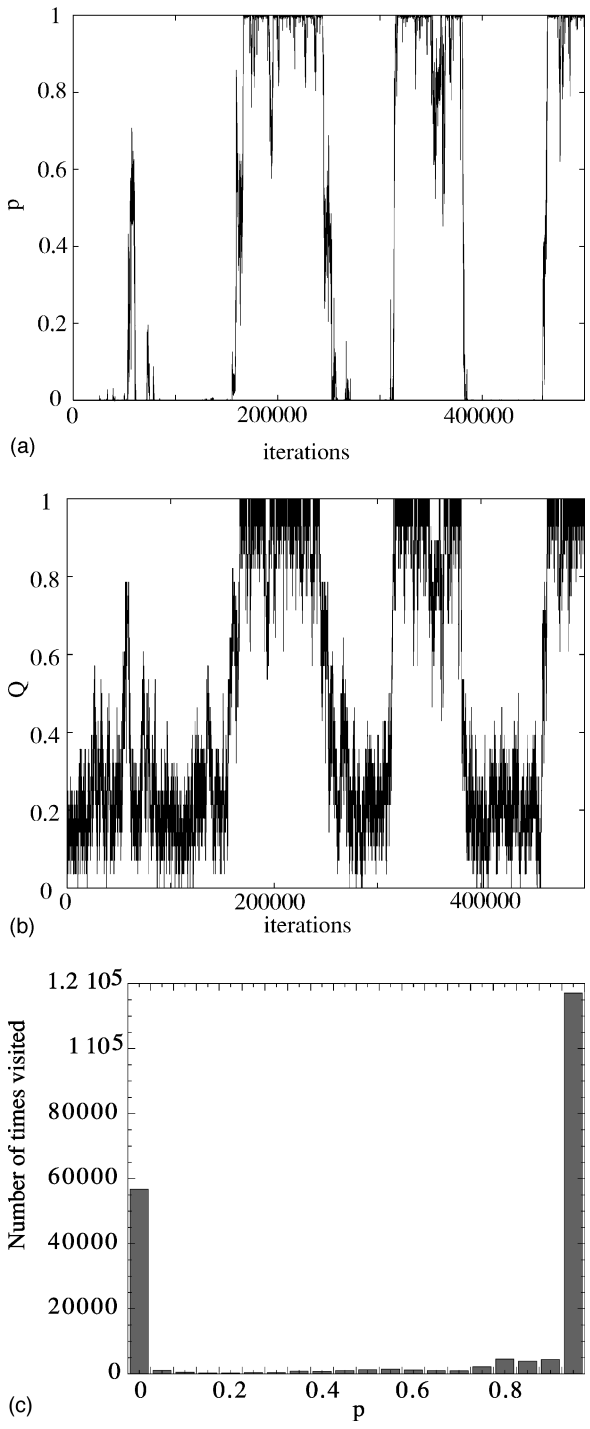

FIG. 18. Equilibrium Monte Carlo trajectory of a 27-mer under the Go interactions on the $p-Q$ plane at $T=2.9$. The transmission coefficient and Q were calculated every five Monte Carlo steps. Due to the large number of p's that must be calculated, 500 folding runs were done to calculate each p. This gives us an error of approximately 4% in p. Unfolded states are defined to be conformations with $Q\leqslant 8$ and the folded state is the native state. Q is normalized so that the maximum number of native contacts corresponds to 1. The simulation begins with a completely unfolded (straight) polymer which is allowed to equilibrate in 10^7 steps. After the initial equilibration, p and Q were taken every five Monte Carlo steps. A total of 10^6 Monte Carlo steps was performed after the initial equilibration. The first 5×10^5 steps are shown in (a) and (b). (a) p as a function of Monte Carlo time. (b) Q as a function of Monte Carlo time. Note that Q follows the large scale behavior of p very closely but that Q fluctuates much more than p. (c) The distribution of visited states over p during the Monte Carlo run.

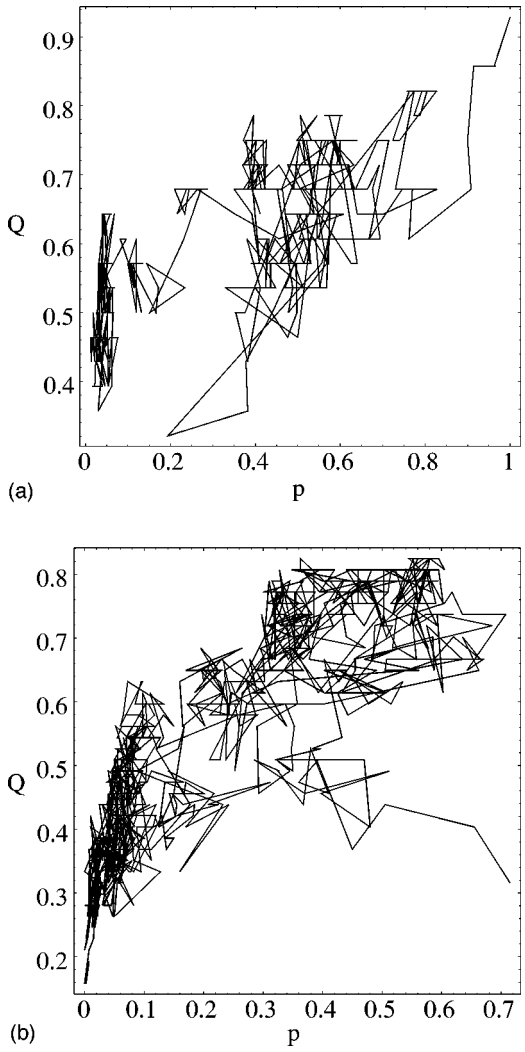

FIG. 19. A portion of a Monte Carlo trajectory of p vs Q for (a) a 27-mer under the Go interactions (see Table I) at $T=2.9$ and (b) a 48-mer under the Go interactions at $T=0.8$. Note that there is no apparent correlation between p and Q on the small scale (compare with Fig. 18).

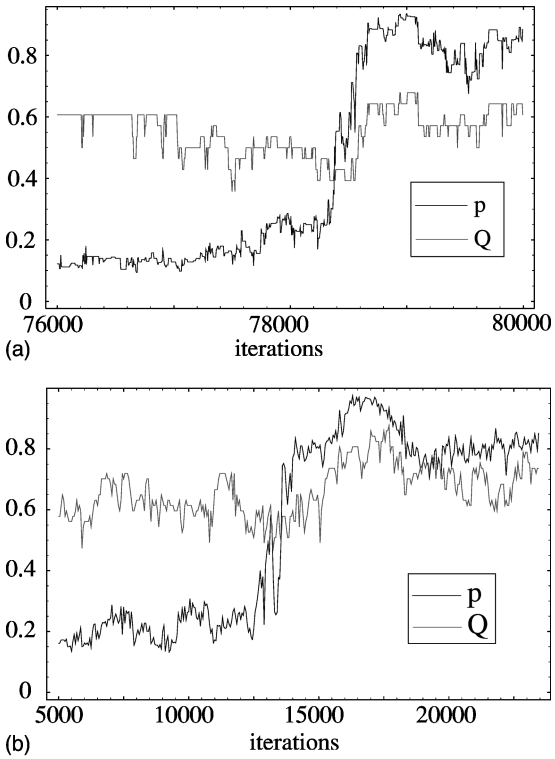

FIG. 20. A closeup look at the equilibrium Monte Carlo trajectory of (a) a 27-mer under the Go interactions at $T=2.6$ and (b) a 48-mer under the Go interactions at $T=0.8$. Note that Q remains approximately constant while p changes from ≈ 0 to ≈ 1 in both cases. The inability of Q to accurately follow p on the small scale is seen under other conditions as well (e.g. at $T=2.9$ for the 27-mer) but is seen more clearly here. This illustrates the deficiency of Q in describing the kinetics of the system.

where $0.2<p<0.8$, then the correlation is very low, $\langle pQ\rangle_c = 0.49$, which reflects the inability of Q to follow p in the transition region. On the other hand, since the p and Q are not completely uncorrelated, Q is not orthogonal to p. Thus we see clearly that even though Q is somewhat correlated with p, it cannot be used as a transition coordinate as it cannot describe the transition region.

Note also that Q fluctuates much more than p. We can quantify this through the diffusion coefficient which is obtained by using Einstein's formula,

$$6D_x t = \lim_{t\to\infty} \langle (x(t)-x(0))^2 \rangle, \tag{5}$$

with $x=p$ or $x=Q/Q_{\max}$. Since t in Eq. (5) refers to real time, the results obtained from the Monte Carlo simulation have to be multiplied by the ratio of Monte Carlo time to real time to obtain the true diffusion coefficient. However, since we are interested only in the ratio of D_Q to D_p, this factor cancels out. In Fig. 21, we see that $8>D_Q/D_p>1$ over the range of Monte Carlo step intervals used to calculate the diffusion coefficients. This implies that the system diffuses more rapidly in Q than in p which is consistent with our assertion that p is the slowest coordinate.

In addition to examining the trajectory of the system, we can use the Monte Carlo histogram technique to calculate the

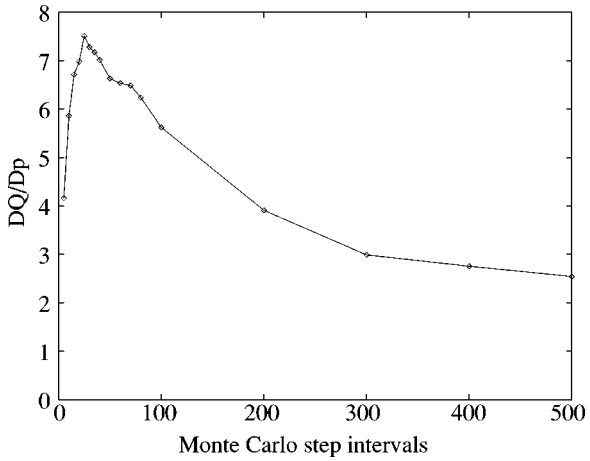

FIG. 21. Ratio of the diffusion coefficient in Q to the diffusion coefficient in p for 27-mers under the Go interactions at $T=0.29$. Note that D_Q/D_p is always greater than 1 implying that the system always moves faster along Q than it does along p.

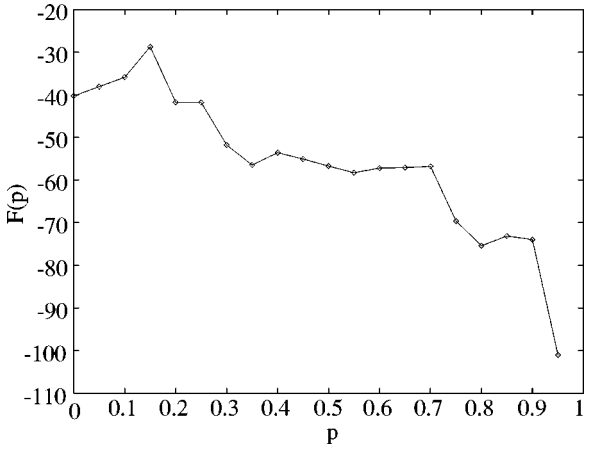

FIG. 22. Free energy profile along p for a 27-mer under the Go interactions at $T=2.9$. Note that the maximum occurs near $p=0.15$. This is because there is no dominant peak in the free energy profile.

free energy.[34] The free energy along the transmission coefficient, $F(p)$, is the free energy which has been averaged over all other coordinates (see Eq. (1)). It is the most relevant free energy landscape for the kinetics. We determine $F(p)$ for the 27-mer using the Go model at $T=2.9$ (Fig. 22). Note that while $F(p)$ does have two minima at $p=0$ and $p=1$ and a global maximum as we expected, the global maximum is not pronounced. Instead, we have a broad maximum which extends across a wide range of p's. This is consistent with the fact that the maximum occurs at $p=0.15$ as opposed to $p=0.5$ (see Section I). As discussed earlier, the transition state ensemble (defined to be at $p=0.5$) is generally not at the peak of the free energy profile along the transition coordinate. Fig. 22 clearly illustrates this concept.

One advantage of this second method is that it does not require enumeration of the polymer conformations which is impractical when we are examining larger polymers. In addition, by looking only at conformations visited during a Monte Carlo run, we avoid looking at conformations which are not kinetically accessible and therefore unimportant under normal folding conditions. Note that these kinetically inaccessible conformations do not correspond to kinetic traps. A kinetic trap is a state which the system can reach but cannot escape from immediately because it is a local minimum in energy. A kinetically inaccessible state, however, is not reachable by the system at all. Furthermore, using this method we are able to obtain the free energy profile along p. This is not possible with exhaustive enumeration because it is not possible to enumerate all the states with low Q. Finally, by looking at the trajectory, we are able to obtain information on the magnitude of fluctuations of p relative to Q which tells us how slow p is compared to Q and therefore how well Q can serve as a transition coordinate.

V. CAN THE TRANSMISSION COEFFICIENT SERVE AS THE TRANSITION COORDINATE?

So far, we have demonstrated a number of methods with which we can determine how closely various parameters approximate the transmission coefficient. We will now address the question of how well the transmission coefficient itself serves as the transition coordinate which will tell us if a good transition coordinate exists for the system. To address this question, we first compare the transmission coefficient to the folding and unfolding time. If p describes how ''close'' conformation C is to the native state in terms of folding, then p should be related to the folding and unfolding times (t_f and t_u, respectively). We find this to hold for the systems we have studied (18-mer IIM and Ising). In particular, we show in Fig. 23 a plot of the mean t_f and t_u vs p (with error bars denoting the standard deviations of the ensemble).

In addition, recall that the transition coordinate is defined as the slowest coordinate. We saw in Section IV that the system fluctuates more rapidly in Q than it does in p. While this supports our assertion that p is the slowest coordinate, it is not sufficient to prove it. Thus the question of whether a good transition coordinate exists for the system remains open.

VI. DISCUSSION

To study the kinetics of protein folding, we have introduced the concept of ''transition coordinate'' which is the slowest coordinate and the transmission coefficient, p. In general, a good single transition coordinate may or may not exist for a system. If there is such a coordinate, then it can be characterized by p. If there is no single transition coordinate, that is, if there are multiple transition coordinates, then all systems at a particular point in the multidimensional transition coordinate space will have a Dirac delta distribution of p as in the single transition coordinate case. Furthermore, even if we do not know what each of the multiple transition coordinates is, p would still be among the slowest coordinates of the system. Thus p is a robust kinetic variable. However, a theory based upon p alone is not very useful because the calculation of p involves examining the details of the kinetics directly which is not very illuminating. In particular, in order to connect the thermodynamic energy landscape with the kinetics, we need to define the transition coordinate in terms of a thermodynamic variable. In addition, the transmission coefficient is not physically intuitive in geometrical terms and is computationally intensive to calculate. Therefore, we proposed two methods by which we can determine how well any given parameter (or set of parameters when considering multiple transition coordinates) approximates p. Since we found that the transmission coefficient is dominated by geometrical factors, we have illustrated the methods by applying them to the two basic geometrical variables of the system, the number of native contacts, Q, and the looplength distribution. Many authors have proposed Q as a transition coordinate for the heteropolymer system[13,14,35,36] because of its simple geometrical interpretation and because it appears naturally in the replica approach to the thermodynamics of protein folding.[23,37–41] However, we found that the concept of contacts, or any other simple geometrical variable such as looplength, is fundamentally inadequate in describing heteropolymer kinetics, either as a local or as a global

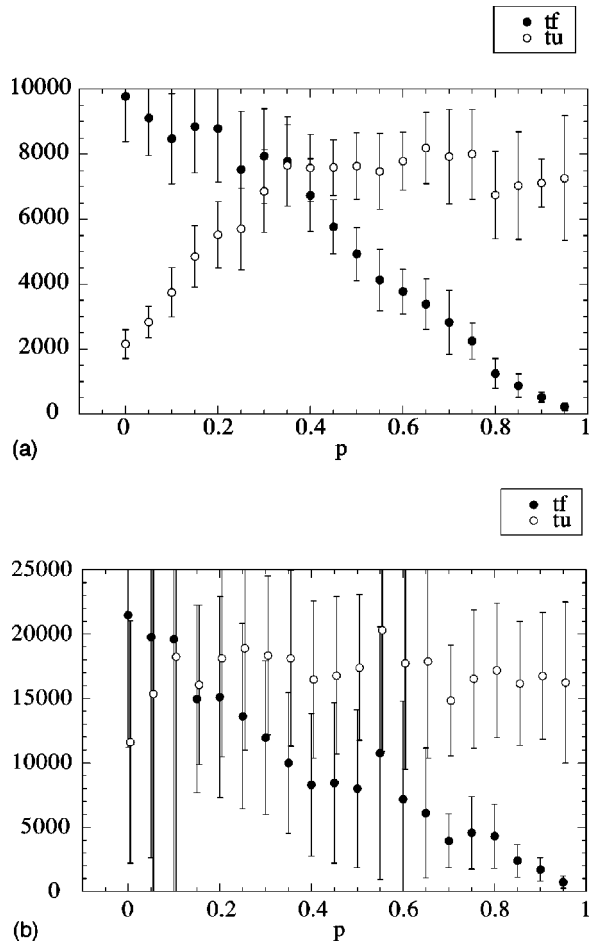

FIG. 23. Mean first passage folding time, t_f, and unfolding time, t_u, vs the transmission coefficient, p. (a) 403 18-mer conformations under the IIM (see Table I) at $T=0.75$ with $Q=9$ and no nonnative contacts (same as Fig. 11) and (b) 4944 18-mer conformations under the Ising interactions (see Table I) at $T=0.7$ and $Q=9$ with and without nonnative contacts were examined. 1000 Monte Carlo runs were used to calculate p, t_f, and t_u for each conformation. The mean t_f and t_u, averaged over all conformations with a given p, exhibit similar trends in both the IIM and Ising models although the standard deviations in the Ising model are much larger. The trend remains the same but with smaller standard deviations when the Ising model is restricted to conformations without nonnative contacts. The strong correlations between the folding and unfolding times and p in both cases indicate that p is a good measure of kinetic distance. Each error bar gives the width of the distribution of t_f's and t_u's over the conformations at that p.

parameter. It is, at best, only a rough approximation of the transition coordinate. While this effect may be partially due to the short-chain length of our model, Gutin *et al*. have found that the dynamic behavior of a polymer depends only weakly on the chain length (power law).[42] Furthermore, our results for the 18-mer and 48-mer do not differ significantly. Thus we expect that our conclusions will hold even at longer chain lengths.

A possible alternative to Q and looplength can be found by using the vapor-liquid phase transitions as an analogous model. In this model, the transition coordinate is the radius of nucleation. The concept of a nucleus according to the vapor-liquid analogy is much more complex than Q which corresponds to the density. It is very difficult to define what is meant by the ''nucleus'' or by the ''radius of the nucleus'' in polymers.[2] In particular, while we are able to see the ''nucleus'' by eye, there is no obvious geometrical coordinate by which we can characterize it.

Nevertheless, despite the intractability of the transmission coefficient as a transition coordinate due to its computational intensiveness and unclear geometrical interpretation, the study of p provides insight to the nature of the transition coordinate of the system. In particular, since we can always directly compute p using simulations, we can examine the kinetic properties of the system even if we do not understand the microscopic meaning of p.

As an illustration of the usefulness of the transmission coefficient, consider the vapor-liquid system. In this model, the quantity that is analogous to Q in the polymer system is the density of the vapor. Like Q, the density is a global parameter which cannot adequately describe the kinetics of the system. In particular, it predicts an energy barrier which scales linearly with system size so that the time needed for the vapor-liquid transition grows exponentially with chain length. This exponential increase in transition time would prevent the system from ever condensing. Instead, it is well-known that the transition coordinate for this system is the radius of nucleation. The free energy profile as a function of this radius contains two minima and a prominent maximum. Suppose, however, that we are unaware of the radius of nucleation as the transition coordinate; we can still completely describe the kinetics of the system through the transmission coefficient. In particular, we can obtain the free energy profile along the transmission coefficient and we should find that it is similar to the free energy profile along the radius of nucleation with two minima and a pronounced maximum. This suggests that a good transition coordinate exists for the system. Furthermore, the existence of a dominant maximum suggests that we can assign the transition state to the peak of the free energy profile.

When we apply the same idea to the polymer system, however, we find that the free energy profile does not contain a pronounced maximum. This alone cannot tell us whether p serves as a good transition coordinate since it is possible for the free energy profile along p to be very complicated. However, even if there is a good transition coordinate, the lack of a single pronounced barrier means that the transition state does not reside at the peak of the free energy profile. Thus the idea of a transition state as used in chemical kinetics theory may not be applicable to the polymer system.

To conclude, we stress that we do not suggest using p as a transition coordinate for practical purposes as it is very computationally intensive. The reason for using p is that, apart from the difficulties in its computation, p behaves in the same way as the best, that is, the slowest, possible transition coordinate does. Thus as long as we do not know a convenient transition coordinate, we can examine p and establish the physical properties of the transition coordinate without knowing what the transition coordinate is. In this sense, our work is purely experimental: we do not know how the system works and we measure a quantity which will re-

veal the physical characteristics of the system. This quantity is the transmission coefficient, p.

ACKNOWLEDGMENTS

The work was supported by the NSF (DMR 94-00334) and by the NIH (R01 52126). One of the authors (R.D.) acknowledges support from the MSTP fellowship and the Harvard-MIT Division of Health Sciences and Technology. A second author (V.S.P.) acknowledges funding from the Miller Institute for Basic Research at UC Berkeley.

[1] M. Karplus and E. Shakhnovich, *Protein Folding* (Freeman, New York, 1992), Chap. 4, pp. 127–195.
[2] M. Karplus and A. Sali, Curr. Opin. Struct. Biol. **5**, 58 (1995).
[3] M. Hao and H. Scheraga, J. Chem. Phys. **102**, 1334 (1995).
[4] E. I. Shakhnovich, Folding & Design **1**, R50 (1996).
[5] E. I. Shakhnovich, Curr. Opin. Struct. Biol. **7**, 29 (1997).
[6] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[7] H. S. Chan and K. A. Dill, J. Chem. Phys. **95**, 3775 (1991).
[8] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, Rev. Mod. Phys. (to be published).
[9] B. Derrida, Phys. Rev. B **24**, 2613 (1981).
[10] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[11] V. S. Pande, A. Yu. Grosberg, C. Joerg, and T. Tanaka, Phys. Rev. Lett. **76**, 3987 (1996).
[12] E. I. Shakhnovich and A. V. Finkelstein, Biopolymers **28**, 1667 (1989).
[13] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, J. Mol. Biol. **252**, 460 (1995).
[14] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, Folding & Design **2**, 109 (1997).
[15] H. Eyring, *Modern Chemical Kinetics* (Reinhold, New York, 1963).
[16] E. M. Lifshits and L. P. Pitaevskii, *Physical Kinetics* (Pergamon, New York, 1981).
[17] S. W. Koch, *Dynamics of First-Order Phase Transitions in Equilibrium and Non-equilibrium Systems* (Springer, Berlin, 1984).
[18] In chemical kinetics, the transmission coefficient is defined differently as the number of states that progresses directly from A to B through the transition state divided by the number of states that actually moves in the same direction through the peak of the free energy profile but may be reflected back later to A.
[19] I. N. Levine, *Physical Chemistry* (McGraw-Hill, New York, 1983).
[20] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[21] S. P. Obukov, Phys. Rev. A **42**, 2015 (1990).
[22] R. Du, A. Grosberg, and T. Tanaka (unpublished).
[23] E. Shakhnovich and A. Gutin, Biophys. Chem. **34**, 187 (1989).
[24] N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).
[25] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[26] A. Sali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[27] K. Binder, *Applications of the Monte Carlo Method in Statistical Physics* (Springer, Berlin, 1984).
[28] E. Shakhnovich and A. Gutin, Proc. Natl. Acad. Sci. USA **91**, 12 972 (1994).
[29] V. S. Pande, A. Yu Grosberg, and T. Tanaka, J. Chem. Phys. **103**, 9482 (1995).
[30] A. Gutin, V. Abkevich, and E. Shakhnovich, Proc. Natl. Acad. Sci. USA **92**, 1282 (1995).
[31] Thirumalai suggested that the folding time grows exponentially with N^{ν} where $\nu=0.5$, if a protein, is assumed to follow the random energy model or Potts glass model (Ref. 32). The exact value of ν does not affect our analysis as $\exp N^{\nu}$ is still much larger than N^{α}.
[32] D. Thirumalai, J. Phys. I **5**, 1457 (1995).
[33] R. Bird, O. Hassager, R. Armstrong, and C. Curtis, *Dynamics of Polymeric Liquids* (Wiley, New York, 1977).
[34] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **103**, 4732 (1995).
[35] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[36] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, J. Chem. Phys. **104**, 5860 (1996).
[37] K. Binder and A. P. Young, Rev. Mod. Phys. **58**, 801 (1986).
[38] T. Garel and H. Orland, Europhys. Lett. **6**, 307 (1988).
[39] E. I. Shakhnovich, G. M. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[40] S. Ramanathan and E. I. Shaknovich, Phys. Rev. E **50**, 3907 (1994).
[41] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, Macromolecules **28**, 2218 (1995).
[42] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996).