Folding kinetics of proteins and copolymers

Robert Miller, Craig A. Danko, Michael J. Fasolka, and Anna C. Balazs
Department of Material Science and Engineering, 848 Benedum Hall, University of Pittsburgh, Pittsburgh, Pennsylvania 15261

Hue Sun Chan and Ken A. Dilla
Department of Pharmaceutical Chemistry, Box 1204, University of California, San Francisco, California 94143

(Received 2 August 1991; accepted 23 September 1991)

We model the kinetic processes by which globular proteins and other heteropolymers fold to compact states. We perform Monte Carlo dynamics simulations on short self-avoiding copolymer chains on two-dimensional square lattices. The driving force for collapse is the aversion of nonpolar monomers for water. The chain monomers are of two types, H and P; favorable interactions occur among HH contacts. One respect in which this study differs from previous Monte Carlo folding studies is that the chains are sufficiently short that: (i) we can know unequivocally which conformations are at global minima (the "native" states) and which are at local minima of free energy, (ii) we can explore "pathway space" densely to determine the relative probabilities of all the possible pathways, and thus we establish that the model is ergodic and gives the equilibrium distribution in the long-time limit. We find that any individual molecule passes through a wide range of conformational states, often many times. Nevertheless, this meandering is not inconsistent with the observation that proteins fold through specific pathways involving particular sequences of events. Some pathways are strongly favored; i.e., folding is "cooperative" in that a "nucleating" HH contact acts as a constraint that restricts local conformational freedom and speeds the "zipping up" of other contacts nearby. Which pathways are favored depends on the sequence and the solvent. How does a chain fold to its native state so quickly? For these short chains of fixed length, we observe three regimes of folding kinetics. In the "Levinthal limit," as the HH attraction approaches zero, folding is slow because the molecule searches randomly through the large ensemble of open conformations. In the "multiple minima" limit, when the HH attraction is very strong, folding is also slow because the chain becomes stuck in local minima of wrong compact conformations. However, we find that folding is relatively fast for intermediate HH attraction because the driving force "directs" the molecule toward a small ensemble of compact states, and yet incorrect potential wells are not too deep to slow this process.

I. INTRODUCTION

The kinetic processes by which a protein folds to its native state are not yet understood. Some of the main puzzles are the following: (1) How does a protein chain fold so quickly (milliseconds to hours) to find the single "native" conformation of lowest free energy, starting from an initially large ensemble of denatured configurations? This is the so-called Levinthal paradox.1 (2) Experiments show that a given protein folds according to a relatively small number of "pathways," i.e., by specific sequences of molecular events.2-4 The existence of pathways implies "cooperativity," the increased probability that a chain will enter a particular conformation, given that it had previously been in some particular other predecessor conformation. How does this cooperativity arise from the driving forces? (3) How is the pathway determined by the sequence of amino acid monomers in the chain and by external conditions? (4) What is the rate-limiting step?

Mainly two types of interaction contribute to protein folding: (i) the helix–coil propensities5,6 among monomers that are connected neighbors in the chain sequence, and (ii) the hydrophobic and solvent interactions among monomers that may either be near each other or far apart in the chain sequence. In the absence of better terminology,7 we refer to these as local and contact interactions, respectively. Local thus refers to torsional, rotational, and steric interactions among neighboring bonds, and includes hydrogen bonding among groups 24 monomers apart. Contact interactions derive from the differential interactions among monomers and solvent molecules, as measured for example by oil–water-type transfer experiments. Contact interactions can occur between monomers that are either close together or far apart in the sequence. For proteins, the dominant contributors to contact interactions are the hydrophobic and hydrogen-bonding and dispersion interactions that drive nonpolar and polar contacts. The main effects of solvent are on the contact interactions.8,9

Several current models of folding kinetics attribute considerable importance to the local interactions in initiating early events in protein folding. In this view, the first stages of
folding involve the formation of helices\textsuperscript{3,10} or turns.\textsuperscript{11,12} This accounts for some observations that helices form quickly\textsuperscript{2} and provides an obvious basis for cooperativity; the zipping of a helix might be rapid once the first turn is formed. However, the following puzzles are not resolved by the view that local interactions are predominant. (1) The thermodynamic intrinsic tendencies for short peptides to form helices in aqueous solution are weak,\textsuperscript{13-17} whereas the contact interactions are relatively strong.\textsuperscript{8,24} (2) The “jump” procedure used in most experiments to monitor protein folding kinetics usually involves a change in concentration of guanidine hydrochloride or urea; the predominant effect of these solvents on protein folding is generally assumed to be on the nonpolar driving force. We are unaware of evidence that jumps in these solvents could induce correspondingly large changes in the local interactions, e.g., to initiate helix formation. Moreover, such an explanation would be unlikely to account for the folding of all-sheet proteins, which have relatively few local interactions. (3) Since the end point of folding (the native structure) is uniquely determined by the amino acid sequence, then the folding pathway must also be determined by the amino acid sequence. But the propensities to form helices are relatively independent of amino acid sequence, since the main interactions in helices are hydrogen bonds among backbone atoms. The manner in which the folding pathway depends on the amino acid sequence thus requires explanation in side-chain interactions.

In the present paper, it is our aim to explore an alternative hypothesis. We ask: What are the rate processes of copolymer collapse for chains that interact exclusively by contact interactions? The equilibrium aspects of homopolymer and heteropolymer collapse have been widely studied.\textsuperscript{8,19-24} The several previous Monte Carlo simulations that aim to model protein folding kinetics\textsuperscript{25-33} do not address this question since they involve combinations of physical and nonphysical local and contact interactions. The present model also differs in one other respect from earlier models. Previous Monte Carlo studies on longer chains explore only a small fraction of conformational space. It follows that those studies: (i) cannot distinguish global from local minima in the free energy, and (ii) cannot determine the relative probabilities of different pathways since the necessarily sparse Monte Carlo sampling of the conformational space explores only a small fraction of them.

In the present work, we study short chains on two-dimensional lattices. In this way, we can (i) exhaustively explore the full conformational space to identify the conformations of global minimum in free energy, and (ii) explore densely all the possible pathways the chain could take to the native state. We find here that chains driven to collapse solely by contact interactions resemble protein folding processes in the following respects. (1) The short chains studied here ultimately find their unique native structures. (2) Even though a given molecule follows an apparently random process, there is nevertheless a cooperativity resulting in highly favored folding pathways for the ensemble of folding molecules. The cooperativity arises from chain compactness and thus differs significantly from the type of cooperativity that arises from local interactions in helix-coil transitions.

We study the following model. A linear chain molecule is configured as a self-avoiding walk on a two-dimensional square lattice. The chain has a specific sequence of H (hydrophobic) and P (other) monomers. It is in an H-aversive solvent (such as water, if II represents nonpolar monomers). For each HH contact in a given spatial configuration, there is a favorable energy $\epsilon < 0$. This energy will depend on the solvent and temperature. The dependence on temperature may be complex, but that is of no consequence for the present investigation. Each amino acid monomer (residue) is represented as occupying one lattice site, connected to its chain neighbor(s), and unable to occupy a site filled by any other residue. We define “connected” neighbors as the units $j$ and $j + 1$ adjacent in the chain and “topological” neighbors as units $(i, j)$ one lattice site apart in space, but not adjacent in position along the sequence.\textsuperscript{3} We study only short chains, for which all possible conformations have previously been enumerated to determine the equilibrium ensemble. In this way we find the conformation(s) of global minimum in free energy. These are identified as the “native state(s).”

There are several reasons we believe that this model, which is obviously highly simplified, has some bearing on the folding kinetics of proteins. The two main simplifications are that: (i) the chains are short, and (ii) the model is two dimensional. The shortness of the chains is necessary for our present purposes, which require exhaustive enumerations in order to establish which conformations are at global and local minima, and in order to explore densely the pathway space. For these purposes there is presently no feasible alternative to using short chains. The fact that the model is two dimensional (2D), rather than three dimensional (3D), offers certain advantages. The principal variable that characterizes the driving forces is the surface/interior ratio. For real proteins with 100 monomers in three dimensions (that are approximately spherical), simple geometric arguments show that only about 20%-30% of the monomers can be in the core; 70%-80% of the monomers must be on the surface of a compact conformation. This surface/interior ratio (75% surface/25% core) is achieved on the 2D square lattice for chains of only 16 monomers in length, and thus can be explored in exhaustive simulations in 2D. On the other hand, in 3D the smallest meaningful model is the 27 monomer cube.\textsuperscript{3} But such a 3D model has two limitations for our present purposes: (i) there is only a single monomer comprising a hydrophobic core, thus the driving forces are not well represented, and (ii) even chain length 27 in 3D is too long for exhaustive simulation, so the native structures for the present HP model cannot be known, or pathways broadly explored, or the processes of kinetic trapping and escaping from local minima determined. Thus, for our purposes, a 2D model is more suitable than a 3D model. The main differences between 2D and 3D models are the coordination number of the lattice, which determines the neighbor interactions and the configurational entropy per polymeric bond, and the greater severity of steric constraints in two dimensions. These factors are important determinants of the quantitative energetics of conformational transitions, but less important, we believe, for qualitative questions ad-
dressed here: (i) Are pathways a natural consequence of collapse driven by contact interactions? (ii) How do trapping and escape processes upon folding to the native state depend on the HH energy? Hence we believe the restriction to two dimensions is not an important limitation for the purpose of the principles we study here.

There are several respects in which this simple 2D short chain H/P model mimics the equilibrium properties of proteins.\(^{36-38}\) For a small degree of attraction between the nonpolar units (small \(|\epsilon|\), most of the chain conformations are denatured. For molecules comprised of certain sequences (referred to as folding sequences), increasing the attraction between the nonpolar units leads to a transition to a relatively small number of native conformations (most often, only one or two). The native conformations have the properties that they are maximally or near-maximally compact, are comprised of a core of predominantly nonpolar residues, and have considerable amounts of two-dimensional equivalents of secondary structures: helices and sheets. The length distributions of helices and parallel and antiparallel sheets in the short-chain 2D lattice model are similar to those that are observed in the known proteins.\(^{39}\) Any model of the forces in proteins must account not only for protein structures but also for protein mutability and evolution, i.e., for how the changes in amino acids lead to changes in native structures. The short chain H/P 2D lattice model is consistent with experiments on protein mutability in the following respects:\(^{37,38,40}\): (i) surface sites are highly mutable, (ii) core H sites are sensitive to mutation, (iii) most mutations among sequences that code for unique native states are neutral insofar as they do not change the native structure, implying that proteins should be highly plastic to mutation, (iv) there are many convergent sequences that should fold to any given native structure, and (v) sequence space is relatively smooth and connected. It is in these respects that the 2D lattice model mimics protein behavior.

II. METHODOLOGY: THE MONTE CARLO SIMULATION

Our aim is to determine the series of conformational changes a chain undergoes in going from an arbitrary initial state to its final native structure. We refer to this series of events as the folding "pathway." To avoid confusion, we use the term "sequence" only to refer to the order in which the monomers are linked in the chain. To determine these pathways, we use the Monte Carlo computer simulation approach described below. Through this simulation, we study the folding (and unfolding) pathways of different copolymer chains that are 13 and 14 lattice monomers in length. In order to determine which particular copolymers to examine, we first performed exhaustive searches\(^{36-38}\) over the full conformational space to find a few H/P sequences that have only a single native structure. The selected H/P sequences are shown in their corresponding native state structures in Fig. 1. The filled squares represent H monomers and the open circles denote the P sites. Having chosen these structures, we then performed the Monte Carlo simulations to follow the folding pathway for each particular copolymer.

The simulation begins with a chain in an arbitrary open state with no HH topological contacts. A specific monomer of the chain is chosen at random and is allowed to move via the algorithm for chain dynamics developed by Verdier and Stockmayer\(^{41}\) and Hilhorst and Deutch.\(^{42}\) Specifically, four types of local chain motion are allowed at each attempted move: an end monomer (bead) can either rotate 90° or 180°; an interior monomer (bead) can move either through a three-bead "flip"\(^{44}\) or a four-bead "crankshaft" ("double flip").\(^{42}\) All moves must obey the excluded volume criteria: no lattice sites can be doubly occupied. Whenever a move causes two Hs to become topological neighbors, the two sites form a (noncovalent) contact.

After a monomer of the chain has been selected by the random process, the appropriate move is attempted. To determine whether this move is accepted, we invoke the standard Metropolis algorithm.\(^{43,44}\) In particular, the total number of HH contacts in the system is calculated before and after the specific move. If the number of HH contacts is increased or unchanged, the free energy of the system will be lowered or unchanged, thus the move is accepted (subject to the exclusion volume criterion). If the number of HH contacts is decreased by the attempted move, the free energy of the system is increased, and thus the move is only accepted with probability \(P^N\), where \(N (> 0)\) is the decrease in the number of HH contacts. The parameter \(P\) defines a "detaching" probability; it can be related to temperature and the energy of the HH interaction through the equation

\[
P = \exp(\epsilon/kT),
\]

where \(\epsilon < 0\) is the HH interaction energy (see above), \(k\) is Boltzmann's constant, and \(T\) is absolute temperature.\(^{43-45}\) Thus a change in \(P\) corresponds to the physical process of

![Diagram](image-url)
changing the temperature or strength of the interaction between the nonpolar groups (either through a change in solvent or through a change in the chemical nature of the species). \( P \) is specified in advance and can be varied from 0.0 (irreversible binding) to 1.0 (no HH attraction).

After each accepted move, we collect data characterizing the conformation of the chain, such as the number of native HH contacts (those present in the native state) and non-native HH contacts. We also record the specific contacts that have been made or broken, and the number of Monte Carlo time steps that have been executed. The clock counts time steps, independent of whether the move attempted is successful or not. For brevity, the number of time steps will be referred to below as “time,” even though there is no simple conversion between the number of Monte Carlo steps and real time. The simulation is run until the molecule reaches its native state, which, as noted above, is known from prior exact enumerations. 36-38 For each of the three copolymer sequences examined, and for different values of \( P \), this procedure was carried out 20 to 100 times each.

One goal of these studies is to investigate the relationship between the folding and unfolding processes. To determine the unfolding pathways, we start with the chain in the native state and apply the above Monte Carlo rules until the chain reaches a conformation that is completely “open,” i.e., for which there are no HH topological contacts.

III. RESULTS

Figure 2 shows an example of an observed folding pathway for sequence S1. It is clear that any particular individual contact may be made or broken many times in the folding process.

Table I shows the distribution of states generated by the Monte Carlo dynamics compared with the exact equilibrium distribution, determined by exhaustive enumeration. The results indicate that 20 folding runs (pathways) were sufficient to achieve agreement between the Monte Carlo and exact equilibrium distributions to within a few percent. Thus the present Monte Carlo procedure is ergodic.

![FIG. 2. A representative folding pathway for sequence S1. Each vertical bar indicates that the particular HH contact shown on the y axis was made during the corresponding time interval shown on the x axis. Boxes at the right end identify contacts which are native. The native structure of sequence S1 is given in Fig. 1.](image)

We find that, on average, the molecules do not explore all the possible pathways. Some pathways are rarely or never traversed. The full distribution of pathway frequencies from our simulations are too complex to display, so we simplify them in the form of “reduced” pathway maps such as Fig. 3 which is shown for sequence S1. Shown here are only diagrams having correct native contacts and no incorrect contacts. In this map, the only pathways shown are those that

<table>
<thead>
<tr>
<th>Number of wrong contacts</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of right contacts</td>
<td>24518*</td>
<td>9197</td>
<td>1767</td>
<td>253</td>
<td>25</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>60.36%</td>
<td>22.64%</td>
<td>4.35%</td>
<td>0.62%</td>
<td>0.06%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>60.42%</td>
<td>22.75%</td>
<td>4.33%</td>
<td>0.56%</td>
<td>0.04%</td>
<td>0.01%</td>
<td>0.0%</td>
<td></td>
</tr>
<tr>
<td>2556</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6.29%</td>
<td>3.13%</td>
<td>0.57%</td>
<td>0.04%</td>
<td>0.007%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.34%</td>
<td>3.1%</td>
<td>0.58%</td>
<td>0.03%</td>
<td>0.01%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>386</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.95%</td>
<td>0.64%</td>
<td>0.13%</td>
<td>0.005%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.88%</td>
<td>0.62%</td>
<td>0.12%</td>
<td>0.01%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.07%</td>
<td>0.08%</td>
<td>0.015%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.09%</td>
<td>0.09%</td>
<td>0.02%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* indicates an exact value.
FIG. 3. Pathway maps for sequence S1: (a), (b), and (c) correspond to different sticking energies, which are related to the detaching probabilities by Equation (1). Each native HH contact (with energy $e < 0$) is indicated as a pair of filled squares, each hollow square indicates an H residue not involved in a native contact, and each circle indicates an P residue. Only one example conformation is shown for each state with a specific set of HH contacts. Lines joining different diagrams indicate conformational adjacency and the number of times the path is traversed in 100 pathways.
are accessed. The numbers along the connecting lines on pathway maps such as Fig. 3 show the number of times that the given pathway was traversed. This number does not explicitly count all the traversals. In most pathways, there is considerable oscillation between two adjacent diagrams. Therefore, we have found it preferable as an indicator of frequency of traversal to count only the number of times an entrance/exit occurs for the step indicated by a given pair of diagrams. For example, if a molecule configures as represented by a diagram I, then as represented by diagram II, back-and-forth 100 times, then exits from II to III, we increment the counter on the I → II step by one, rather than by 100. We do this only for the purpose of illustration and for removing “irrelevant” actions; for calculations below for counting traversals, we count each step explicitly.

Figure 3 shows that some pathways are traversed more often than others. Comparison of the two sets of numbers in Fig. 3 shows that there are some differences in the relative importances of specific pathways in the folding and unfolding directions. The relative importance of a pathway also depends on the HH sticking energy; compare Figs. 3(a), 3(b), and 3(c) for sequence S1. These preferences can be rationalized on the basis of the conditional probability for a jump from one conformation to another. In the absence of a sticking energy, a particular HH sticking event will compete on equal footing with all other possible conformational changes, including those that make or break other HH contacts. On the other hand, if the sticking energy is strong, then a particular HH sticking event will occur with much greater frequency than HH-contact breaking events, and will pre-
dominantly only compete with other HH-contact making events. Thus the folding process is more "directed" to go energetically downhill if the HH sticking energy is strong, although not necessarily always toward the native structure. We observe that chains appear to follow a broader distribution of pathways for intermediate values of the HH sticking energy than in the weak or strong sticking limits. Sequences S2 and S3 exhibit similar behavior, as is evident from their pathway maps (Figs. 4 and 5).

The reason that certain pathways are favored relative to others, i.e., the cooperativity of the folding process, is apparent from the six configurations in Fig. 3(a) labeled A, B, C, D, E, and F. Given that the chain reaches configuration A, it much prefers to pass through intermediate state C or D than B. Two factors determine the favorability of one path relative to another: the advantage in sticking energy, and the advantage in configurational entropy. The entropy effect is as follows. Though B, C, and D have the same number of HH contacts, the preferred process is from A to C or A to D, rather than from A to B because C and D are much less restricted in their conformational entropies than B. The sticking energy contribution to the cooperativity is as follows. Once the chain reaches D, it much prefers to pass through immediate state F than E. This is because F represents a gain of two HH contacts whereas E represents a gain of only one. The relative importance of the energetic and entropic factors in determining the favored pathways depends on the sticking energy. Comparison of the above transitions in Figs. 3(a)-3(c) shows that as the sticking energy weakens, the energetic advantage diminishes.

The cooperativity observed here for copolymer collapse kinetics is of a very different nature than the cooperativity in...
helix-coil processes. In the latter processes, there is only a single avenue of "flow": the first step is "nucleation," involving the formation of the first turn of the helix (which is entropically costly), and the second and subsequent steps involve "propagation" in a single direction (which is entropically less costly). In copolymer collapse there are multiple avenues of flow because at each stage there can be different HH contacts that can form, and each may have a different entropy loss. There are three main observations about these copolymer collapse kinetics that can be seen in the pathway maps. (1) The pathway of folding is determined by the monomer sequence. Comparison of the three pathway maps in Figs. 3–5 shows that sequence S1 folds mainly via a single pathway [A–D–F in Fig. 3(a)], whereas sequences S2 and S3 fold through multiple converging and diverging channels. (2) Most contacts formed on the main pathways are the most local ones that can be formed from the given partial conformation. For example, the first steps in the dominant paths for folding sequences S1 and S3 are order 3, i.e., (i,i + 3), HH contacts.34 For sequence S1, every subsequent step in the dominant pathway involves only successive sets of contacts that are "effectively order 3" relative to the given conformation at that stage: i.e., the two free bonds immediately connected to the existing hydrophobic cluster are first folded to the native position by forming an extra HH contact, resulting in an enlarged hydrophobic cluster, then the next two free bonds are folded, etc. For sequence S3, two (i,i + 3) HH contacts form first, then the two ends of the chain assemble together (Fig. 5). These results can be summarized in terms of a principle of minimum conformational entropy loss per step. Some justification for "most probable path" processes is given by Kikuchi.46 We believe the importance of this observation is that it may offer an explanation for the Levinthal paradox, and how long chain H/P copolymers can collapse in a time that is much shorter than would be required by exhaustive search processes. In the present study, however, we do not further consider the issue of the chain length dependence of folding time. (3) However, there are also examples of HH contacts on the main pathway that are not local. The dominant first step in the path for folding sequence S2 is that the ends come together. The explanation for this is that the favorability of a pathway consists of a product of conditional probabilities. The probability of an intermediate state depends not only on the probability of its predecessor conformations, but also on the probability of its successor conformations. For sequence S2, the intermediate with a single order-3 HH contact (at the left of the second row in Fig. 4) should have a high probability of formation, because, in general, low-order contacts are more favorable than high-order contacts.34 However, after the adoption of conformations with this specific HH contact, the chain is often forced to backtrack from these conformations because there is only one way that this intermediate can fold further. In this connection, it is interesting to note that among the first rapid steps of folding cytochrome C, the ends come together to form a hydrophobic contact among N-terminal and C-terminal helices.47

Some models of protein folding pathways, including the diffusion–collision model,48,49 and the "framework" model, assume that helix and sheet structures form early in pathways of folding. The chains in our simulations were too short to test these predictions.

Figure 6 shows how frequently different numbers of native and non-native contacts are made (and broken), averaged over all the pathways, for sequence S1. Results from the other sequences show similar features (not shown). Two general features of the pathways are clear from these figures. First, when the HH sticking energy is weak (P → 1) the folding pathways pass very often through conformations that have few HH contacts because these constitute most of the states in conformational space. Second, when the HH sticking energy is strong, the folding chain passes through any one state far less frequently than when the sticking energy is weak, and most often it passes through states that are more compact and have more native contacts than when the sticking energy is weak.

Whereas Fig. 6 shows the frequency of passing through a given state, Fig. 7 shows the average time spent in each state. There are two general conclusions from these results. First, when the HH sticking energy is weak, the molecules spend little time in any given state, and greater time is spent in more open conformations than in compact ones. Second, when the sticking energy is strong, chains spend far more time in any given state, and most of the time is spent in compact states. In this case, the chains become compact very quickly. Thus in this model the compact states are those that account for the kinetic bottleneck on the way to the native structure. The time spent in a state depends most importantly on the total number of HH contacts, and is much less dependent on whether contacts are native or non-native. This is readily understood since the time spent in a given state should just depend on the Boltzmann weight of that state, and hence on the number of HH contacts, rather than on any other details of their distribution.

Figures 8 and 9 show the time required (averaged over pathways) for the molecule to proceed to the native state, as a function of the sticking energy, from the last appearance (prior to the native state) of conformations of a given set of native and non-native contacts. When the sticking energy is weak (P → 1), the time required to reach the native state is long because the chain performs an undirected random walk to find the single native conformation. Interestingly, the time required to reach the native state is longest if the chain starts from the compact states. In this case, the most important determinant of the time required to fold is the compactness of the starting conformation; it is less important whether HH contacts are "right" or "wrong." This is the case for all three sequences. This result would seem to be puzzling. Why should a chain that has 4 out of 5 possible native contacts take longer to reach the native state than a chain that has 0 native contacts? A more detailed look at our pathways shows that this arises because the number of native contacts is an insufficient measure of the "closeness" of a molecule to its native state. Many conformations that have several native contacts are "poor" in that many specific conformational moves are required to break the native contacts and re-form them into a more proper conformation in order that the process can proceed toward the native structure.
Thus the number of native contacts alone does not tell the whole story. Moreover, the independence on sequence arises because the main constraint, in the weak sticking limit, must be the presence of severe steric constraints, rather than the HH interactions.

At the opposite extreme, when the sticking energy is strong ($P \rightarrow 0$), comparison of Figs. 8 and 9 shows that the time required to get to the native state differs depending on the sequence. In the strong sticking limit, the time to the native state becomes increasingly determined by the number of wrong contacts: the more non-native contacts, the longer it takes to get from that state to the native conformation. From the fact that this time depends on the relative number of right and wrong contacts, it follows that there is a dependence on the amino acid sequence, because only the sequence can determine right from wrong. This arises because when the number of non-native contacts is large, the potential well that must be escaped in order to get onto a pathway toward the native state is deep. On the other hand, conformations with native contacts are of two types: those that are on the pathway to native, which can fold quickly, and those that are not, which are therefore also stuck unproductively in local minima. Hence in the strong sticking limit, most time is spent in sluggish exploration of the deep wells of compact conformations; it is relatively infrequent that a wrong conformation can escape its potential well to break a contact and to then proceed toward the native state.

These simulations also show a most interesting result that for intermediate values of the sticking energy ($P \approx 0.5$), the folding time is relatively fast compared to either the weak...
or strong sticking limits. In this case, the formation of HH contacts, native or non-native, directs the folding energetically downhill to compact states, but incorrect local minima are not too deep and the chain can readily escape to better conformations. Again, the longest time to the native state results from starting in the most compact configurations.

Figure 10 provides an overview summary of these results, and shows the time-to-native state from any configuration, as a function of the HH detaching probability. This shows more clearly the point made above that the time to fold is longest in either the weak or strong sticking limits, and that for intermediate sticking energy, the time to fold to the native structure is shortest. The difference in folding times is large; even for these short chains, the difference in the time required to fold is nearly 2 orders of magnitude.

IV. CONCLUSIONS

We have developed a simple model for the collapse kinetics of proteins and copolymers. Short chains are modelled as self-avoiding walks on 2D square lattices. Chains are copolymers of monomer types H and P; HH contacts are favored by a sticking energy, $e$. By Monte Carlo simulations,
these chains undergo conformational change until they ultimately reach the native configuration at the global minimum of free energy.

This copolymer collapse model shows the following features of protein folding kinetics:

1. For sufficient IIII attraction, the chains spend much time in compact but non-native states on their way to the native structure.

2. The time required to fold to the native state is considerably faster than would occur by the random search process of Levinthal, if there is some HH attraction. In the Levinthal limit of no HH attraction, the equilibrium disfavors the native structure. However, if the HH attraction is too strong, the kinetics is slow because chains stick in local minima. We do not address here the important problem of how the folding time scales with the chain length; this will be considered in subsequent work.

3. A most interesting result is that there are specific pathways of folding in copolymer collapse processes. The pathway of folding is dependent on the monomer sequence and on the strength of the HH attraction. The cooperativity is of a fundamentally different nature than that which arises in classical helix-coil transition processes that are driven by local interactions. The cooperativity observed here derives entirely from solvent-mediated contact interactions. Although the specific pathways that we observe are undoubtedly dependent on the Monte Carlo move set and the lattice details, we believe the most important and robust result is

FIG. 8. The average time required for the chain to reach the unique native state from initial states of given numbers of native (right) and non-native (wrong) HH contacts (for sequence S1).
that the existence of sequence-dependent favored pathways should be expected to be a natural consequence of copolymer collapse processes driven by contact interactions.

ACKNOWLEDGMENTS

A. C. B. gratefully acknowledges financial support from the National Science Foundation through Grant No. DMR-9100818, and the Holchst Celanese Corporation. H. S. C. and K. A. D. thank the NIH, the URI Program of DARPA for financial support.

FIG. 9. Same as Fig. 8 for sequence S2.

FIG. 10. Time-to-native state vs sticking energy, averaged over initial configurations, for all three sequences.
(i) and "through-solvent" for (ii) are not completely accurate because helix-coil propensities (i) also include a considerable contribution from through-space hydrogen bonding among (i,i + 3) neighbors and because interactions among connected neighbors in polymers are potentials of mean force (Refs. 8 and 9) and are not completely independent of the solvent. The terms "local" and "nonlocal" are also not completely accurate for (i) and (ii). Local refers to interactions among connected and near neighbors in the monomer sequence, and nonlocal refers to interactions among monomers that are far apart in the chain sequence (Ref. 8). This terminology is not suitable here because hydrophobic interactions (ii) can occur among monomers near each other in the sequence.


For the 27 monomer cube, there are 4,960,608 conformations (or 4,960,608/48 = 103,346 conformations if rigid rotations and inversions are counted as identical); see H. S. Chan and K. A. Dill, J. Chem. Phys. 92, 3118 (1990).


