PROTEINS: Structure, Function, and Genetics 32:425–437 (1998)

Statistical Mechanics of Protein Folding by Exhaustive Enumeration

Gordon M. Crippen* and Yoshiaki Zenmei Ohkubo
College of Pharmacy, University of Michigan, Ann Arbor, Michigan

ABSTRACT **It is hard to construct theories for the folding of globular proteins because they are large and complicated molecules having enormous numbers of nonnative conformations and having native states that are complicated to describe. Statistical mechanical theories of protein folding are constructed around major simplifying assumptions about the energy as a function of conformation and/or simplifications of the representation of the polypeptide chain, such as one point per residue on a cubic lattice. It is not clear how the results of these theories are affected by their various simplifications. Here we take a very different simplification approach where the chain is accurately represented and the energy of each conformation is calculated by a not unreasonable empirical function. However, the set of amino acid sequences and allowed conformations is so restricted that it becomes computationally feasible to examine them all. Hence we are able to calculate melting curves for thermal denaturation as well as the detailed kinetic pathway of refolding. Such calculations are based on a novel representation of the conformations as points in an abstract 12-dimensional Euclidean conformation space. Fast folding sequences have relatively high melting temperatures, native structures with relatively low energies, small kinetic barriers between local minima, and relatively many conformations in the global energy minimum's watershed. In contrast to other folding theories, these models show no necessary relationship between fast folding and an overall funnel shape to the energy surface, or a large energy gap between the native and the lowest nonnative structure, or the depth of the native energy minimum compared to the roughness of the energy landscape. Proteins 32:425–437, 1998.** © 1998 Wiley-Liss, Inc.

Key words: theory of protein folding; folding funnel; folding thermodynamics; folding kinetics; conformation space; sequence/structure compatibility; thermal denaturation

INTRODUCTION

What are the main ideas we can glean from the vast protein folding literature? For a recent review, see Dill et al.[1] To succinctly summarize the experimental situation, it is generally held that a small fraction of all possible amino acid sequences having sufficient chain length fold reversibly in dilute solution, without external guidance, to a fairly unique native conformation, as long as the temperature and solvent composition fall within certain ranges that can vary considerably from one protein to another. Folding occurs cooperatively over a narrow range of conditions, although not as sharply as a first-order phase transition in a macroscopic system. Intermediates in folding are hard to detect for many proteins. Although the denatured state is highly disordered, the folding process requires only an amount of time that allows each molecule to sample some 10^{10} conformations, far fewer than the total possible. In fact, if each molecule of an entire mole of a 100-residue protein sampled that many conformations, this is still far fewer than a rough estimate of 3^{100} total possible conformations.

The theoretical side of protein folding likewise has an enormous literature stretching over many years. The major problems addressed by statistical mechanical theories of protein folding are the causes of cooperativity, lack of folding intermediates, the high rate of folding, the ability to converge on the native from so many starting conformations in the unfolded state, and the pervasiveness of secondary structure. Much of the work is based on the statistical mechanics of polymers, which was originally developed to describe average properties of disordered systems, and then this has been adapted to globular proteins, where only very specific sequences fold to unique, compact conformations that depend in some complicated way on the sequence. As a sampling of some of the major ideas that have come from analytical statistical theories, Karplus & Weaver[2] showed how their diffusion-collision model could account for the

Grant sponsor: Vahlteich Research Award Fund (College of Pharmacy, University of Michigan); Grant sponsor: NSF; Grant number: DBI-9614074.
*Correspondence to: Dr. G.M. Crippen, College of Pharmacy, University of Michigan, Ann Arbor, MI 48109–1065. E-mail: gcrippen@umich.edu
Received 25 November 1997; Accepted 14 April 1998

cooperativity of folding, and it fit with the appealing idea that first secondary structure segments formed at least approximately, and then subsequently these packed together to form the native structure. On the other hand, Ptitsyn[3] advocated the view that the first step in folding was a collapse to a "molten globule" state, followed by rearrangement of the fairly compact chain and secondary structure formation. Bryngelson and Wolynes[4] suggested that for any given folding sequence, the energy surface as a function of conformation must be rugged to some degree, and the important factor required for rapid folding is that the general energy well around the native (the "folding funnel") must be deep compared to the overall ruggedness. This was based on a random energy model where the energy of each conformation was taken to be a random value from a distribution having a certain mean and standard deviation.

While current statistical mechanical theories have certainly captured some important features of the physical chemistry of real proteins, such as cooperativity of folding and rapid folding from the random coil state, their derivation requires making some broad assumptions about the average behavior of polypeptides. Clearly some simplifying assumptions must be made because neither nature nor computer has sufficient time to exhaustively explore all conformations and all sequences for even small proteins.

The idea here is to simplify an otherwise realistic representation of polypeptides by reducing chain length, number of conformational states per residue, and number of choices of amino acids until all sequences and all conformations can be exhaustively enumerated. By varying these parameters in the computationally feasible range, general conclusions can be detected and extrapolated to parameter values corresponding to real proteins. Questions to be addressed include: is the energy landscape really a funnel aimed at the native conformation? What energetic features correspond to rapidly folding sequences, for example the energy difference between the native and the mean nonnative conformations? Is there a general folding mechanism for all folding sequences, or do some proceed by a recognizable pathway while others have innumerable routes?

METHODS

Conformations

Each residue in the polypeptide chain is represented by the five nonhydrogen atoms of an alanyl residue, so that the sidechain is indicated only by the C^{β} atom, regardless of the residue type. Peptide bonds are taken to be planar and trans, and all bond lengths and angles are fixed at standard values[5] without any special treatment of glycine or proline.

Fig. 1. Conformation BHBGBGA, the compact native conformation of sequence LKLPL SFPSA LFKIL NNALK LPLSF PSNPP CEKIM, drawn as a schematic ribbon diagram (UCSF, MidasPlus).

The only remaining variables are the ϕ and ψ dihedral angles for each residue, but if we were to adopt a rotational isomeric model allowing only a three-way choice of helix, extended, or coil conformational state for each residue, we would have too many conformations for even a 30-residue chain.

In order to keep the total number of conformations down to a manageable level while still allowing protein-like folded states and a reasonable random coil state, we take the $\phi\psi$s for whole contiguous segments of chain to be those given in one of several building blocks. One satisfactory set consists of ten building blocks, namely a 20-residue segment of perfect α-helix ($\phi,\psi = -57°, -47°$) denoted by A, an 8-residue segment of perfectly extended β-strand ($\phi,\psi = -129°, 124°$) denoted by B, and eight different 2-residue turn segments, C-J. The lengths of the α and β blocks were chosen to give similar end-to-end distances appropriate to small globular proteins. Turn $\phi\psi$s were not taken from crystal structures, but rather from a grid search for those values that gave self-avoiding and sometimes compact conformers ACA, ACB, and BCB. Each conformation consists of alternating α/β and turn blocks until the desired chain length has been reached, even though that might come in the middle of a block. For 35 residues there are 762 self-avoiding conformations, running in alphabetical order: ACBCA, ACBCB, ACBDA, ..., BJBJBIB, BJBJBJA, BJBJBJB. A compact one of these, BHBGBGA, is shown schematically in Figure 1. Allowing successive turn blocks produces too many conformations because they are so short. Allowing successive AA or BB blocks biases the set of conformations toward overly extended structures. Otherwise, all combinations of blocks are generated, and those with steric overlaps are discarded.

This way of generating a finite set of conformations meets several goals. The number of conformers grows rapidly with chain length l, but is neither ridiculously small nor infeasibly large in the size

TABLE I. Sets of Conformations Generated by Ten $\phi\psi$ Blocks

No. residues l	No. self-avoiding conformations	No. compact conformations	σ^2 (l)
15	13	1	3.3
20	69	0	3.7
25	104	1	3.5
30	488	0	4.6
35	762	3	4.1
40	3,358	11	4.6
45	5,430	30	4.4
50	22,544	169	5.1
55	38,359	186	4.8
60	151,109	621	5.3
70	1,009,345	1,789	5.6

range of very small proteins (see Table I). There is a small but nonzero fraction of conformers that are as compact as real native proteins. We take as a compactness criterion that the radius of gyration is no more than 30% greater than the minimal observed value.[6]

$$r_{gyr,min}(l) = -1.26 + 2.79l^{1/3} \quad (1)$$

These compact structures even resemble real folding motifs, including helical bundles, β sandwiches, and α/β. Finally, the full set of conformers, including those with long-range steric overlaps, corresponds to polyalanine under Flory Θ conditions, and therefore the characteristic ratio

$$\sigma^2(l) = \frac{\langle(\mathbf{r}_1 - \mathbf{r}_l)^2\rangle}{3.80^2\,(l-1)} \quad (2)$$

for long chains ($l \geq 40$) should approach the experimental value[7,8] of 9 ± 1. The numerator in the definition of the characteristic ratio is simply the mean square end-to-end distance (in Å) for a chain of l residues. Since Table I is the result of a coarse and nonuniform sampling of all conformations, the characteristic ratio is not a smooth function of l, but it extrapolates to roughly 9 at somewhere beyond 130 residues.

Conformation Space

It is all very well and good to draw vague diagrams of energy as a function of some axis labelled "conformation," but here we need to construct a well-defined conformation space so we can discuss the smoothness of the energy surface and whether it really funnels down to the native conformation. Of course, what the energy surface looks like depends to some degree on how conformations are parameterized. Most work has concentrated on comparing pairs of conformations (of the same chain length), such as the ubiquitous RMSD, the root mean square distance between corresponding C^αs after optimal rigid superposition of the two structures. We have refined this idea in our ρ measure of conformational similarity, which is based on RMSD but compensates for its misleading biases depending on the size of the structures being compared.[9] Then $\rho = 0$ for identical conformations, $\rho < 0.5$ for obvious visual similarity, and $\rho = 2$ for maximally dissimilar conformations. What is not widely appreciated is that any such measure of similarity, based on optimal superposition, cannot be used to construct a Euclidean conformation space. That is, there is no ordinary Cartesian space $\mathbb{R}^n$ of any dimension n such that conformations are points and RMSD or ρ is the distance between them. The underlying reason is that we can optimally translate and rotate structure B onto structure A, and C onto A, but that is not in general the optimal superposition of B onto C. What is needed instead is a way to uniformly position all structures and then abstract the few most significant parameters describing them so that n is kept small.

Of course opinions vary as to what are the most significant conformational features. Shakhnovich and coworkers have emphasized the fraction of interresidue contacts in a nonnative conformation that are also seen in the native, Q in their notation.[10] This assumes a particular choice of sequence and energy function, as well as a global search for the native conformation. Here we want to separate conformational and sequence considerations as much as possible, so we will focus on the overall chain fold as the most important feature, employing the same methods we used in an earlier analysis of all possible protein folds,[11] namely the discrete cosine transform (DCT). Like the Fourier transform, the initial terms of the DCT are large and correspond to the C^α trace at low resolution, while later terms tend to be small in magnitude and correspond to finer details, such as the winding of the chain in a helix. We convert conformations into coordinates of points in a 12-dimensional conformation space according to the following procedure. Every conformation is translated so that the centroid of its C^α atoms is at the origin. Then let the first principal axis, $\boldsymbol{u}_1$, be the normalized vector to the centroid of the first third of the chain. The second principal axis, $\boldsymbol{u}_2$, is the normalized component orthogonal to $\boldsymbol{u}_1$ of the vector to the centroid of the last third of the chain. Finally, $\boldsymbol{u}_3 = \boldsymbol{u}_1 \times \boldsymbol{u}_2$, as in Figure 2.

All $3l$ terms of the full DCT for each conformation are calculated in its respective principal axis coordinate system, and those 12 terms that have the greatest range over all conformations are kept, in order of their ranges. Generally, the first coordinate of a conformation in this space is the DCT term for the slowest variation down the chain of the position

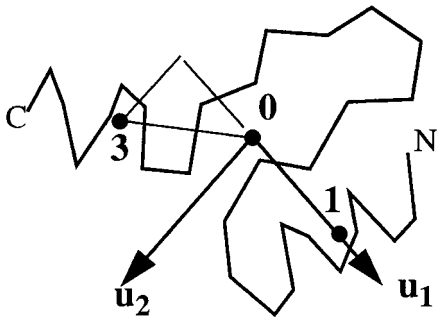

Fig. 2. Standard positioning of a protein conformation. Point 0 is the centroid of the whole chain, 1 is that of the first third, and 3 is that of the last third. The standard reference frame consists of unit vectors, $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and $\boldsymbol{u}_3$, where $\boldsymbol{u}_3$, points into the plane of the paper in this example and is not drawn.

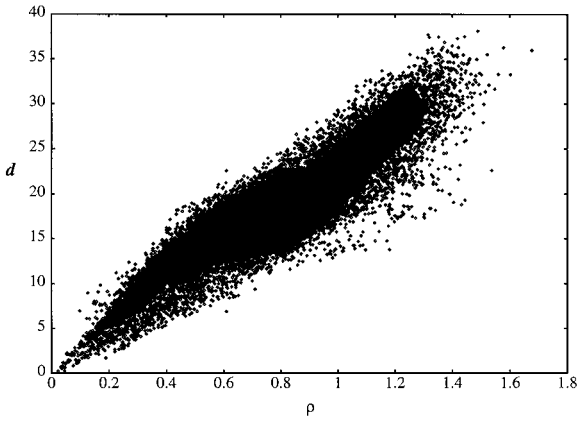

Fig. 3. The correlation of overall conformational similarity (ρ) for all pairs of the 762 structures of 35 residues vs. the distance (*d*) in the abstract conformation space of DCT coefficients between the corresponding pairs of points. (For clarity, only 1/20 of all dots are drawn, but the overall scatter is still well represented.)

along the axis of greatest elongation. Keeping only 12 terms for 30–60 residues means the finer details of residue-residue packing tend to be overlooked, but if two conformations are close in this space, no great rearrangement of the chain is required to convert from one to the other. In fact, Figure 3 shows that ordinary Euclidean distances in this space have a 90% correlation with ρ, since both measures emphasize similarity in overall chain fold. (Incidentally, rotating each conformation to its more customary inertial principal axes gives rise to some pairs of conformations that are close in ρ but distant in DCT space because some of the axes of one conformation are reversed compared to those of the other.) Furthermore, it can be shown that points near the origin correspond to compact conformations, i.e. those with small radius of gyration.[11] Figure 3 also shows that our conformational ensemble is a broad sampling, in that there are some pairs of conformations that are nearly maximally dissimilar ($\rho = 2$).

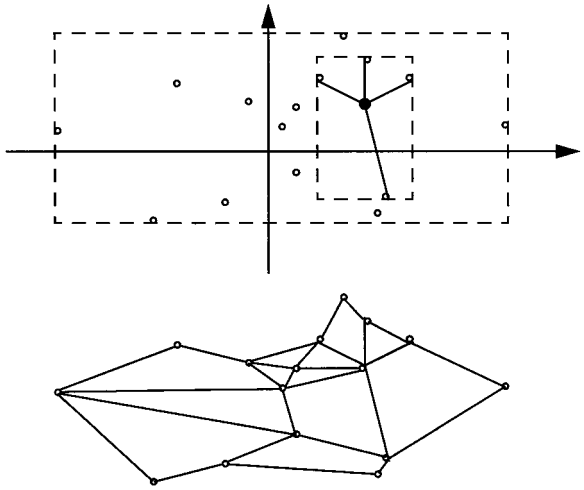

Fig. 4. Determination of the neighbors of each point scattered across the plane. Contracting boxes are indicated by dashed lines, and the neighbors thus determined are joined by solid lines. Above: construction of the immediate neighbors of the one point drawn as a solid dot; below: the complete set of mutual neighbors for all points.

Now our ensemble of conformations has been reduced to a scattering of points in a 12- dimensional space, and this scattering is in general not very uniform. Later when we calculate the kinetics of protein folding, we will assume that any one conformation will convert at various rates to its nearest neighboring conformations. The apparent kinetics therefore depend on the working definition of neighbors. At one extreme, if all conformations are neighbors of all others, then most sequences will fold rapidly because there are never kinetic bottlenecks. At the other extreme of very few neighbors, conformation space may break up into two mutually inaccessible parts because there are no kinetic pathways between the two, or at least there will tend to be many slow folding sequences because there are so few pathways leading to the native. In lattice simulations of protein folding, two conformations are taken as neighbors if there is some local chain perturbation, such as corner flips or crankshaft moves, that transforms one into the other.[12] For collections of points in space, there are a number of ways one can define neighbors, and each definition has associated an algorithm for determining the neighbors of all the points.[13] Because some appealing definitions are computationally expensive for many points and high dimensions, we have adopted the following expedient approach. It at least ensures that all pairs of points are somehow connected together by a sequence of neighbor links, and if point A is a neighbor of B, then B is a neighbor of A. These two conditions assure kinetic accessibility of all states and microscopic reversibility. Figure 4 shows how the algorithm works in two dimensions. First observe that the coordinate axes were chosen such that the scat-

ter along the first axis is greatest. This is the horizontal axis in the figure. If the points are sorted according to their first coordinate, neighboring points tend to be nearby in the sorted list, which greatly speeds the algorithm for many points. Next, there is always a minimal box having sides parallel to the axes that encloses all points. Consider the interior point with all its neighbors drawn. If there is another point within the current box to the right (left, above, or below) the central point, contract the corresponding side of the box to that interior point. When the box cannot be further contracted, the points on the faces of the box are the neighbors of the central point. In n dimensions, there are at most $2n$ neighbors of the central point at this stage (fewer if the central point was on the perimeter of the scatter). After this process has been applied to all points, it is possible that A has B as a neighbor, but not vice versa. In such cases, A is added to the list of neighbors of B, so that in the end, some points may have more than $2n$ neighbors or as few as one.

This algorithm gives satisfactory results, for example on the 762 conformations of 35 residues. The mean number of neighbors for a conformation is 9.95, implying that many structures are on the perimeter of the cloud in 12-dimensional space so that they have many fewer than the 24 neighbors one would expect in the interior of the distribution. The mean distance d in conformation space between neighbors is 13.4, which according to Figure 3 translates to an average ρ of around 0.5, which in turn means fairly apparent similarity by eye. However, the maximal d between neighbors is 40.7, so that in rare instances, very isolated conformers in this uneven scattering are assigned an extremely dissimilar neighbor. If our sampling of conformation space was more dense, there would not be such distant neighbors. Instead, they would interconvert via a series of closer intermediates that our current sampling lacks.

Sequences

We have experimented with both hydrophobic/polar sequences of just two residue types and random sequences having all 20 residue types represented in the observed frequencies of occurrence.[14] In any event, the sequence possibilities are exhaustively generated by stringing together blocks of residue types from a set of choices, just as conformations were generated. However, there is no necessary connection between the lengths of the sequence and structure blocks. Random sequences tend to favor native conformations having rather low melting temperatures (see the section on equilibrium thermodynamic properties, below). Given a set of conformations, we have found better sequences by a rudimentary genetic algorithm where the fitness function f to be maximized is the calculated statistical weight of the native structure $\boldsymbol{R}_i$ at a moderately high temperature $T_0 = 50$ for a sequence $\boldsymbol{S}_j$, without specifying in advance which structure should be the native.

$$f(\boldsymbol{S}_j) = \max_{\boldsymbol{R}_i} \left[\frac{\exp\left(-\dfrac{E(\boldsymbol{R}_i, \boldsymbol{S}_j)}{T_0}\right)}{\displaystyle\sum_k \exp\left(-\frac{E(\boldsymbol{R}_k, \boldsymbol{S}_j)}{T_0}\right)} \right] \qquad (3)$$

Starting with a small initial population of random sequences, random point mutations are tried, always keeping any improvement. The goal is not to locate the very best sequence, but merely to have some that are modestly stable thermally. This is also not an attempt to design a sequence that folds to a given target structure. For 35 residues, our mildly "designed" sequence was LKLPLSFPS NPPCEKIMA ALFKILNNA CRVCPPAP, which we then divided into four blocks of 9, 9, 9, and 8 residues. If we denote the blocks by a-d, then aaaa has 36 residues, the first 35 of which constitute the first sequence choice. However, dddd has only 32 residues, so we must add on one more block, say ddddа, to get a string that can be truncated to 35 residues. In all, the combinations aaaa, aaab, ..., ddddc, ddddd, make up a total of 457 rearrangements of four and five blocks, resulting in our full set of 35 residue sequences.

The factor linking sequence and conformation is the energy function. Ideally, the "energy" function would be an accurate estimation of the Gibbs' free energy of the solvated polypeptide in very dilute solution, averaging over solvent and sidechain configurations for the given backbone conformation. This is still an unreliable and extremely lengthy calculation, but on the other hand, a detailed molecular mechanics force field treatment for the polypeptide alone would also be a bad approximation to the free energy. In the interests of speed and simplicity, we instead take the view that the energy could be a nearly arbitrary function of sequence and structure that determines the effective temperature scale, the effective ionic strength and pH, and decides which sequences fold to which structures. Given that energy function, we want to learn the distinguishing characteristics of proteins that fold rapidly to stable native conformations. This is a less demanding objective than requiring the energy function to favor the crystal structures of naturally occurring sequences. As a not implausible choice—but certainly not a close approximation to the true free energy—we have used our potential function[15] that does correctly distinguish between the native and many nonnative conformations for many folding

sequences. This function matches our choice for representing the polypeptide chain because it depends only on the positions of the backbone heavy atoms and the C^{β}, regardless of residue type. Because the energy function is taken to be independent of temperature and solvent composition, only heat denaturation can be simulated, not cold denaturation or urea denaturation.

Equilibrium Thermodynamic Properties

At this point, we have a large number of self avoiding conformations $\{\boldsymbol{R}_i(l)\}$ for various chain lengths l and similarly many sequences $\{\boldsymbol{S}_i(l)\}$ for those same chain lengths. Connecting these is our arbitrarily chosen energy function $E(\boldsymbol{R}_i(l), \boldsymbol{S}_j(l))$ that we take as defining the unit of absolute temperature T. This amounts to setting Boltzmann's constant $k_B = 1$. Then the partition function for any sequence is

$$Z(\boldsymbol{S}_j(l)) = \sum_i \exp\left(-\frac{E(\boldsymbol{R}_i(l), \boldsymbol{S}_j(l))}{T}\right), \tag{4}$$

and hence the statistical weight of any particular conformation is

$$P(\boldsymbol{R}_i(l), \boldsymbol{S}_j(l)) = \frac{\exp\left(-\dfrac{E(\boldsymbol{R}_i(l), \boldsymbol{S}_j(l))}{T}\right)}{Z(\boldsymbol{S}_j(l))}. \tag{5}$$

From the partition function, we can calculate any equilibrium thermodynamic functions we want. Suppose for the moment that the set of conformations is such a coarse sampling of all (smoothly variable) conformations that each member of the set can be viewed as "substantially different" from all the rest. As $T \to 0$, $P \to 1$, for the conformation of globally minimal energy for that sequence. We take the melting temperature, T_m, to be the point where $P = 0.5$ for that conformation. Let ΔT be the width of the melting transition, defined by the temperature interval from $P = 0.9$ to 0.1 (see Figure 5). Since the transition sigmoid is not necessarily symmetric, it is possible to have $\Delta T > T_m$. Each sequence may have in general a different "native" conformation of globally minimal energy, and T_m and ΔT may vary greatly with sequence. Any sequence having adequately high T_m and small ΔT will be referred to as a "folding sequence" and its lowest energy structure as its corresponding "native conformation."

On the other hand, one might view the sampling of continuous conformation space to be dense enough that the immediate neighbors of the native conformation should be included as small fluctuations around the global minimum. Then T_m would be the temperature where the sum of statistical weights of the native and its neighbors is 0.5.

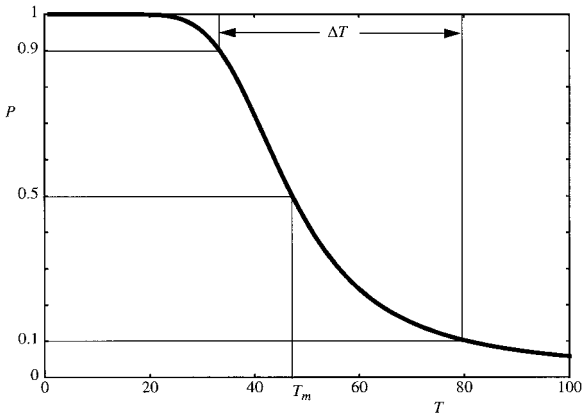

Fig. 5. Typical melting curve (T = absolute temperature, P = statistical weight of the native) for a 35-residue sequence having a high melting temperature ($T_m = 47.1$) and a relatively cooperative denaturation ($\Delta T = 47.2$).

Kinetics

Typically in cubic lattice studies, the rate of folding is equated with the mean number of Metropolis Monte Carlo iterations required to reach the native from a random starting conformation.[10] Here we instead set up the sparse set of linear differential rate equations for the unimolecular reactions between each conformer and its neighbors, where the rate constant for going from conformation $\boldsymbol{R}_i$ to $\boldsymbol{R}_j$ is taken to be

$$k_{ij} = \begin{cases} CT \exp\left(-\left(E(\boldsymbol{R}_j) - E(\boldsymbol{R}_i)\right)/(k_B T)\right) & \text{if } E(\boldsymbol{R}_i) < E(\boldsymbol{R}_j) \\ CT & \text{otherwise} \end{cases} \tag{6}$$

for some constant C, in accord with standard absolute rate theory.[16] This equation assumes that the rate constant is independent of the distance between the two conformations, and it assumes no activation energy for moving from one conformation to one of its neighbors, although in the broader landscape some intermediate conformations represent the transition states between different energy minima. Then the kinetics of conformational change can be expressed as a system of first order rate equations, where c_i is the time-dependent concentration of $\boldsymbol{R}_i$, and the sums run over the neighbors $\boldsymbol{R}_j$ of $\boldsymbol{R}_i$.

$$\frac{dc_i}{dt} = -\sum_j k_{ij} c_i + \sum_j k_{ji} c_j \tag{7}$$

Taking the initial populations of each conformational state to be equal, i.e. the high temperature random coil state, we crudely integrate the rate equations to get the populations of important states as a function of time, where the time scale depends on the arbitrary value of C. After a large number of integration steps taken at the temperature corre-

sponding to the native $P = 0.9$, indeed the relative concentration of the native state is 0.9, independent of time step size. As a general measure of the rate of folding, we use the relative concentration, F, of the native state after a fixed number of moderately small time steps, say 100. Sequences folding to relatively unstable native conformations do so slowly because of the preexponential factor of T in the expression for the individual rate constants. On the other hand, a sequence having a high melting temperature may still fold slowly due to being caught in kinetic traps.

Usually when people discuss kinetic mechanisms, it is in terms of the predominant pathway from some initial state to some final state. The pathway is often characterized in terms of a sequence of intermediate states. This is not really appropriate for protein folding because the initial unfolded state is a very disordered, more or less random coil, and there may be many different pathways leading to the native (and its immediate neighbors). Consider the abstract graph in Figure 4, where the nodes are conformations corresponding to points positioned in our 12-dimensional space, and the edges are bidirectional kinetic pathways between neighbors. At every step in the integration of the rate equations (equation 7), protein concentration flows up and down the edges, and we note the net summed flow along each edge over the whole simulation. Two neighboring conformations at equal energy would be in dynamic equilibrium but would have net zero flow between them. Important folding mechanisms would correspond to a sequence of edges having high flow running down to the native state. An extremely diffuse folding mechanism would correspond to small flows along many edges leading to the native.

RESULTS

Thermodynamics

For 35-residue chains we find a wide range of T_m, from nearly 0 to 53.1 (in absolute temperature, but arbitrary degree size determined by the potential function and not to be equated with degrees Kelvin), depending on the sequence. It so happens the designed sequence had $T_m = 52.9$, so rearrangement of sequence blocks found a thermally more stable protein. In general ΔT is uncorrelated with T_m, except that always $\Delta T > T_m$. Figure 5 shows a typical denaturation curve for a sequence having a relatively high T_m and low ΔT, where the sigmoid is steeper below the melting temperature than above it. This compares poorly to experimental thermal denaturation of small, stable, globular proteins,[17] where T_m may be 300–350 K, and often ΔT is only 10 K. Taking $\Delta T/T_m$ as a measure of cooperativity of the thermal denaturation transition, one can start with equation 5 and show that the smallest ratio (most cooperative) is obtained when all the n_{non} nonnative conformations have the same energy, E_1, where $E_1 - E_0 = g_0$ is the famous gap in the energy spectrum between the native conformation and the lowest nonnative one. While T_m increases with increasing g_0, the *ratio* $\Delta T/T_m$ is independent of it in this limiting case of just two energy levels. To put it precisely,

$$\text{minimal}\,\frac{\Delta T}{T_m} = \frac{2(\ln 9)(\ln n_{non})}{(\ln n_{non})^2 - (\ln 9)^2} \qquad (8)$$

where the ln 9 terms come from our definition of ΔT running from the temperature having 90% native to 10% native. This derivation is confirmed by our computational result that there is a clear positive correlation between g_0 and T_m, but no significant correlation between g_0 and ΔT over a set of 457 sequences for 35-residue chains (Figure 6). This contrasts with the conclusion of Sali et al. that "a sparse [energy] spectrum with a large [g_0] leads to a cooperative curve."[10] Instead, a sharper transition is obtained when not only are all the nonnative conformations at the first excited state, but the degeneracy is high, i.e. the number of nonnative conformations, n_{non}, is large. Since according to equation 8 the minimal ratio is roughly proportional to 1/log n_{non}, our restricted conformation space exhibits less cooperativity than real proteins have. For our set of 762 conformations of a 35-residue chain, the best possible ratio is 0.74, compared to the best observed values of 1.09, so it is conceivable that some other sequence would show better cooperativity. For example, if $g_0 = 351.6$, then $T_m = 53$, and the most cooperativity would occur if all $n_{non} = 761$ nonnative conformations had energies just 351.6 above the native. Then the ensemble of conformations would be 90% native at $T = 40$, 50% at 53, and 10% at 78.7, so that the melting curve is clearly sigmoidal but somewhat broader at the high temperature end, as in Figure 5.

In contrast, the melting of a very cooperative, two-state, single domain protein can be simulated by

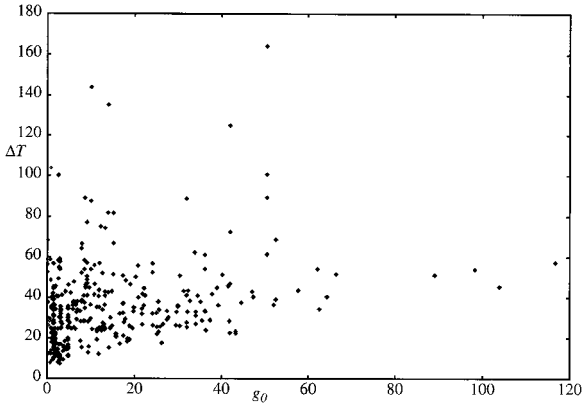

Fig. 6. There is no correlation between the energy gap (g_0) from the native to the best nonnative and the sharpness of the melting transition (ΔT).

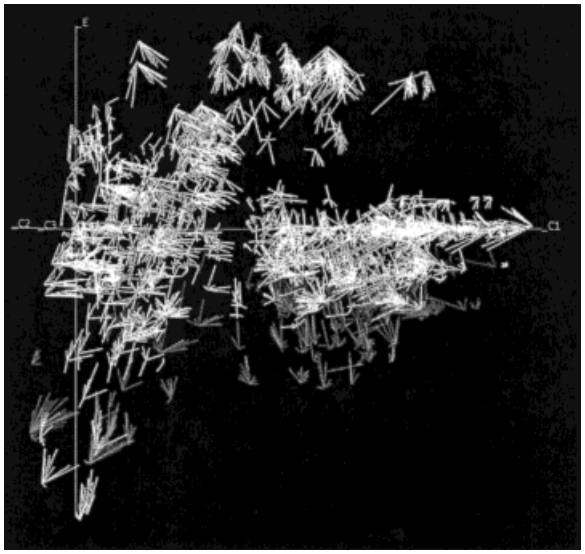

Fig. 7. Conformation/energy space for the fastest folding sequence of 35 residues. The vertical axis represents conformational energy (E), and the approximately horizontal axes correspond to the 12 dimensions of the conformational parameters, the most important one being labelled C1. Then each conformation is drawn as a starburst with rays extending a tenth of the way toward its nearest neighbors. Conformations belonging to the watershed of the global minimum are colored white, and the other 72 minima are given a variety of different shades of gray.

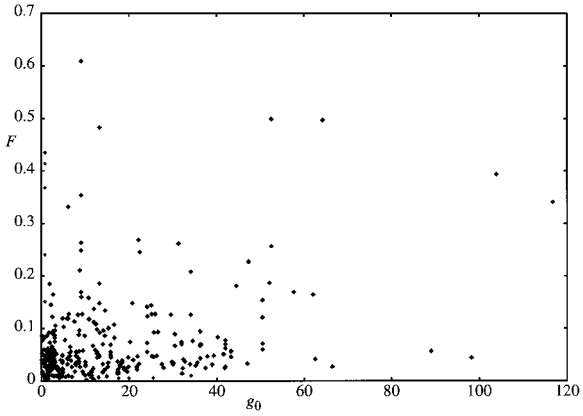

Fig. 8. There is no correlation between the rate of folding (F) and the energy gap (g_0) from the native to the best nonnative.

exactly the same calculations starting with $g_0 = 53995$ and $n_{non} = 10^{67}$. Then there is 90% native at $T = 345.2$, 50% at 350, and 10% at 354.8, so that ΔT is less than 10 K and the sigmoid is very symmetric about T_m. Clearly the width and asymmetry of the thermal denaturation curves in our model arise directly from the great restrictions placed on the total number of conformations. Of course, all the above discussion focuses on the chain entropy of unfolding, and takes no account of entropic changes of the solvent upon unfolding the polypeptide chain, beyond what may be implicit in the potential function.

The picture is much the same when the native state is considered to be the global minimum conformation plus its immediate neighbors. Denoting the melting temperature of this enlarged native state by $T_{m,nbr}$, we always find $T_{m,nbr} > T_m$, sometimes by as much as 35 degrees. The largest observed $T_{m,nbr}$ was 75.

When conformations for a given sequence are sorted by energy, much has been said of the size and location of large energy gaps. Here we find that the largest energy gap occurs either between the lowest few conformations or very high on the scale, above 600th place out of 762 total conformations. Apparently either some of the most energetically favorable conformations can be changed only by a loss of many favorable interactions, or an unfavorable conformation can be made worse only by adding many unfavorable interactions. Mediocre conformations tend to have near neighbors on the energy scale. Having a large initial energy gap is not a necessary condition for thermal stability. It is possible, for example, to have the largest energy gap of 139 energy units at 659th place, compared to the initial gap $g_0 = 117$, and still have $T_m = 53$.

Kinetics

The fastest folding sequence having $F = 0.609$ was once again not the designed sequence, but a rearrangement of it. Figure 7 shows a view of the energies of all 762 conformations as a function of their 12-dimensional coordinates. Although there are a total of 73 different local minima, 74.7% of all conformations lie in the watershed of the global minimum, including many very high energy structures. Folding for these conformations is rapid because there is at least one monotonically descending pathway across the energy surface for each of them to reach the native state. Folding from other starting conformations requires the crossing of kinetic barriers, the mean barrier being a relatively low value of 16.1 for this sequence. Notice the native conformation is compact (near the energy axis) and rather close to another low energy conformation. This causes a low $T_m = 29$, but a relatively high $T_{m,nbr} = 64$.

There are many other sequences that fold much slower, and it is instructive to focus on one of these having $F = 0.07$. Strangely enough, it has fewer local minima (55), an even greater fraction of all conformations in the native watershed (82%), and a lower mean energy barrier for the rest, only 9.2 energy units. The problem is that another local minimum lies near in energy and conformation to the native, giving rise to low thermal stability ($T_m = 15.5$). The refolding has to take place at a rather low temperature, and the near-native minimum competes with the native for a long time.

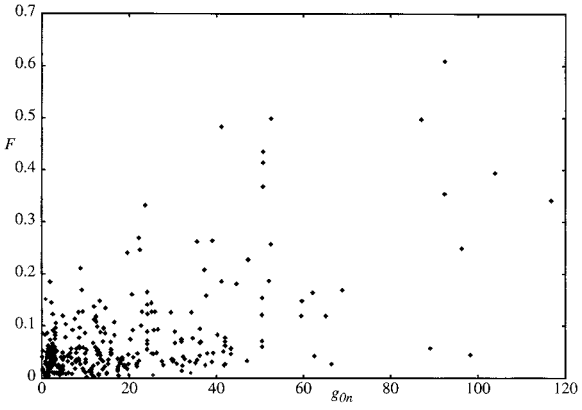

Fig. 9. There is no correlation between the rate of folding (F) and the energy gap (g_{0n}) from the native to the lowest conformation that is not a neighbor of the native.

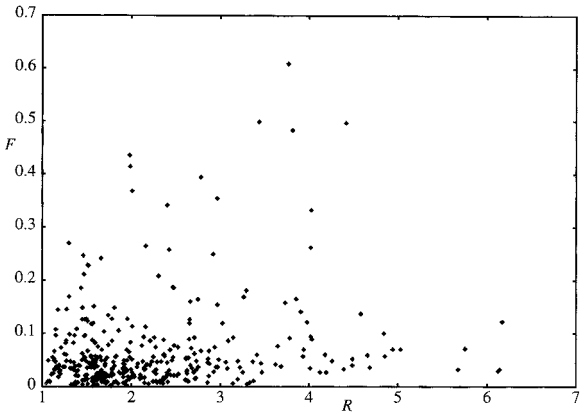

Fig. 10. There is no correlation between the rate of folding (F) and the foldability parameter (R) from spin-glass theory.

The energy gap is nearly irrelevant

Sali et al. have stated that "the necessary and sufficient condition for a sequence to fold rapidly in the present model is that the native state is a pronounced energy minimum."[10] In that work, their measure of a pronounced energy minimum was the difference in energy between the global minimum (the native conformation's energy) and the energy of the first excited state, that is, g_0 in our notation. Figure 8 shows no correlation between F and g_0. Of course, their model for a protein was a self-avoiding walk on a cubic lattice with isotropic contact vs. no contact interactions between point residues when exactly one lattice step apart. Their measure of folding speed was the number of Monte Carlo steps required on average to visit the native conformation. It is not surprising that their results should differ from ours, given the great differences in the simulation of proteins. Shakhnovich has more recently argued that the essential feature of fast-folding sequences is that there be a large energy gap between the native state and the lowest structurally distinct native state.[18,19] The closest equivalent in this work would be $g_{0n} = E_{non} - E_{nat}$, where E_{non} is the energy of the lowest conformation that is neither the native nor an immediate neighbor of the native. Figure 9 shows that having a very small g_{0n} is sufficient to cause slow folding, but beyond that there is little correlation with F. Large g_{0n} is associated with both fast and very slow folding sequences.

The folding funnel is irrelevant

Consider $\rho(E, d^2)$, the correlation coefficient between the energy of a conformation and the squared distance in conformation space from it to the native. (This is not to be confused with ρ, the measure of conformational dissimilarity.) If the energy surface has an overall funnel shape down to the native,[20] compared to a small scale roughness of the surface, $\rho(E, d^2)$ would be large and positive, where of course a correlation coefficient always lies in the range [–1, 1]. Over all 457 sequences, we find values from –0.5 to +0.7, and there is no correlation between F and $\rho(E, d^2)$. In particular, for the fastest folding sequence, $\rho(E, d^2) = 0.024$ even though obviously the native has the lowest energy in Figure 7. The reason for the low correlation is also clear from the figure, namely the balance of high and low energy conformations near the native as well as far away. The one slow folding sequence has a fine folding funnel with $\rho(E, d^2) = 0.623$ (see Figure 13).

As in the previous section, we must ask whether this result is indicative of an error in our model, or in those very different models that require a folding funnel, or whether all these simplified models of protein folding are suspect. In building a protein folding theory from spin-glass theory, it is important to remember that there is simply a single parameter that describes the roughness of the energy landscape and hence the sorts of energy barriers that must be surmounted as molecules seek the native state. If these barriers are high compared to thermal energy fluctuations, and if there is no pronounced slope of the averaged energy surface down toward the native, then folding will be slow. Roughness is equated with kinetic barriers. In our model, this equivalence of roughness to barrier is not built in, but rather some sequences have energy surfaces with kinetic barriers as a very indirect consequence of our energy function and the positions of the different conformations in our conformation space. Our model is unique in its ability to ask whether roughness should be equated with kinetic barriers, and whether an overall funnel shape of the energy surface is essential for folding. Perhaps real proteins follow neither model, but at least ours points out a possibility that has been excluded a priori from other models.

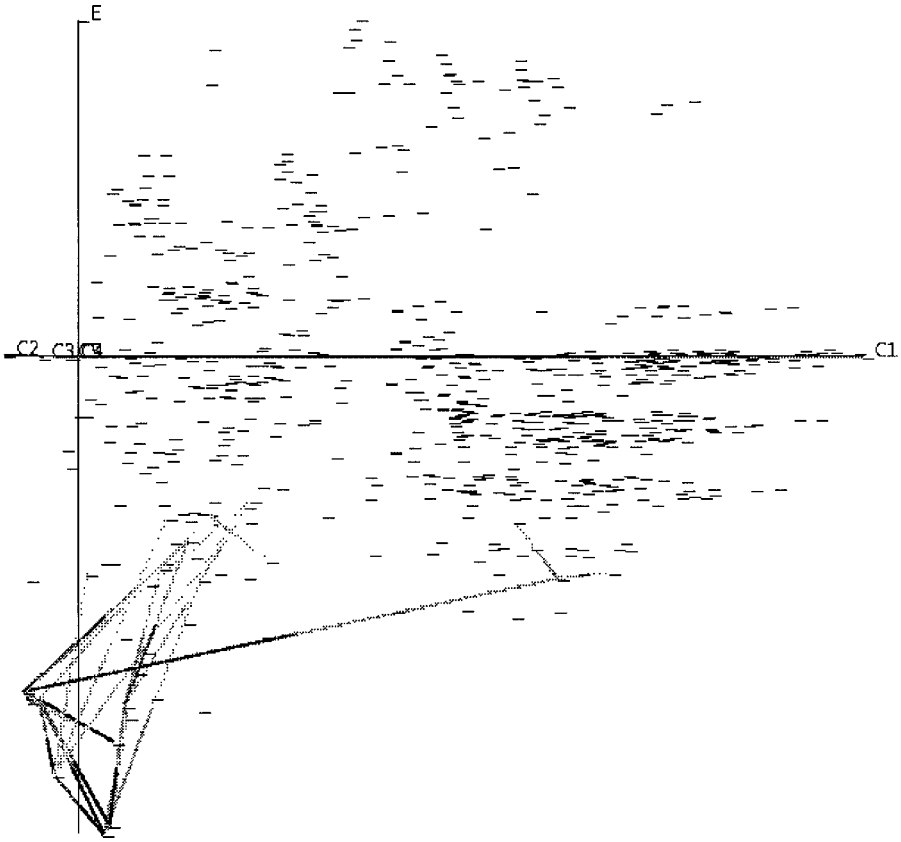

Fig. 11. Kinetic pathway of folding for the fast folding sequence in a view close to that of Figure 7. The locations of all conformations are marked by short dashes. Neighbor edges are drawn if their net concentration flows are at least 10% of that of the edge having greatest flow. Small flows are drawn faintly, large ones are drawn darker, and the flow direction along the edge goes from light gray to dark gray.

If conformation space is even so simple as the continuous, two-dimensional Euclidean plane, many different sorts of hindrances to folding can be envisaged, as recently illustrated by Chan and Dill.[21] If conformation space is a contiguous piece of $\mathbb{R}^n$, then an energetic barrier between two contiguous parts of that space (i.e., two ensembles of conformations representing two macroscopic thermodynamic states) must be an $n - 1$ dimensional "fence" of high energy that cuts the conformation space into two pieces. If conformation space is viewed as discrete, as in our model where individual conformations are connected if they are neighbors by our definition, then a barrier must be a cut set in graph theory terms, namely a set of high-energy conformations that break the conformation space network into two pieces if they are removed.

Simple reasoning by topographic analogy is very helpful. Suppose the energy surface is the altitude of a sparsely wooded park. The landscape is extremely rough because there is a great height difference between the tops of the trees and the lawn between them. Nevertheless, they pose negligible obstacles to a person strolling across the grass because he can easily go around them. The trees represent 0-dimensional bumps on a 2-dimensional surface, and we have the analogous situation in Figure 7. In contrast, a vast plain may be on the whole quite smooth, but a low fence restricts the wandering of all but the most energetic sheep.

Likewise, it is easy to understand why our model shows no folding funnel for fast folding sequences. Imagine a mountainous region having one deep lake fed by a complicated network of small streams winding throughout the region. Since there is only one lake, rainwater quickly flows off the sides of the mountains, into the tributaries, and eventually down into the lake. A hiker wandering up and down the mountains, paying no attention to the rivers, would not notice any general decrease in elevation that would lead him toward the lake because little space is taken up in river valleys, and mountains near the lake may be as high as those far away. There is a funnel in the sense that the bottoms of the stream beds have a consistent slope along their winding paths toward the lake, but these downward gradients do not consistently point toward the lake, and their elevation differences are tiny compared to the heights of the immediately adjacent mountains.

"Foldability" is nearly irrelevant

The application of spin-glass theory to protein folding concludes that fast, stable folding is associated with high values of a foldability parameter, $R = (\bar{E} - E_{nat})/\sigma_E$, which is just the separation between the native energy and the mean energy over all conformations divided by the standard deviation of the energy distribution.[22,23] Over all sequences it ranges between about 1.0 and 6.2 and shows no correlation with F (Fig. 10). The fastest folding sequence has an intermediate value of 3.8. The slow folding sequence is much better at 5.0. In terms of our landscape analogy, it doesn't matter much how rugged the mountains are or how high they are above the lowest lake. Rainwater can still run off quickly down to the lake by following streams winding around the feet of the mountains, as long as there is at least a moderate elevation drop and there aren't dams in the way.

Low kinetic barriers and large native watersheds are necessary but not sufficient

The fraction of conformations in the native watershed ranges between 5% and 85%, and certainly a low value implies slow folding. However, high values are associated with any folding rate, as seen for our fastest and example slow folding sequences. Similarly, high kinetic barriers imply slow folding, but low barriers alone aren't enough to achieve fast folding, for example the case of our slow folding sequence. Simple thermal stability, such as high T_m or $T_{m,nbr}$ is neither necessary nor sufficient, there being no correlation with F. Over all our sequences, the only way to avoid $F < 0.2$ was to demand $T_m > 29$, $R > 3.0$, more than 70% in the native watershed, and kinetic barriers less than 17 units. While no three of these four factors was sufficient to ensure at least moderate folding rates, the combination of all four was sufficient. To put it another way, none of the single factors proposed by different authors is necessary and sufficient for rapid folding to a stable native state because there are at least four relatively independent ways to prevent it. A successful sequence must avoid all of these major mistakes simultaneously.

Folding mechanisms vary greatly

Figure 11 shows the concentration flows calculated for the refolding of the fast folding sequence over the time required to achieve 80% native, at a temperature that gives 90% native at equilibrium. The high energy conformers (and some low in energy!) quickly convert to much lower energy states by such a great variety of paths that no buildup of intermediates is observed, and most edges have little flow. Finally, near the native in conformation and energy, the concentration is funneled into a few states that interconvert by a few, high-flow paths.

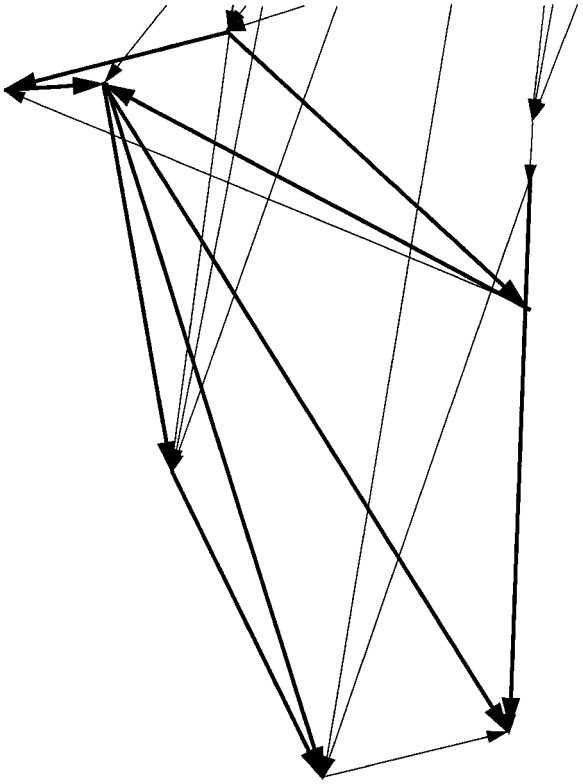

Fig. 12. A close-up view around the native of Figure 11. Flows are drawn with arrows, thicker ones denoting heavier flows.

With the exception of a cluster of conformations at the lower right, this picture is consistent with an almost unanimous rapid collapse to a few compact conformations (those that lie near the energy axis), followed by slightly slower rearrangements to achieve the native. In the close-up view in Figure 12 near the top, the four conformations connected by strong flows might be considered significant intermediates in the last stages of folding.

For comparison, Figure 13 shows the kinetics of folding for our example slow folding sequence. Note how one conformation at medium energy on the left side of the illustration is fed by a variety of other slightly higher states, and it in turn converts to a few but rather different low energy states. Neither the kinetic bottleneck nor the native conformations are as compact as the native of the fast folding sequence. Zooming in on the native in Figure 14, we see how one nonnative conformation of extremely low energy competes with the native, and interconversion between the two is through an intermediate much higher in energy. This picture is consistent with an examination of the concentrations of different conformations as a function of time, where some intermediates build up concentration and hold it for a long time.

In general, we have not yet made a broad survey over all sequences in search of consistent folding

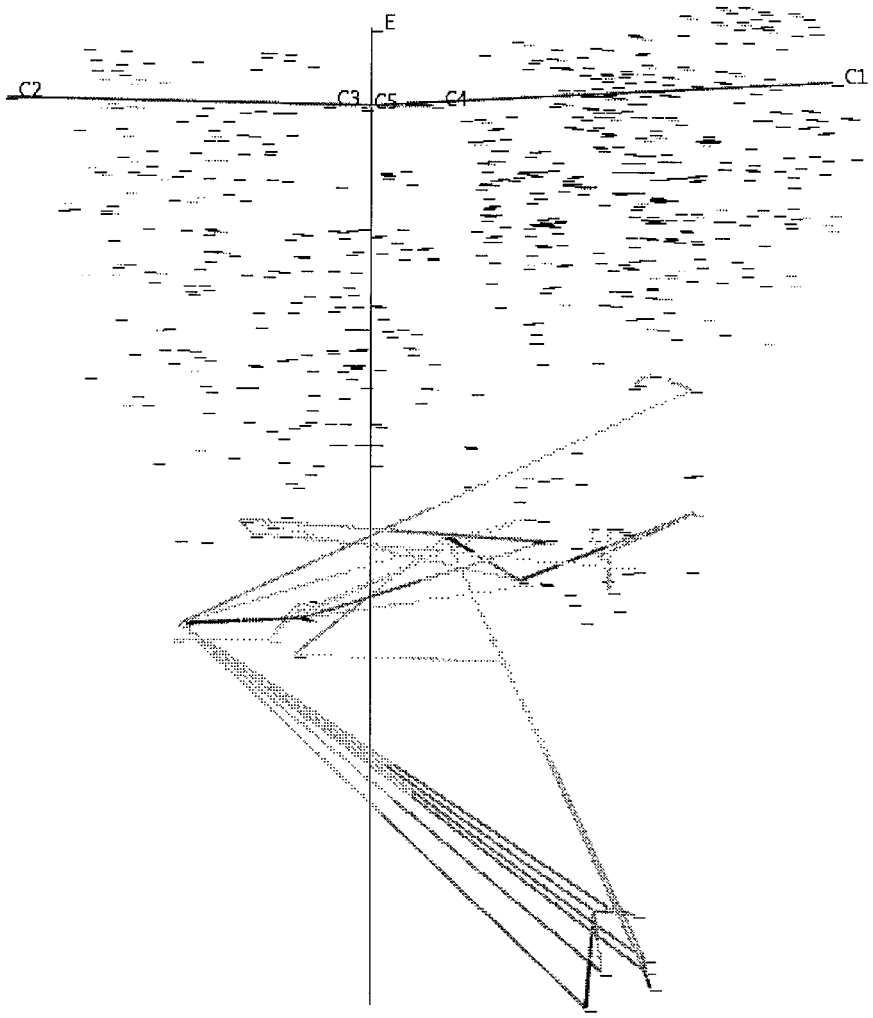

Fig. 13. Kinetics for the slow folding sequence, drawn as in Figure 11. An example of a kinetic bottleneck.

mechanisms that correlate with fast folding. Of course, it is possible but by no means certain that a denser sampling of conformations might remove some of the kinetic barriers. Given that all our conformations consist of preformed secondary structure elements joined by turns, we are unable to address the questions about whether secondary structure forms early or late in folding, or whether long-range[23] or short-range[24] contacts are key factors in folding. Aside from these concerns, just examining the folding of two sequences reveals great variety. Certainly it is inadequate to think about the folding of all proteins as a simple linear sequence of events. Neither can we characterize the folding of all our model proteins as a diffuse collapse without any sort of observable intermediates.

CONCLUSIONS

Statistical mechanical theories and computational studies of protein folding have employed several different models, each one involving some drastic simplifications. This work is no exception, but at least its simplifications are different, and its results are precise in that they do not suffer from flawed statistics, biased or incomplete sampling, or fallible searches for global optima. It has often been argued that since a model gives rise to some features of real protein folding, such as unique native conformations and cooperative folding, conclusions from that model can be immediately extended to real proteins. Our model system also exhibits many of these protein-like features, but our results disagree with most other theoretical studies. One can always argue about which model is better, but clearly it is necessary to validate a model in more detail before coming to sweeping conclusions about the principles of folding of real proteins.

There are many causes of slow folding, and no single, simple statistic so far proposed seems to capture the necessary and sufficient condition for fast folding to a stable native state in this study. When even such a simple model study as this can

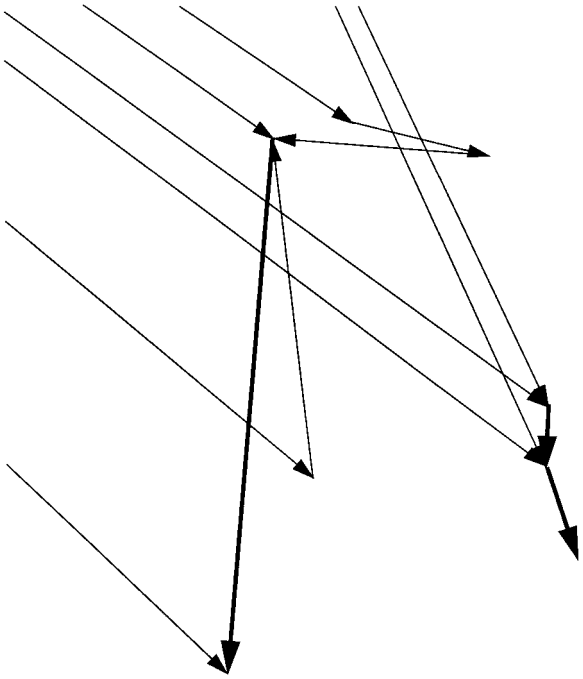

Fig. 14. A close-up view around the native of figure 13 drawn as in Figure 12. The very low energy nonnative conformation competes with the native.

produce such striking counterexamples to prevailing theories, we must realize that our mental pictures of the conformational energy surface and folding process have been oversimplified.

REFERENCES

1. Dill, K.A., Bromberg, S., Yue, K. et al. Principles of protein folding—A perspective from simple exact models. Protein Sci. 4:561–602, 1995.
2. Karplus, M., Weaver, D.L. Protein-folding dynamics. Nature 260:404–406, 1976.
3. Ptitsyn, O.B. Protein folding: Hypotheses and experiments. J. Protein Chem. 6:273–293, 1987.
4. Bryngelson, J.D., Wolynes, P.G. Spin glasses and the statistical mechanics of protein folding. Proc. Natl. Acad. Sci. U. S. A. 84:7524–7528, 1987.
5. Park, B.H., Levitt, M. The complexity and accuracy of discrete state models of protein structure. J. Mol. Biol. 249:493–507, 1995.
6. Maiorov, V.N., Crippen, G.M. Contact potential that recognizes the correct folding of globular proteins. J. Mol. Biol. 227:876–888, 1992.
7. Brant, D.A., Miller, W.G., Flory, P.J. Conformational energy estimates for statistically coiling polypeptide chains. J. Mol. Biol. 23:47–65, 1967.
8. Miller, W.G., Brant, D.A., Flory, P.J. Random coil configurations of polypeptide copolymers. J. Mol. Biol. 23:67–80, 1967.
9. Maiorov, V.N., Crippen, G.M. Size-independent comparison of protein three-dimensional structures. Proteins 22:273–283, 1995.
10. Sali, A., Shakhnovich, E., Karplus, M. Kinetics of protein folding. J. Mol. Biol. 235:1614–1636, 1994.
11. Crippen, G.M., Maiorov, V.N. How many protein folding motifs are there? J. Mol. Biol. 252:144–151, 1995.
12. Chan, H.S., Dill, K.A. Transition states and folding dynamics of proteins and heteropolymers. J. Chem. Phys. 100: 9238–9257, 1994.
13. Edelsbrunner, H. "Algorithms in Combinatorial Geometry." vol. 10, EATCS Monographs on Theoretical Computer Science. Brauer, E.W., Rozenberg, E.G., Salomaa, A. (eds.). Berlin: Springer-Verlag, 1987.
14. McCaldon, P., Argos, P. Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences. Proteins 4:99–122, 1988.
15. Maiorov, V.N., Crippen, G.M. Learning about protein folding via potential functions. Proteins 20:167–173, 1994.
16. Amdur, I., Hammes, G.G. "Chemical Kinetics." New York: McGraw-Hill, 1966.
17. Pfeil, W., Privalov, P.L. Thermodynamic investigations of proteins. III. Thermodynamic description of lysozyme. Biophys. Chem. 4:41–50, 1976.
18. Shakhnovich, E.I. Theoretical studies of protein-folding thermodynamics and kinetics. Curr. Opin. Struct. Biol. 7:29–40, 1997.
19. Gutin, A.M., Abkevich, V.I., Shakhnovich, E.I. Evolution-like selection of fast-folding model proteins. Proc. Natl. Acad. Sci. U.S.A. 92:1280–1286, 1995.
20. Onuchic, J.N., Wolynes, P.G., Luthey-Schulten, Z., Socci, N.D. Toward an outline of the topology of a realistic protein-folding funnel. Proc. Natl. Acad. Sci. U. S. A. 92:3626–3630, 1995.
21. Chan, H.S., Dill, K.A. Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. Proteins 30:2–33, 1998.
22. Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G. Protein tertiary structure recognition using optimize Hamiltonians with local interactions. Proc. Natl. Acad. Sci. U. S. A. 89: 9029–9033, 1992.
23. Govindarajan, S., Goldstein, R.A. Optimal local propensities for model proteins. Proteins 22:413–418, 1995.
24. Unger, R., Moult, J. Local interactions dominate folding in a simple protein model. J. Mol. Biol. 259:988–994, 1996.