Scaling approach to the folding kinetics of large proteins

Erik D. Nelson* and Nick V. Grishin
Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 6001 Forest Park Blvd., Room ND10.124, Dallas, Texas 75235-9050, USA

(Received 15 July 2005; revised manuscript received 11 November 2005; published 10 January 2006)

We study a nucleation-growth model of protein folding and extend it to describe larger proteins with multiple folding units. The model is of one of an extremely simple type in which amino acids are allowed just two states—either folded (frozen) or unfolded. Its energetics are heterogeneous and Gō-like, the energy being defined in terms of the number of atom-to-atom contacts that would occur between frozen amino acids in the native crystal structure of the protein. Each collective state of the amino acids is intended to represent a small free energy microensemble consisting of the possible configurations of unfolded loops, open segments, and free ends constrained by the cross-links that form between folded parts of the molecule. We approximate protein free energy landscapes by an infinite subset of these microensemble topologies in which loops and open unfolded segments can be viewed roughly as independent objects for the purpose of calculating their entropy, and we develop a means to implement this approximation in Monte Carlo simulations. We show that this approach describes transition state structures (ϕ values) more accurately and identifies folding intermediates that were unavailable to previous versions of the model that restricted the number of loops and nuclei.

DOI: 10.1103/PhysRevE.73.011904 PACS number(s): 82.37.Rs, 02.40.Pc

I. INTRODUCTION

In a recent paper [1], we described a simple experiment to unfold proteins by a kind of topological implication [2]. There, proteins were expanded mechanically along a path of steepest increase in available free space (in order to represent maximum entropy or minimal entropy loss [3,4]) so that the path of unfolding just reflected the shape of the initial (native) state. For large proteins, it was clear from inspection that certain buried or frustrated fold segments would remain folded until other parts of the molecule could unfold to create space for them to move. We were interested to know whether this "topological order" characteristic of unfolding a topological defect needs to occur in reverse for a protein to fold its native shape. We found that for many proteins, even those with very simple topologies, the mechanic unfolding paths compared well with the key events described in protein folding kinetics experiments, and the results suggested a potentially useful division of protein structures into fragments with different dynamic characteristics: namely, (i) cooperative parts in which the unfolding of each group of cross-links simultaneously supports the unfolding of all others, and (ii) frustrated parts in which unfolding two or more groups of cross-links dynamically conflict. Frustration leads to dispersion against the entropic order of contact formation or dissociation [3], and consequently, models that use contact order to describe folding may improve if the kinetics of cross-linking is somehow renormalized [5,6] in terms of these cooperative units [7].

Here we begin to investigate these ideas using a simple nucleation growth model of protein folding developed by Finkelstein and co-workers [8–12], one of our aims being to schematize this model in a way that allows more accurate treatment of larger proteins with multiple folding units [13,14]. While our previous work focused purely on topological constraints (dihedral angle, residue geometry, and topology), the model we investigate here [10] is thermodynamic and contains very limited constraints corresponding to the surface burial of residues and the entropy cost of loop closure. In this model, each amino acid occupies one of only two states—either folded (frozen) or unfolded. Each collective state of the amino acids is intended to represent a small microensemble, consisting of the configuration states of unfolded segments constrained by the frozen amino acids and the cross-links that form between them. The free energy of a microensemble, or "microstate," γ is defined as

$$F(\gamma) = \epsilon \sum_{i<j} {}^{\gamma}\delta(i,j) - T\left[(L - q(\gamma))\sigma + \sum_{p} {}^{\gamma}s(p) \right], \quad (1)$$

where in the first (energetic) part of this expression, $\delta(i,j)$ is the number of heavy atom contacts (including main-chain atoms) between residues i and j in the native crystal structure [15], and the sum $\Sigma_{i<j}{}^{\gamma}$ includes all pairs of amino acids that are frozen in γ. In the entropic part of the expression, L is the chain length, q is the number of folded residues, $\sigma = 2.3R$ is the entropy cost to freeze an amino acid, and $s(p)$ is the entropy cost to link the ends of an unfolded segment into a loop p as described in Eqs. (2) and (3) below. The entropy is approximated as if the loops are independent objects [16] (only two loops were allowed in this formula) and their interaction with folded parts of the molecule is excluded volume only.

Each microstate consists of one or more folded nuclei, or native droplets [17] decorated by whatever unfolded loops, open segments, and ends are formed by the cross-links, and can be represented diagramatically as shown in Figs. 1–3. Authors of this type of model limited the space of their sys-

*Corresponding author. FAX: 214-645-5948. Email address: enelson@spirit.sdsc.edu

 ©2006 The American Physical Society

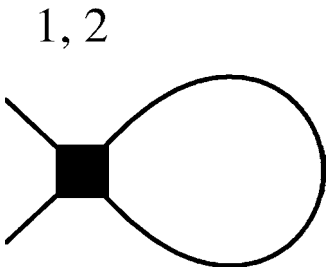

FIG. 1. Topology diagram formed by two cross-linked fold blocks $\alpha^{(1)}$ and $\alpha^{(2)}$.

tems to at most a couple of unfolded loops in order to speed computations, however, the largest proteins studied in kinetics experiments (length $L \sim 400$ amino acids) can require terms with up to $\lesssim 10$ unfolded loops [8] and even moderately large proteins can contain multiple folding domains that nucleate independently. The complexity of the diagram topologies increases rapidly with the number of unfolded segments, and estimating the entropy for all these terms is an extremely complex if not forbidding task. However, if one follows the diagram topologies out to a sufficiently large order, patterns emerge that suggest how an approximation based on diagram connectivity could be established.

To start with, the zeroth order approximation (o) is just a single folded droplet with two unfolded ends [9]. Continuing the independent loop approximation above leads to a first order approximation—(i) that of a droplet decorated by a halo [18] of unfolded loops (Fig. 1). The next anzatz could be (ii) to allow multiple droplets of the form (o) and (i) connected by open unfolded segments, then (iii) insertion of forms (o) and (i) within loops [Fig. 2(a)], and finally (iv) cross links between loops and insertions (i.e., all possible

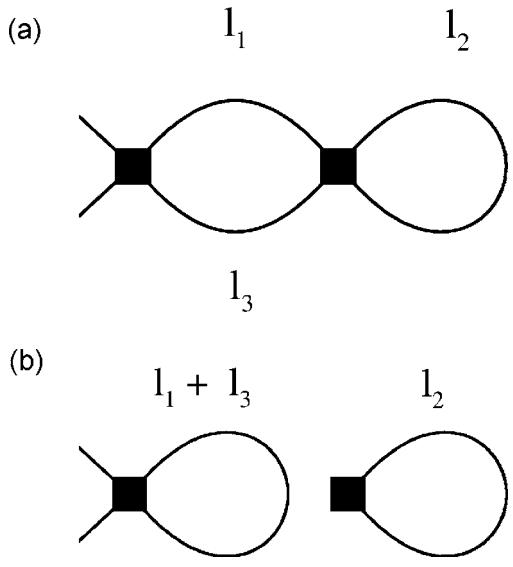

FIG. 2. (a) Simplest reducible diagram formed by two pairs of cross-linked fold blocks, $\alpha^{(1)}, \alpha^{(4)}$ and $\alpha^{(2)}, \alpha^{(3)}$ (block labels not shown) and (b) reduction to independent loops. The parameters l_n label the lengths of adjacent unfolded segments. This step favors local nucleation or zipping [3] so that either $l_1 + l_3 < l_2$ or $l_2 < l_1 + l_3$ is favored entropically.

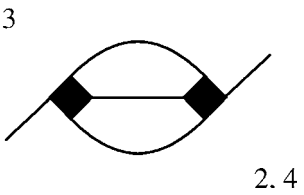

FIG. 3. Simplest irreducible diagram formed by permuting the cross connections between fold blocks in the previous Fig. 2.

diagrams). Each of these approximations corresponds to a landscape of microstates and intuitively one expects that entropy loss considerations might allow us to terminate resolution of the landscapes at a tractable level. For example, diagrams with cross links among the loops (the simplest of which is shown in Fig. 3) scale in a way that tends to conflict with the entropic order of cross linking, while in the rest of the diagrams, loops and open unfolded segments tend to scale as independent objects as in the initial approach above (we will refer to these terms as "reducible" diagrams in Sec. II).

Given that we could estimate the loop entropy for these diagrams, the first problem is computational—is there an efficient way to select them during the simulations? Here we develop a Monte Carlo algorithm to solve this problem and we use it to study about ten small to moderately large proteins including those investigated in Refs. [10,11].

Bearing in mind that Eq. (1) describes the unfolded segments of proteins as random coils, there are entropic reasons why increasingly complex diagrams starting with those in Figs. 2 and 3 should have a low frequency of occurrence in the simulations, and accordingly, we find that type (ii) diagrams dominate the examples we study. It is difficult to say to what degree this result extends to proteins since protein unfolded states are thought to be partially structured [19]. In collapsing to a partially structured ensemble (globule), part of the entropy difference between random coil and folded states is balanced by the attractive energy that causes the chain to collapse. In its present form, the model would allow this situation to be expressed only as a superposition of folded and unfolded amino acid states. And, if we were to construct an entropy approximation to fit the globule state [20] (assuming this state had nativelike structure), the energy would be scaled down in Eq. (1) to meet the entropy approximation because the model adjusts the free energy landscape $F(\gamma)$ to the condition of equilibrium between folded and unfolded states (see below). Therefore, the results of this step may be very similar to those for the random coil except that the unfolded state of a residue now corresponds to a partially structured state in the protein. In particular, although the entropy model changes, it may still stratify according to whether the boundary conditions on unfolded segments are fixed (loops) or free to move (open segments and ends) so that the two models (random coil and globule) roughly correspond. It should be noted, however, that the model has many small problems and has only been applied to proteins of about $L \sim 100$ amino acids, so it is difficult to take the coarse agreement between it and protein transition state structures as confirmation of this sort of correspondence.

In any case, recent work suggests that some of the parameters in this model, such as the persistence length [19], the form of the interaction energy [21], and also the response of the model to situations where non-native interactions [21–24] control the folding process need to be investigated before attempting to reapproximate the entropy this way. These types of investigations would have little value if the model were restricted to its original form, and indeed the present approach leads to significant improvements in the transition state structures (ϕ values [25]) over the results for protein crystal structures calculated in Refs. [10,11] precisely because of the added multiple loop and droplet terms described above. As expected, by including these more complex terms the model picks up more of the "residual" features of a protein's native structure that define its intermediates and is now able to identify intermediates [26,13] that were unavailable to restricted versions of these models [10–12].

In Secs. II–IV we now describe this approach and our results. Later we discuss some of the basic problems encountered with two-state nucleation-growth models and how these and other Gō-like models [14,26,27] may interpret the landscapes of proteins with non-native intermediates.

II. REDUCIBLE DIAGRAMS

Each collective state of the amino acids can be represented by a list of folded and unfolded blocks,

$$\rightarrow \alpha^{(n)}(i,j) \rightarrow \alpha^{(n+1)}(k,l) \rightarrow \alpha^{(n+2)}(m,n) \rightarrow , \qquad (2)$$

where arrows stand for the unfolded parts and the index pairs (i,j), etc., number the end points. The fold blocks, $\alpha^{(n)}(i,j)$, are cross linked according to Eq. (1) and the spatial connectivity of nuclei, loops, open segments and so on can be represented by diagram as shown in Fig. 1.

It is simple to enumerate all the diagrams allowed for a given number of folded blocks, and in doing so, one finds the set of approximations listed above. Again, type (ii) diagram topologies can be constructed by bonding the protruding ends of simpler type (o) and (i) diagrams like Fig. 1 together in sequence. For this class of diagrams it seems reasonable to continue the scheme defined in Eq. (1), in which loops are independent, and the segments joining the nuclei (open unfolded segments) are treated the same way as ends. In other words, in this approximation type (o) diagrams are reducible to open unfolded segments of length equal to the sum of unfolded amino acids on their ends.

All that occurs in this step is that unfolded segments are divided into looplike and endlike types for the purpose of calculating their entropy, but clearly this division becomes less convincing with increasing diagram complexity. For example, type (iii) diagrams include nested loops, the simplest case of which is shown in Fig. 2(a). The diagram in this figure is, topologically, almost identical to that for a nucleus decorated by two loops [one loop containing a type (o) diagram] so it would be within the approximation used in the original model [10] to approximate the loops in Fig. 2 as independent objects. However, type (iv) diagrams include terms that cannot be reduced to either loops or ends (the first nontrivial diagram of this type is shown in Fig. 3), and although it would appear as if we could interpet this diagram as, say, three ends or some fractional number of loops, this step would most likely lead to confusing results since the boundary conditions (ends of the segments) are perfectly correlated.

While we know that these more complex terms are small, it is still of interest to sample them to gain perspective on the situation of folding from a partially ordered globule, and the simplest, most consistent approximation we can take is to restrict the system to type (iii) diagrams where it still makes some sense to approximate loops and open segments as independent objects [16]. In this case, both type (o) segments and loops are considered reducible [see Fig. 2(b)] so that the whole unfolded part of the molecule is viewed as a noninteracting soup of open segments (or just one long open segment) and loops. Irreducible terms like those in Fig. 3 are not included but we still keep track of how often they are attempted during the simulations. The basic kinetic algorithm and the screening method are described in the Appendix.

III. MONTE CARLO

In each Monte Carlo attempt, an amino acid is selected at random and flipped. If the resulting diagram is reducible, a list of loops is created and used to compute the free energy difference between the initial and target microstates (steps that result in irreducible diagrams are not counted in thermal averages). The free energy is calculated exactly as in Eq. (1) except that now the entropy sum includes reduced loops.

Each loop (or reduced loop) p is closed off between two folded blocks $\alpha(i,j)$ and $\alpha'(k,l)$ by the frozen amino acids j and k and contains one or more unfolded blocs n of length $l_n(p)$ due to the possibility of insertions along its length. Let

$$l(p) = 1 + \sum_n l_n(p) \qquad (3)$$

be the effective length of the loop and let $r(p)$ be the space distance between its bracketing alpha carbons. The entropy cost to close a loop is defined as

$$s(p) = \frac{5}{2} \ln l(p) - \frac{3}{2} \frac{r^2 - a^2}{2aAl(p)}, \qquad (4)$$

where $A=20$ is the persistence length and $a=3.8$ is the chain spacing between alpha carbons. This expression describes a random coil with ends fixed to an impenetrable plane [9] that enforces the excluded volume of the droplet. Although it may improve things to consider the excluded volume of droplets joined to the ends of open segments, our intent is just to isolate the effects of relaxing the diagram restrictions.

To collect statistics, we sample by small ensembles [28] that restrict the protein to sliding windows $q \in [i-\omega, i]$ of width ω. The windows are about the size of a typical nuclear region in the largest proteins studied below. For each protein, we sample all windows in the index range $i \in [\omega, L]$ for $\sim 10^6$ Monte Carlo steps (MCS) per amino acid, per interval. The window ensembles have a precise relationship to the global ensemble as long as there are no ergodicity problems (damping) within windows (which may be the case if the

landscape is sufficiently grooved or textured). For our system, it was necessary to discard data recorded near the edges of the windows to recover the equilibrium condition, $F(0)=F(L)$. Again, as in Ref. [10], all the simulations are conducted at the condition of equilibrium between native and folded states so that ϵ in Eq. (1) is expressed in terms of temperature and F appears in units of RT.

IV. RESULTS

We report results for about ten α, β, and α-β protein topologies having both two-state and multistate kinetics ranging in size from about 50–200 amino acids [29]. In this section, we discuss the general features of these results and we compare effective ϕ values (cross-link probabilities) calculated using restricted and scaling versions of Eq. (1) for the set of five proteins studied in Ref. [10].

The cross-link probability is

$$\Phi(j) = \sum_{\gamma \in q^{\ddagger}} P(\gamma) \frac{C(\gamma,j)}{C(j)}, \tag{5}$$

where $P(\gamma)$ is the probability of occupying γ, $C(\gamma,j)$ is the number of *active* contacts [$C(j)$ the total number of contacts in the native state] with amino acid j [$C(\gamma,j)=0$ if j is unfolded]. The sum includes conformational states within the transition ensemble [4] which we select to coincide with the maximum in the free energy profile, $F(q^{\ddagger})=\max F(q)$. Since conditions can vary strongly in ϕ-value measurements, it is of interest to know how well the results agree with experiments in the immediate vicinity of the transition state. Therefore, to indicate this we also report the maximum correlation values obtained in the small window $q^{\ddagger}\pm 2$.

The results for the three nucleation-condensation-type proteins (CheY, Barnase, and CI2 [26,30,31]) are shown in Fig. 4. The correlation coefficients for all five proteins at maximum correlation (including src and α-spectrin sh3, respectively) are 0.73, 0.73, 0.61, 0.16, and 0.77. At the trantition state $q^{\ddagger}$, the correlation values are, 0.62, 0.67, 0.61, 0.16, and 0.77—a typical improvement of about 30–40% over the restricted model [32]. If we exclude 1srl, a NMR structure that is poorly described both here and in Ref. [10], the correlation coefficients we calculate at $q^{\ddagger}$ against the data set in Ref. [11] are just as accurate as the more recent (but still restricted) version of the model revised to include hydrogen atoms in the contact energy [11].

Inspection of the diagram statistics reveals that the improvements cited above are directly linked to releasing the diagram constraints. For example, while the occupation probability of diagrams with three or more loops is less than 1% at the transition states of CheY and CI2, the occupation of multiple droplet diagrams is of the same order of magnitude (10–20%) as improvements in the correlation coefficients. At the transition state of Barnase, the frequency of states with 0–4 unfolded loops is 0.030, 0.158, 0.559, 0.245, and 0.009 (and with 1–3 droplets is 0.876, 0.118, and 0.006)—again, the higher order terms are of the same order as improvements in the correlation coefficients.

Releasing the constraints allows the model to pick up transition state features that cannot be resolved by the original model. For example, the alternating α-β protein CheY has an off-pathway intermediate with helical local order [26,13]. The intermediate is thought to result from competition between the β-interior and α-helical exterior of the protein and may involve non-native interactions. We observe an intermediate in which helices are formed but not β strands (except within the nucleus) at around $q=24$. In crossing this region, the number of droplets jumps (the probability of four droplets reaching about 10%) and the spectrum of droplet sizes changes abruptly from bimodal (centered around 2 and

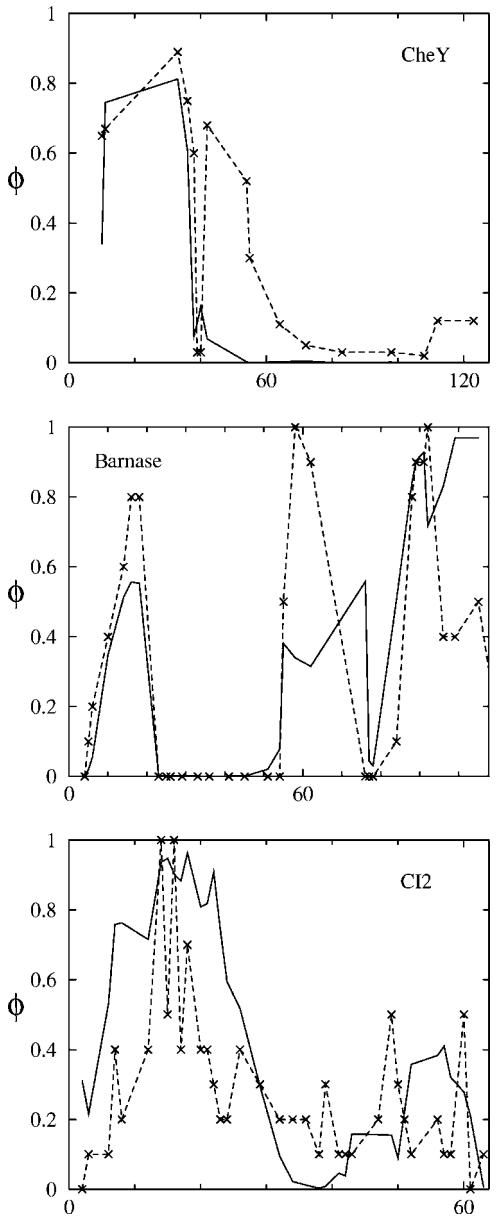

FIG. 4. Correlation of measured ϕ values and contact probabilities, $\Phi(j)$ (crosses and solid lines, respectively) for CheY, Barnase, and CI2. The solid lines correspond to the point of maximum corellation in the immediate neighborhood $q^{\ddagger}\pm 2$ of the transition state. The data is identical to that reported in Ref. [10] and the simulations are likewise conducted with segments of two amino acids. The indices in the figures (x axis) label the amino acids and the data from 1 to L.

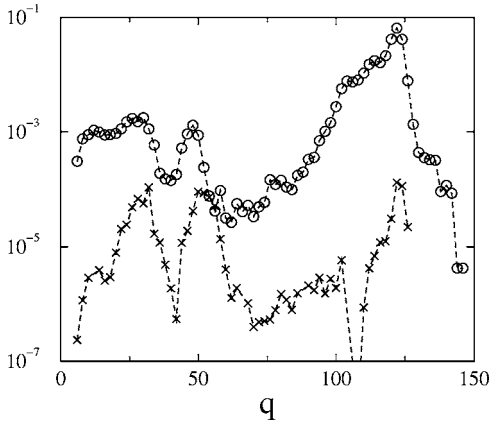

FIG. 5. Attempt frequencies for reducible conformations containing nested loops (circles) and irreducible parts (crosses) for T4 lysozyme. The relative size and magnitude of the attempt frequencies in this figure are somewhat typical of the proteins studied in this work except near the folded edge of the plot. The *occupation* probabilities of reducible nested diagrams is typically about a decade smaller than the attempt frequencies in this figure.

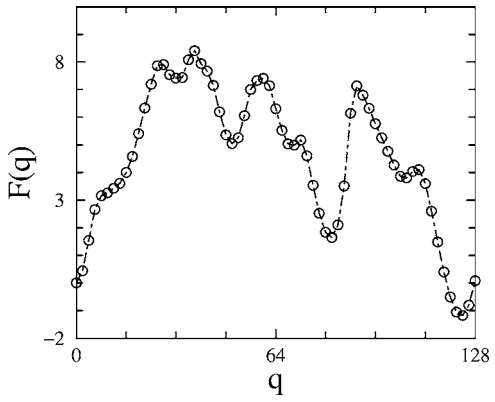

FIG. 6. Free energy profile, $F(q)$, for CheY. The off-pathway "helical" intermediate is detected near $q=24$. At the transition state, $q^{\ddagger}=38$, the occupation probabilities for states with 0–2 unfolded loops (1–3 droplets) are 0.767, 0.215, and 0.016 (0.726, 0.218, and 0.054). After the transition state, α-β layers progressively accrete onto the nucleus, and the formation of helices registers with (begins just before) maxima in the profile.

q) to unimodal (centered around 2) to bimodal before reaching the transition state. The location of this intermediate and the major features of the folding profile, while more pronounced, correspond with the simulation results of Clemente and co-workers [26].

The number of unfolded loops near the transition states of the proteins in our sample agreeswith the simple $\frac{1}{6}L^{2/3}$ estimate given in Ref. [8], but not as a strict rule. For example, across the transition state of α-spectrin sh3, the occupation of diagrams with more than two loops can be as large as 0.29. Diagrams that require reduction of loops are very infrequent during these simulations (typically $<10^5$ percent), and the *attempt* frequency for irreducible diagrams is comparable to, but typically smaller than that for nested diagrams, suggestive of the results we may obtain if we include these terms in the simulations. A more unusual example of our results for attempt frequencies is shown in Fig. 5. Again, it is difficult to predict the contribution of these terms for the case of a structured globule unfolded state, and one cannot rule out rare instances where a specific diagram like that in Fig. 2 may play an important part in nucleating a certain large protein fold. Nevertheless, there may be some hope of describing this situation more convincingly in terms of an approach like that in Ref. [20].

In closing, it may be worthwhile to list some of the problems we have so far neglected to discuss for the two-state models [12]. First, the dynamical scheme of this approach still inhibits some parallel [33] kinetic processes (specifically, independent folding of subdomains that interpenetrate, or are otherwise strongly connected in the native fold) because any pair of amino acids that are in contact in the crystal structure are also in contact if they both freeze into a folded state. This type of situation occurs maybe once (staphylococcal nuclease) out of the 20 or so proteins we have simulated so far, so it does not yet seem to present a significant limit to the usefulness of this model. Also, the entropy cost to freeze an unfolded segment or close an unfolded loop depends on amino acid composition [34,35]. Since our model does not include this dependence, it is likely that the fine scale features of our results (lengths of order several amino acids) on which differences among amino acid entropies do not average could be improved. Finally, non-native interactions can play a significant part in [21,22] and even control the folding dynamics of certain proteins [23], and Gō-like models are not, of course, able to describe these situations very accurately. Nevertheless, by simulating these types of proteins [24], we find that non-natively stabilized intermediates can leave a specific signature that is simple to explain [36] and serves as an indication of where this model is working as intended.

To summarize, although we have allowed for very complex terms within the simulations, it turns out that just releasing some of the restrictions [including type (ii) diagrams perhaps with type (o) insertions in loops] can improve ϕ-value calculations quite a bit. This is also what allows the model to identify more coarse grained features such as the helical intermediate in Che Y (see Fig. 6). But it is interesting that proteins for which ϕ-value estimates improve drastically after adding hydrogen atoms (U1A [11], for example), also improve substantially by releasing these restrictions (E. Nelson, unpublished data [37]), which shows that the model is more responsive to the fine scale features of the native structure as well. Indeed, although the constraints on loops and nuclei seem applicable to small proteins of about $\lesssim 100$ amino acids, we have seen that they do not apply to some of the smallest proteins studied by this type of model. More complex diagrams may become significant if the unfolded state is considered as a partially structured globule, but it seems unlikely that this step leads to any drastic differences in the problem for the reasons given above. In this vein, it should be noted that even high temperature unfolding simulations give accurate ϕ-value estimates, so it may be better to investigate some of the more basic problems with this model noted above. But, even in its present form, the approach here appears accurate enough to decode the

intermediates [24] of some of the larger proteins [13,14] studied in kinetics experiments.

ACKNOWLEDGMENT

We are grateful to Ken Dill for helpful comments during the completion of this work.

APPENDIX

The scaling approximation is relatively straighforward but the computational steps needed enact it are complex and may be worth explaining.

The main problem encountered in setting up the background kinetic part of the code is that diagram topologies are defined in terms of folded blocks, which are impermanent objects. As a result, it is necessary to assign each of the blocks $\alpha^{(n)}(i,j)$ a temporary (integer) identity p while it exists during the simulation. The number of possible identities is conserved, each p being drawn from a stack $p \in [0, L/2]$ when a fold block is created and returned to the stack when the block dissolves. This constraint makes it possible to define a matrix, $m(p,q)$, whose rows and columns are indexed by the identities $[0, L/2]$ to record the current number of cross-links between each pair of fold blocks, so that the topology of the diagram corresponding to a given state of folded blocks can ultimately be accessed.

To screen for irreducible diagrams, it is necessary to create an ordered list of droplets (nuclei) for every input state. The list is constructed recursively: to start, the first fold block $\alpha^{(0)}$ in the block list in Eq. (2) is placed into an empty container $\mu^{(0)}$ (for brevity we use C++ language terms here) that will ultimately hold a completed droplet (note that within a droplet every pair of blocks is joined by a path of cross-links). If the next fold block $\alpha^{(1)}$ is cross-linked to $\alpha^{(0)}$ it is added to $\mu^{(0)}$, otherwise a new container $\mu^{(1)}$ is created and $\alpha^{(1)}$ is added to it. This process is continued until all of the blocks in Eq. (2) are inserted into droplet containers, merging droplets [but preserving the order in Eq. (2)] if a block $\alpha^{(n)}$ pulled from the list links two droplets together. The result is an ordered list of droplets,

$$\mu^{(0)} \rightarrow \mu^{(1)} \rightarrow \mu^{(2)} \rightarrow \cdots , \tag{A1}$$

where in each droplet, the blocks are ordered as they appear in sequence and front $(\mu^{(p<q)}) <$ front $(\mu^{(q)})$.

For reducible diagrams, each droplet is either completely independent of (joined by a single unfolded segment), or completely contained by, another droplet. A droplet $\mu^{(p)}$ contains $\mu^{(q)}$ if the front and back of the list $\mu^{(q)}$ are bracketed in sequence by two adjacent blocks in the list $\mu^{(p)}$ [as in Fig. 2(a)]. Thus, ordering the list speeds computations and, once given, it is simple to test whether a droplet $\mu^{(p)}$ brackets only a *subset* of blocks in $\mu^{(q)}$ (as in Fig. 3) signifying an irreducible part in the microstate.

[1] E. D. Nelson and N. V. Grishin, Phys. Rev. E **70**, 051906 (2004).

[2] D. Bohm and B. Hiley, *The Undivided Universe* (Routledge, New York, 1993), secs. 15.4–6.

[3] In other words, a condition of local equilibrium (see A. Fernandez, A. Arias, and D. Guerin, Phys. Rev. E **52**, R1299 (1995), and reference [4]) that favors sequence local cross-link formation, or zipping; K. Fiebig and K. A. Dill, J. Chem. Phys. **98**, 3475 (1993); M. Cieplak, Phys. Rev. E **69**, 031907 (2004). As is now well understood, the folding rates of proteins inversely correlate with the mean length of native loops, or "contact order;" D. A. Baker, Nature (London) **405**, 39 (2000); D. N. Ivankov *et al.*, Protein Sci. **12**, 2057 (2003).

[4] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1996).

[5] P. G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979).

[6] P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **94**, 6170 (1997).

[7] As in Ref. [1], this is an attempt to explain the physical origin of folding "pathways." Here, the idea of renormalization indicates possible correspondence between mechanically cooperative substructures and thermodynamically cooperative folding units observed in hydrogen exchange experiments [see, for example, M. Krishna, Y. Lin, L. Mayne, and S. W. Englander, J. Mol. Biol. **334**, 501 (2003); H. Maity, M. Maity, M. Krishna, L. Mayne, and S. W. Englander, Proc. Natl. Acad. Sci. U.S.A. **102**, 4741 (2005); and also C. Merlo, K. A. Dill, and T. R. Weikl, Proc. Natl. Acad. Sci. U.S.A. **102**, 10171 (2005)].

[8] A. V. Finkelstein and A. Y. Badretdinov, Folding Des. **2**, 115 (1996).

[9] A. V. Finkelstein and A. Y. Badretdinov, Mol. Biol. **31**, 391 (1997).

[10] O. V. Galzitskaya and A. V. Finkelstein, Proc. Natl. Acad. Sci. U.S.A. **96**, 11299 (1999).

[11] S. O. Garbuzynskiy *et al.*, J. Mol. Biol. **336**, 509 (2004).

[12] The spin glass approach has been used to extract rough nucleation features of small proteins by many other authors [for a review, see V. Munoz, Curr. Opin. Struct. Biol. **11**, 212 (2001)]. More recent work has turned toward a better description of the unfolded state degrees of freedom [see, for example, D. E. Makarov and K. W. Plaxco, Protein Sci. **12**, 17 (2004); P. Das, S. Matysiak, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. **102**, 10141 (2005)].

[13] Y. Bollen and C. van Mierlo, Biophys. Chem. **114**, 181 (2005).

[14] P. Das, C. Wilson, G. Fossati, P. Wittung-Stafshede, K. Matthews, and C. Clemente, Proc. Natl. Acad. Sci. U.S.A. **102**, 14569 (2005).

[15] Contacts are registered between atoms that are less than 5 Å apart and only between amino acids that are separated by at least one site along the chain.

[16] P. G. de Gennes, *Simple Views on Condensed Matter* (World Scientific, Singapore, 1992) Sec. 3.14.

[17] Alternatively, we use the spin-glass language [see for example, D. S. Fisher, Physica D **107**, 204 (1997)].

[18] S. S. Plotkin, J. Wang, and P. G. Wolynes, Phys. Rev. E **53**, 6271 (1996); J. Chem. Phys. **106**, 2932 (1997).

[19] See for example, B. Zagrovic and V. S. Pande, Nat. Struct. Biol. **10**, 955 (2003); J. Mol. Biol. **323**, 153 (2002) and references.
[20] S. F. Edwards, Proc. Phys. Soc. London **92**, 9 (1967).
[21] E. Paci, M. Vendruscolo, and M. Karplus, Biophys. J. **83**, 3032 (2002).
[22] S. S. Plotkin, Proteins **45**, 337 (2001); C. Clementi and S. S. Plotkin, Protein Sci. **13**, 1750 (2004).
[23] M. Gruebele, Nat. Struct. Biol. **9**, 154 (2002); A. Capaldi, C. Kleanthous, and S. Radford, Nat. Struct. Biol. **9**, 209 (2002).
[24] E. D. Nelson and N. V. Grishin (unpublished).
[25] A. Fersht, *Structure and Mechanism in Protein Science* (Freeman and Company, New York, 1999).
[26] E. Lopez-Hernandez and L. Serrano, Folding Des. **1**, 43 (1996); E. Lopez-Hernandez, P. Cronet, L. Serrano, and V. Munoz, J. Mol. Biol. **266**, 610 (1997); C. Clementi, H. Nymeyer, and J. Onuchic, J. Mol. Biol. **298**, 937 (2000).
[27] N. Gō, Adv. Biophys. **18**, 149 (1984); N. Koga and S. Takada, J. Mol. Biol. **313**, 171 (2001).
[28] M. S. Chung, A. F. Neuwald, and W. J. Wilbur, Folding Des. **3**, 51 (1997).
[29] We study protein G, src, and alpha spectrin sh3, CI2, parvalbumin B, Barnase, CheY, cutinase, apoflavodoxin, staphylococcal nuclease, ribonuclease H, and T4 lysozyme.
[30] The free energy profile of Barnase has a minima in the transition region [roughly corresponding to the kinetic intermediate identified in L. Serrano, A. Matouschek, and A. R. Fersht, J. Mol. Biol. **224**, 805 (1992)]. The intermediate is bordered by two roughly equal free energy maxima and the data in Fig. 4 corresponds to the maxima nearest the folded state [N. Vu, H. Feng, and Y. Bai, Biochemistry **43**, 3346 (2004)].
[31] L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, J. Mol. Biol. **254**, 260 (1995).
[32] The pdb files used to generate this data are 3chy, 1a2p, 1coa, 1srl, and 1shg, respectively. The corellation coefficients given for src and α-spectrin correspond to the more recent (more complete) data set [11].
[33] R. G. Palmer, D. L. Stein, E. Abrahams, and P. W. Anderson, Phys. Rev. Lett. **53**, 958 (1984).
[34] J. d'Aquino, J. Gomez, V. Hilser, K. Lee, I. Amzel, and E. Freire, Proteins **25**, 143 (1996).
[35] R. Guerois and L. Serrano, J. Mol. Biol. **304**, 967 (2000); C. Kiel *et al.*, J. Mol. Biol. **348**, 759 (2005).
[36] D. C. Wright and N. D. Mermin, Rev. Mod. Phys. **61**, 385 (1989).
[37] During the review of this paper, we simulated all ten of the available crystal structures studied in Ref. [11] using a contact radius of r_c=5.5 Å and found roughly 30% improvement in ϕ-value estimates over the results for r_c=6.0 Å reported in that work.