PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

GeoFold: Topology-based protein unfolding pathways capture the effects of engineered disulfides on kinetic stability

Vibin Ramakrishnan,[1,2] Sai Praveen Srinivasan,[1,3] Saeed M. Salem,[4] Suzanne J. Matthews,[4] Wilfredo Colón,[1,3] Mohammed Zaki,[4] and Christopher Bystroff [1,2,4*]

[1] Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, New York 12180
[2] Department of Biology, Rensselaer Polytechnic Institute, Troy, New York 12180
[3] Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, Troy, New York 12180
[4] Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York 12180

ABSTRACT

Protein unfolding is modeled as an ensemble of pathways, where each step in each pathway is the addition of one topologically possible conformational degree of freedom. Starting with a known protein structure, GeoFold hierarchically partitions (cuts) the native structure into substructures using revolute joints and translations. The energy of each cut and its activation barrier are calculated using buried solvent accessible surface area, side chain entropy, hydrogen bonding, buried cavities, and backbone degrees of freedom. A directed acyclic graph is constructed from the cuts, representing a network of simultaneous equilibria. Finite difference simulations on this graph simulate native unfolding pathways. Experimentally observed changes in the unfolding rates for disulfide mutants of barnase, T4 lysozyme, dihydrofolate reductase, and factor for inversion stimulation were qualitatively reproduced in these simulations. Detailed unfolding pathways for each case explain the effects of changes in the chain topology on the folding energy landscape. GeoFold is a useful tool for the inference of the effects of disulfide engineering on the energy landscape of protein unfolding.

Proteins 2012; 80:920–934.

Key words: topology; pathways; kinetic stability; DHFR; barnase; lysozyme FIS; configurational entropy.

INTRODUCTION

The configurational space of a polypeptide chain is astronomically large, yet the folding of most proteins is completed within a fraction of a second. This paradoxical observation strongly suggests that well-defined folding pathways exist, dictated by energetic interactions and topological constraints. Since the folding pathway is the inverse of the unfolding pathway, a mechanistic model for the unfolding pathway would be a helpful step towards understanding folding pathways and developing a predictive model for folding.

The folding and unfolding sides of the same energy landscape present distinct modeling issues. Easily computed structural bioinformatic models can predict the folding rate, and with it the height of the energetic barrier. The relative contact order,[1] a simple calculation based on the native structure, is highly anticorrelated to and predictive of the folding rate. However, the unfolding rate depends on topology of the backbone, the oligomerization state, the presence of bound ligands, local sequence-structure propensities, and the protein environment. Models of unfolding based on the native state statistics alone have had only modest success.[2]

The other extreme of the modeling continuum are all-atom, explicit molecular dynamics simulations. These models are manageable for the unfolding process but not the folding process, because of the enormous search space of folding. For example, molecular dynamics simulations starting from the native state have produced realistic intermediate states of unfolding[3–5] and approximate unfolding rates[6] for small proteins. While a few small proteins have been successfully folded by explicit molecular dynamics,[4,7,8] deriving folding rates would require numerous runs and is still too computationally challenging. It is possible to speed up the folding process by

Vibin Ramakrishnan's current address is Department of Biotechnology, Indian Institute of Technology, Guwahati 781039, India
Suzanne J. Matthews's current address is Department of Computer Science & Engineering, Texas A&M University, College Station, TX 77843
Saeed M Salem's current address is Computer Science Department, North Dakota State University, Fargo ND 58102.

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: National Science Foundation; Grant number: CCF-0432098, DBI-0448072, MCB-0519507.
*Correspondence to: Christopher Bystroff, Department of Biology, Rensselaer Polytechnic Institute, Troy, NY 12180. E-mail: bystrc@rpi.edu.
Received 14 June 2011; Revised 7 October 2011; Accepted 4 November 2011
Published online 11 November 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.23249

reducing the energy function to a simple Go model[9,10] and thereby get folding rates, but this oversimplification of the energy landscape creates its own inaccuracies.[9] In short, folding lends itself to statistical models but not to all-atom simulations, while unfolding lends itself to all-atom simulations but not to easily computed statistical models.

The challenge is to distill a simplified model of unfolding based on what we know about the mechanism. Along these lines, Xia *et al.* have identified some of the structural characteristics that appear to correlate with high kinetic stability (very slow unfolding) in proteins.[11] They find that kinetically stable proteins tend to be of mixed secondary structure content (alpha/beta), rather than pure beta sheet or pure alpha helix. They often have a dimeric or high order assembly with the N and C termini buried in the multimer interface. In kinetically stable monomers, the termini are often tucked into the middle of a beta sheet. When not buried, chain termini are often observed wrapped around the protein like a belt or latch. The impression from structure gazing is that the situation of the chain termini somehow speaks to the unfolding rate. Therefore a simplified model must account for steric hindrance to unfolding in the native state and along the pathway.

The mechanistic model described here encodes a strictly tree-like unfolding pathway consistent with the "parsing perspective" of Dill,[12] the "block folding" model of Nussinov,[13] and Finkelstein's "folding nuclei".[14,15] Viewed in the folding direction, all steps are on-pathway, and each condensation step involves previously condensed substructures. We further restrict the pathway such that no intermediate can be formed if it requires the chain to pass through itself or if it forces an unlikely concerted motion of three or more parts, like the act of tying of a knot. Similar arguments may be found in Maity and Englander's work on cytochrome c, describing folding events as the stepwise assembly of foldon units.[16] These basic principles are well established in the literature of folding theory and even taken for granted in some of the most successful algorithms for *de novo* structure prediction.[17,18]

Our program, GeoFold (Geometric unfolding; http://www.bioinfo.rpi.edu/bystrc/geofold/server.php) follows on the conceptual framework of our previous model, UNFOLD.[19] We model all steps in a pathway as two-body condensations. An ordered set of two-body condensations is a tree, and a given protein may have many such trees, comprising an ensemble of pathways. In UNFOLD, protein structures were reduced to weighted secondary structure element graphs. Contact energies were assigned to pairs of secondary structure elements using solvent accessible surface area and other terms. The graph was then hierarchically partitioned at each step, without regard for chain crossing. In nine case-studies, the simple UNFOLD pathways were found to be predictive of phi-values and other experimental data on folding pathways. The new program GeoFold now accounts for chain crossing and contains energy terms that account for most of the known energetic components of protein stability. It carries out unfolding on a graph using a finite difference approach, producing simulated experimental data, whereas the previous method only produced a heirarchy of intermediate states.

Because of the absence of any off-pathway intermediates in our model, and because off-pathway states dominate the unfolded side of the energy landscape, the current model cannot predict folding rates, only unfolding rates.

In this article, we demonstrate that a mechanistic model which accounts for steric interactions is sufficient to explain previously unexplained differences in stability and unfolding rate for four proteins with engineered disulfide linkages.

The new program and server will be useful to explore and anticipate the energetic consequences of protein engineering, and for deciphering the genetic roots of protein aggregation and amyloid formation.[20] As we will show, mutations affect the kinetic stability of a protein and the accumulation of intermediate states depending on how they perturb the protein folding pathway. A better understanding of the structural determinants of kinetic stability may lead to ways to improve the shelf life of proteins by design.

RESULTS

Geofold first determines all geometrically possible pivot, hinge, and break points (Fig. 1) in a protein structure, splits that structure into nonoverlapping substructures, then proceeds recursively, until the substructures are fully unfolded. The series of splits comprise unfolding pathways. Unfolding pathways are structured as directed acyclic graphs (DAGs) with bifurcating edges (one node connecting to two nodes) as shown in Figure 1(c). Each bifurcating edge represents an elemental unfolding step, or "cut," that splits one substructure into two, using a pivot, hinge or break move as shown in Figure 1(a,b). Each node in the graph represents a native substructure, which is a spatially contiguous subset of the native protein structure. The forward and reverse rates of reaction of each elemental subsystem are calculated using the energy function defined in Methods. The combined set of all cuts, along with the associated energy functions, represent a system of simultaneous equilibria whose solution is a set of equilibrium concentrations. The equilibrium state can be found numerically using finite difference methods. The equilibrium state and the rate at which it is obtained are the values we compare to equilibrium stability and unfolding rate in this study.

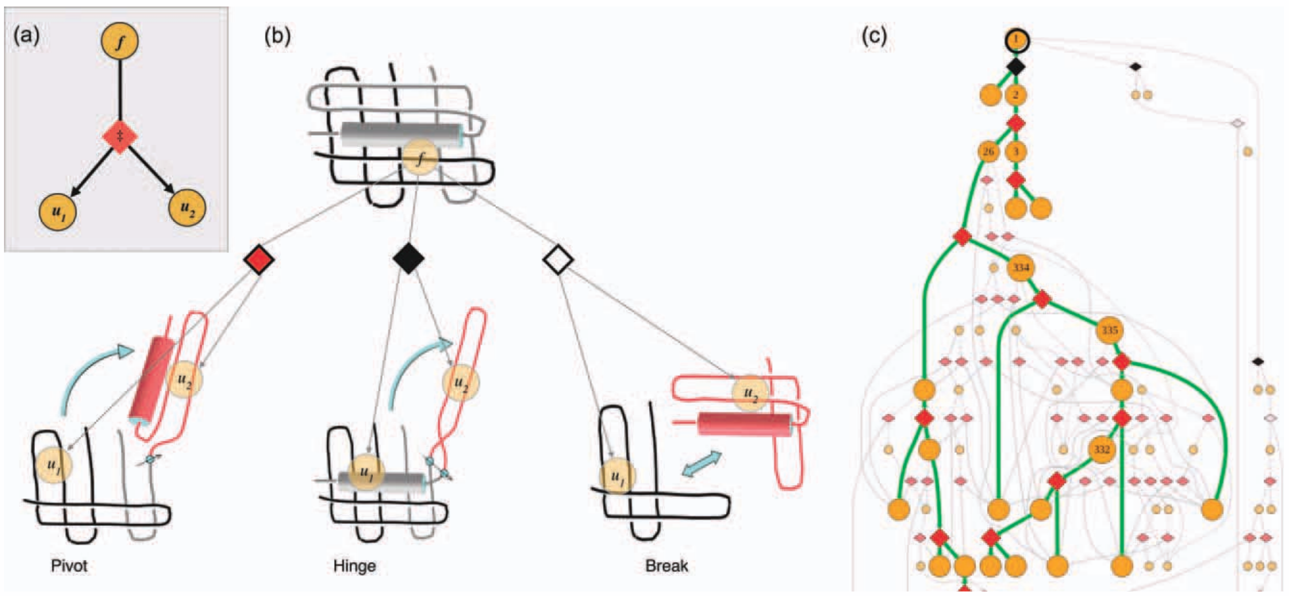


Figure 1

(**a**) Elemental subsystem for the kinetic model, a cut. *f* is any spatially contiguous substructure of a protein, and is partitioned into spatially contiguous substructures u_1 and u_2. (**b**) Diamond shapes represent the cuts, each having a type (color) and an associated energy barrier ‡. A pivot motion is single point revolute joint. A hinge rotates around two points. A break is a pure translational motion. Rotations and translations must not cross chains. (**c**) Top portion of an unfolding pathway DAG for DHFR. Node 1 is the fully folded state. Thick green lines indicate the pathway of maximum unfolding traffic; gray lines are other significant pathways. Unfolding simulations start with all of the protein in node 1, and end when node concentrations reach equilibrium. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

To carry out a virtual unfolding experiment, the energy landscape is tilted from the folded state to the unfolded state by changing either the temperature (T) or the desolvation energy (ω). Finite difference calculations produce a time course of all substructure concentrations. For the purposes of easy analysis, we have grouped substructures together into three classes—folded, unfolded, and intermediate—based on buried surface area. The time course of folded state concentrations from the finite difference simulations were fit to a simple exponential decay to give empirical unfolding rates (k_u). It should be noted that complex, multiphasic unfolding has been observed in these simulations, and this is noted where appropriate, but for most purposes we used a simple half-life analysis.

GeoFold unfolding pathways can be broadly characterized as follows. At high solvation energy (ω) the folded state is always the dominant state at equilibrium, while at low ω the unfolded state always dominates. The transition in ω from folded to unfolded is generally sigmoidal, characteristic of a cooperative process, but the degree of cooperativity varies. Both two-state and three state unfolding behaviors have been observed. In three-state systems, there are values of ω where intermediate states exist at equilibrium. The unfolding rate is estimated from the half-life of unfolding and is often roughly log-linear with ω, but sometimes shows a "roll-over" at very low ω as the kinetics approaches diffusion control. We observe curvature throughout the range of ω when more than two states predominate at equilibrium. The energy landscape of the unfolding pathway generally has a maximum in the middle, characteristic of a two-state system.

Initial studies of small monomeric proteins

Ideally, GeoFold should be able to mimic the unfolding mechanism well enough to reproduce kinetics experiments and predict experimentally determined unfolding rates. To test this, simulated unfolding rates (k_u) were determined for a set of well-studied, small, monomeric proteins.[21] Kinetic simulations were carried out over a range of desolvation free energy values ω. Rates were determined by finding the half-life $t_{1/2}$ at each ω, converting it to an unfolding rate $k_u = \ln(2)/t_{1/2}$, and then finding the rate in "pure water" by log-linear extrapolation of the $\ln(k_u)$ versus ω plot. Unfortunately, the simulated unfolding rates were found to be insignificantly correlated with experimentally determined k_u values (data not shown).

This meant that one or more of the assumptions built into the model were wrong. Since the failure could not be attributed to inaccuracies in the folding pathways versus inaccuracies in the energy calculations, we turned to case studies intended to isolate the effects of topology on the folding pathway. As you will see, the results suggest that the folding pathways are accurate and serve to explain the experimental results, leaving as the cause of failure the inaccuracies in the energy function, although precisely where remains to be determined.

Studies of topological perturbations versus unfolding rate

We asked whether the changes in topology created by disulfide linkages could explain experimentally determined changes in stability or unfolding rate. Four case studies are presented here: three mutants of barnase[22–24] five mutants of T4 lysozyme,[25] two of factor for inversion stimulation (FIS),[26] and one of *E. coli* dihydrofolate reductase (DHFR).[27] In each case, biophysical studies were done to determine the stability or unfolding rate.

Because the task of optimizing the energy parameters to fit the kinetics of unfolding proved to be an impossibly difficult one, the energy function was necessarily left in an unrefined state, with each component set to a physically reasonable value as described in Methods. No attempt has been made with the current case studies to identify parts of the energy function that are responsible for higher or lower rates. Fortunately, the topological effects of disulfides on the unfolding pathway dominate other energetic components, affecting the order of events rather than the energy of each event. For the most part in these case studies, the addition of a disulfide linkage does not change the noncovalent terms of the energy function. For that reason, we can proceed to validate the mechanistic aspect of the program and its prediction of the order of events, without further empirical proof of the validity of the energy function. Images of the unfolding pathways as directed acyclic graphs are provided as Supporting Information.

Disulfide mutants of barnase

In work by Clark *et al.*, barnase has been mutated to add disulfide linkages at three positions, 43–80, 70–92, and 85–102.[22–24] The crystal structures of each have been solved, and the equilibrium stabilities were determined for each mutant, both in the oxidized and reduced forms. The 85–102 mutant (PDB ID: 1bng) was stabilized against urea denaturation due to disulfide formation, unfolding at >7.5*M* urea as compared with 4.5*M* for the reduced form. The oxidized 43–80 mutant (PDB ID: 1bne) was somewhat more resistant to urea denaturation, unfolding at 5.8*M* versus 4.0*M* for the dithiol form. The 70–92 mutant (PDB ID: 1bnf) was destabilized by the disulfide bond, unfolding at 3.0*M* urea versus 3.4*M* for the dithiol form. The loss of stability was attributed to disulfide-induced distortion of local structure and disruption of a salt-bridge near the site of mutation. Further inspection on our part finds that 1bnf in the oxidized state contains local structure that is less favorable than the wild type local structure. Specifically, the Type II beta turn at high propensity sequence KSGR is replaced with a Type I turn at low turn propensity sequence SGRC in the oxidized double mutant. Turn propensity was measured using HMMSTR.[28]

Clarke *et al.* attempted to explain the disulfide stabilization on the basis of decreased entropy of the denatured state, using Flory's formula $\Delta S = -2.1 - 3/2(\ln n)$; where n is the number of residues encompassed by the disulfide.[29] But this theory could not explain how a shorter encompassed loop, 85–102, was more stabilizing than a longer one, 43–80. The differences can be explained by the unfolding pathway predicted by GeoFold, which finds the wild type 85–102 contact to be broken early in unfolding, followed by the 70–92 contact, followed by 43–80. The C-terminal region 50–110 is a 4-stranded beta-meander with an exposed C-terminus, which provides a series of pivot points starting from the C-terminus and working inward [Fig. 2(a)]. The predicted pathway shows the C-terminal strand unfolding first, exposing the neighboring strand (Strand 3), which unfolds, leaving residues 1–90. At this point, the rate limiting step is the pivoting of the N-terminal helix away from the (now partially unfolded) beta sheet. In 1bng, the 85–102 disulfide prevents the unfolding of Strand 3, forcing the N-terminal helix to expose a greater surface area when it separates from the beta sheet. The higher solvation energy cost of this step explains the greater stability of the 85–102 mutant.

GeoFold predicts no difference in the unfolding rate of 43–80 as compared to wild type. It also predicts only a minor difference between the 85–102 mutant and the 70–92 mutant, and both are predicted to unfold much slower than wild type. The 70–92 mutant is predicted to be much more stable than it actually is, but this is believed to be the result of loop distortions observed in the 70–92 mutant. Distortion and local structure propensity are not part of the current GeoFold model.

Figure 2(b) shows the locations of the engineered disulfides relative to the unfolding order of the secondary structure elements. Figure 2(c,d) shows a plots of equilibrium unfolding, using desolvation energy ω, and unfolding rate k_u with respect to ω. The log-linear relationship is steeper for the oxidized 85–102 mutant than for the wild type and 43–80 mutant, consistent with a greater amount of surface area exposed in the transition state of the former.

Disulfide mutants of lysozyme

Matsumura[25] used a rational design algorithm to insert disulfide bonds into phage T4 lysozyme in places where they were predicted to minimally disturb the backbone conformation. Five double mutants were created and then subjected to thermal denaturation under reduced and oxidizing conditions. A cysteine-free mutant (C54T-C97A) served as the wild type (called WT*) and as the base from which the five disulfide mutants were made. All of the reduced mutants were slightly less stable

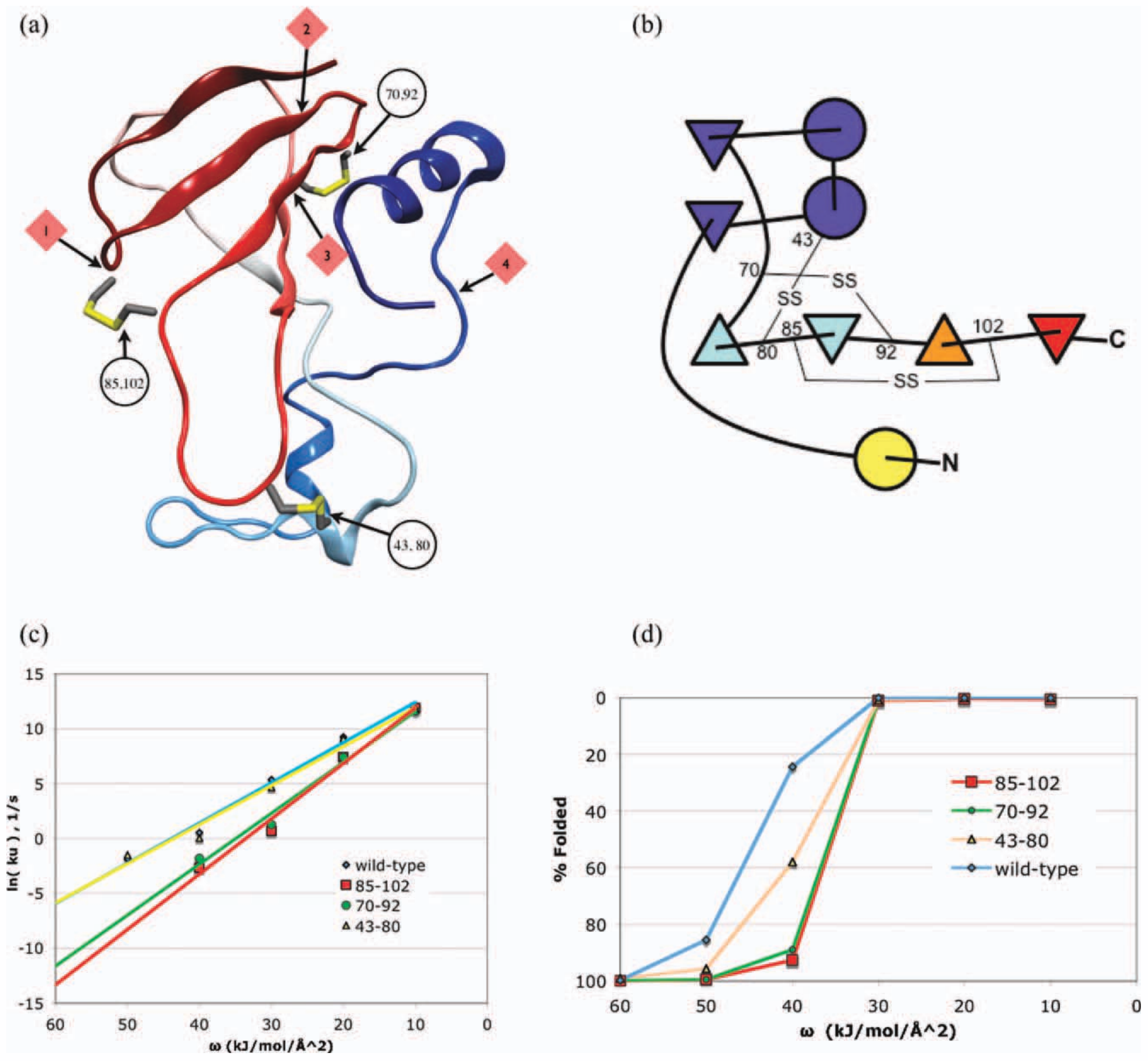


Figure 2

(**a**) Barnase ribbon showing locations of first four pivot locations and the disulfide positions. (**b**) Secondary structure element diagram of barnase showing locations of disulfide linkages. Online color version shows the predominant pathway of unfolding from red (early unfolding) to blue (late unfolding). (**c**) Simulated unfolding kinetics, showing a slowing effect for mutants 85–102 and 70–92, a result of blocking early unfolding steps. Note that slow unfolding rates with $\ln(k_u) < -5$ cannot be measured. (**d**) Equilibrium unfolding simulation, varying desolvation energy ω. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

than WT*, while three of the five mutants were significantly more stable when oxidized [Fig. 3(a)].

Models for the five disulfide mutants were made using the molecular modeling software MOE (Chemical Computing Group, Montreal), using its Rotamer Explorer function plus energy minimization. All models have good disulfide bond geometry and no significant changes in the backbone coordinates.

The wild type unfolding pathway starts with the separation of the N and C-terminal domains with a pivot position located in the middle of a long, domain-crossing helix, at residue 69 [Fig. 3(c), insets]. According to the program, cleavage at this particular location exposes a minimum amount of buried surface area, allows a maximum number of pivoting directions as compared to all other positions in the chain, and maximally exposes new flexible pivot and hinge locations along the pathway of unfolding, as compared with all other cleavage locations. In two of the mutants, 127–154 and 90–122, the initial step of the unfolding

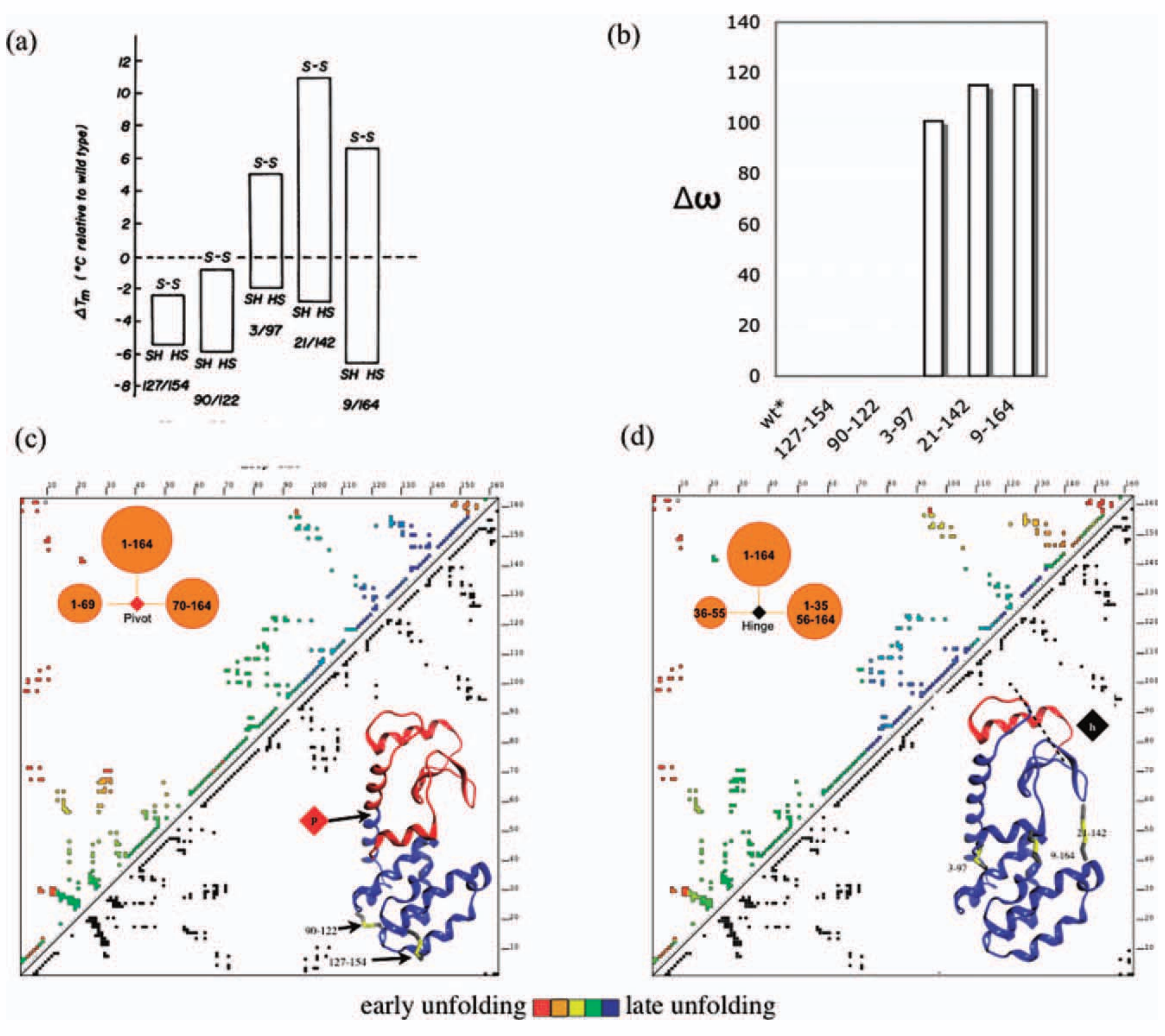


Figure 3

(**a**) Figure 2 from Matsumura *et al*., used by permission, showing the changes in melting temperature for reduced versus oxidized disulfide mutants of T4 lysozyme. (**b**) Changes in equilibrium unfolding point, as ω value, in GeoFold simulations. Mutants 127–154 and 90–122 unfold at the same w at WT*. (**c,d**) Age plot for unfolding pathway of (c) wild type T4 lysozyme, or reduced, or 127–154 or 90–122 mutants, and (d) oxidized 21–142 or 9–164 mutants, with contacts colored red to blue according to unfolding order. Upper inset in (c,d): first unfolding step, a pivot move in (c), a hinge move in (d). Lower inset: ribbon drawing showing how the structure is divided in the first unfolding step by (c) the pivot move *p*, and (d) the hinge move *h*.

pathway was the same as that of WT*, and the equilibrium melting point in ω was also the same, while in the other three mutants the melting point in the oxidized state was increased [Fig. 3(b)], similar to the experimental results.

In the three stabilized mutants, 3–97, 21–142, and 9–164, a pivot at any position between the two cysteines was disallowed by the algorithm, and the unfolding pathway began instead with small N and C-terminal segments. But these small pivot moves did not unlock domain opening steps. Hinge motions, revolute joints with two fixed points, were required for unfolding to proceed. Figure 3(d) (insets) shows the first significant unfolding step in the stabilized mutants. Unfolding in the simulation is slower because the program awards hinges a lower configurational entropy gain than pivots, given that a single-axis revolute joint adds only one rotational degree of freedom, while a pivot motion adds two or three rotational degrees of freedom. Exposing the same amount of buried surface area while gaining less configurational entropy leads to a higher transition state energy and therefore a lower unfolding rate.

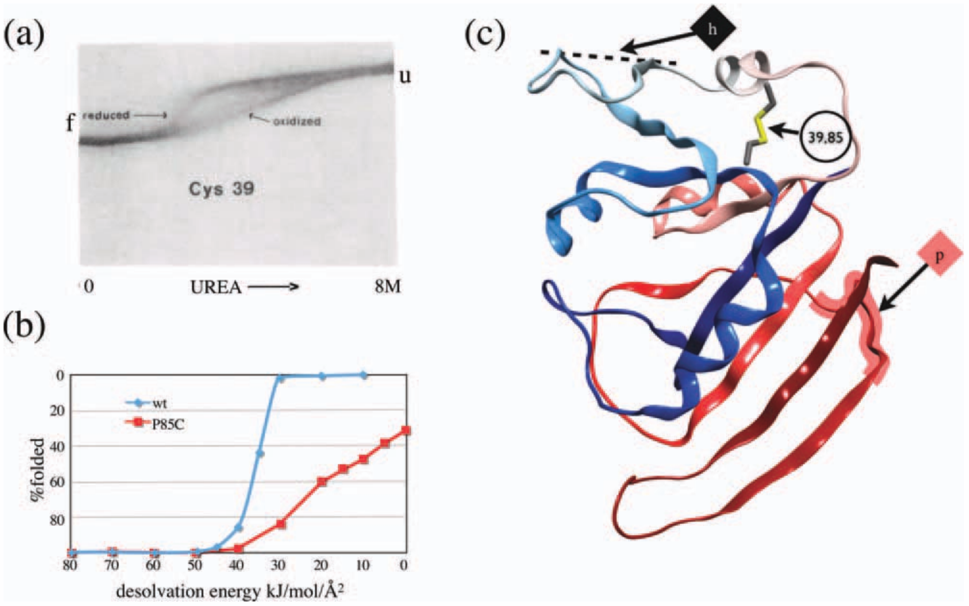


Figure 4

DHFR. (**a**) Figure 4 from Villafranca *et al.*, used by permission, showing urea gradient gel electrophoresis equilibrium denaturation of reduced and oxidized P39C DHFR. (**b**) Simulated equilibrium denaturation curve from GeoFold. The axes have been reversed to match the image. (**c**) DHFR ribbon showing location of engineered disulfide and first unfolding steps in the inside-out pathway, hinge *h*, and the outside-in pathway, pivot *p*. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 3(c,d) also show a summary of the two pathways as “age plots" where contacts are colored in the order they are lost, illustrating the inside-out unfolding order of the stabilized mutants, versus the outside-in unfolding pathway of the WT*. Flory’s equation and our mechanistic approach predict the same overall result in this case, although for barnase they do not.

A disulfide mutant of dihydrofolate reductase

A single mutation in *E. coli* dihydrofolate reductase (DHFR), P39C, allows a disulfide to form with wild type cysteine 85.[27] The wild type and the mutant in both the oxidized and reduced state were characterized by GndHCl and urea equilibrium unfolding experiments, showing that the reduced form was identical in stability to the wild type, and that the oxidized form unfolded at higher urea (or GndHCl) concentrations. Furthermore, the wild type enzyme has a sharp, two-state equilibrium unfolding curve while the oxidized mutant has an extended transition region, suggestive of one or more intermediate states [Fig. 4(a)].

In remarkable agreement, the simulated equilibrium unfolding curve also shows the extended transition for the oxidized state, and the initial unfolding begins at the same urea concentration (desolvation energy ω in the simulations) for both oxidized and reduced states [Fig. 4(b)].

At high ω (low urea) the predicted pathway of DHFR unfolding begins with a series of hinge motions within the loosely packed “adenosine binding domain,” residues 37–91 (ABD), unfolding generally from the middle of the chain outward to the termini [Fig. 4(c)]. This “inside-out” pathway makes sense energetically, given that the terminal segments have extensive contacts and are more topologically tangled than the ABD, and experimentally, given that the ABD is somewhat flexible, rotating between crystal structures.[30] The 39–85 disulfide blocks the inside-out pathway, forcing the unfolding of the remaining protein to proceed from the termini, or “outside-in," which requires a much lower desolvation energy ω. At low ω (high urea), the pathway is outside-in, unfolding from the C-terminus and without the use of hinge motions. Two distinct unfolding pathways lead to the broad transition seen in Figure 4(a,b).

The existence of multiple pathways in DHFR folding (two or four channels) was proposed by Matthews.[31,32] Their descriptions of mutually exclusive channels without equilibrium intermediate states is perfectly consistent with the mutually exclusive inside-out and outside-in pathways observed here. A similar outside-in unfolding scenario was developed based on the kinetics of methotrexate binding and tryptophan fluorescence,[31,33] and was later supported by Go simulations[34] which showed the C-terminus folding last and the adenosine binding domain folding first. On the other hand, hydrogen/deuterium exchange NMR experiments supporting a pathway in which the termini fold first, and unfold last;[35] specifically, a burst phase intermediate of folding contained protected backbone H-bonds in the C-terminal strand. Iwakura,[36] using circular permutants, has suggested that DHFR folding depends only on the presence of early folding units and not on their order along the chain. The NMR experiments and the indifference of DHFR to circular permutation agree with the inside-out pathway. In retrospect, it makes perfect sense that the outside-in pathway that exposes more surface area early would be favored at high denaturant versus the inside-out pathway, which is more sterically hindered, because steric hindrance does not depend on denaturant. Note that the NMR experiments and the inside-out unfolding pathway were carried out at low denaturant, whereas the Trp fluorescence and the outside-in pathways were done at high denaturant.

Symmetric disulfides in the fragment for inversion stimulation dimer

Factor for inversion stimulation (FIS, PDB ID:3jrh) is an intertwined, homodimeric DNA-binding protein. Two single site cysteine mutations were engineered into the dimer, and both mutants formed a disulfide at the twofold symmetry interface.[26] This created proteins with branched, noncyclic topology, in contrast to the other disulfide linkage mutants presented here, all of which produced a cyclic topology. For this reason, the increased stability of these mutants cannot be explained by Flory's formula, which models the loss of entropy in the unfolded state due to cycle formation.[29]

The rate limiting step in wild type FIS unfolding has been shown to occur before dissociation of the monomers,[37] so that any increase in stability must be due to slower unfolding, not to a faster association of monomers or decreased entropy of the unfolded state. Both disulfide bridges were shown to stabilize FIS, but the S30C mutant was more stabilized and denatured more cooperatively than the V58C mutant [Fig. 5(a)]. Equilibrium unfolding curves for wild type and S30C both fit a dimeric two-state model, whereas V58C best fits a 3-state model. In previous studies, a mutation of proline 61 in helix B to alanine increased stability by 4 kcal/mol and changed the folding pathway from 2-state to three-state,[38] and the C-terminal helices C and D were shown by limited trypsin proteolysis to unfold first in this mutant. The equilibrium intermediate was determined to be a trypsin-resistant dimeric fragment consisting of intertwined helices A and B.

In the simulations, the transition state of the wild type protein was dimeric. The slow step in wild type unfolding was the initial pivoting of helix A (either one) away from the rest of the dimer [Fig. 5(a)]. Both mutants were more stable than the wild type, and S30C was more stable and more cooperative than V58C, agreeing with the experimental results [Fig. 5(b)]. The kinetics of wild type and S30C unfolding show a log-linear relationship with ω but V58C has a distinctly curved relationship, suggesting different pathways at high and low denaturant, again similar to the experimental results.

Ignoring the floppy N-terminal hairpin which is disordered in most crystal structures, FIS is an all alpha helical dimeric protein, and unfolding can proceed only at the three junctures between the four helices, A/B, B/C, and C/D [Fig. 5(a)]. The intertwined dimer cannot dissociate before the A/B pivot. We observed an A/B pivot in the wild type versus a B/C pivot in the S38C mutant, where the A helices are linked and not free to pivot. In the V58C mutant, ambiguous pathways, both A/B and B/C, were observed in the simulation.

The simulated pathways agree with experimental data wherever possible. Inasmuch as the P61A and V58C mutations both serve to strengthen the dimeric interaction between helices B, the similarity between the experimental pathway of P61A and the simulated pathway of V58C is supportive of the accuracy of the program. Both mutations serve to block the propagation of the wild type unfolding pathway, forcing B/C in lieu of dissociation.

DISCUSSION

A recent study has found that it is possible to predict the unfolding rates of single-domain proteins using only information about the structural class of the protein and its size.[39] Indeed, kinetically stable proteins have structural class preferences.[11]

A model for protein flexibility has been previously described as a network glass,[40] where unfolding is done by a stochastic simulation using multijointed tethers for hydrophobic interactions and a template-based hydrogen bond potential. This method has been used to identify a transition state cluster in barnase unfolding, and consistently identified the regions most protected from H/D exchange. But H/D exchange only identifies broken hydrogen bonds, not necessarily capturing nonlocal side chain contacts, and in principle, hydrogen bonds in late folding helical regions could be protected from exchange early in folding. The pathway proposed by Rader[40] has the helical domain unfolding first, which disagrees with our prediction (Fig. 2). But predictions of phi-values in barnase by Galzitskaya[15] using dynamic programming agree with our results, finding the high phi-values in the N-terminal helix. GeoFold differs in many ways from Rader's method. It is deterministic and exhaustively samples alternative pathway, like Galzitskaya's method, and it treats pivot and hinge energies differently. This last feature accounts for our barnase pathway, which agrees more with the experimental data. Admittedly, we placed more weight on entropic terms than on hydrogen

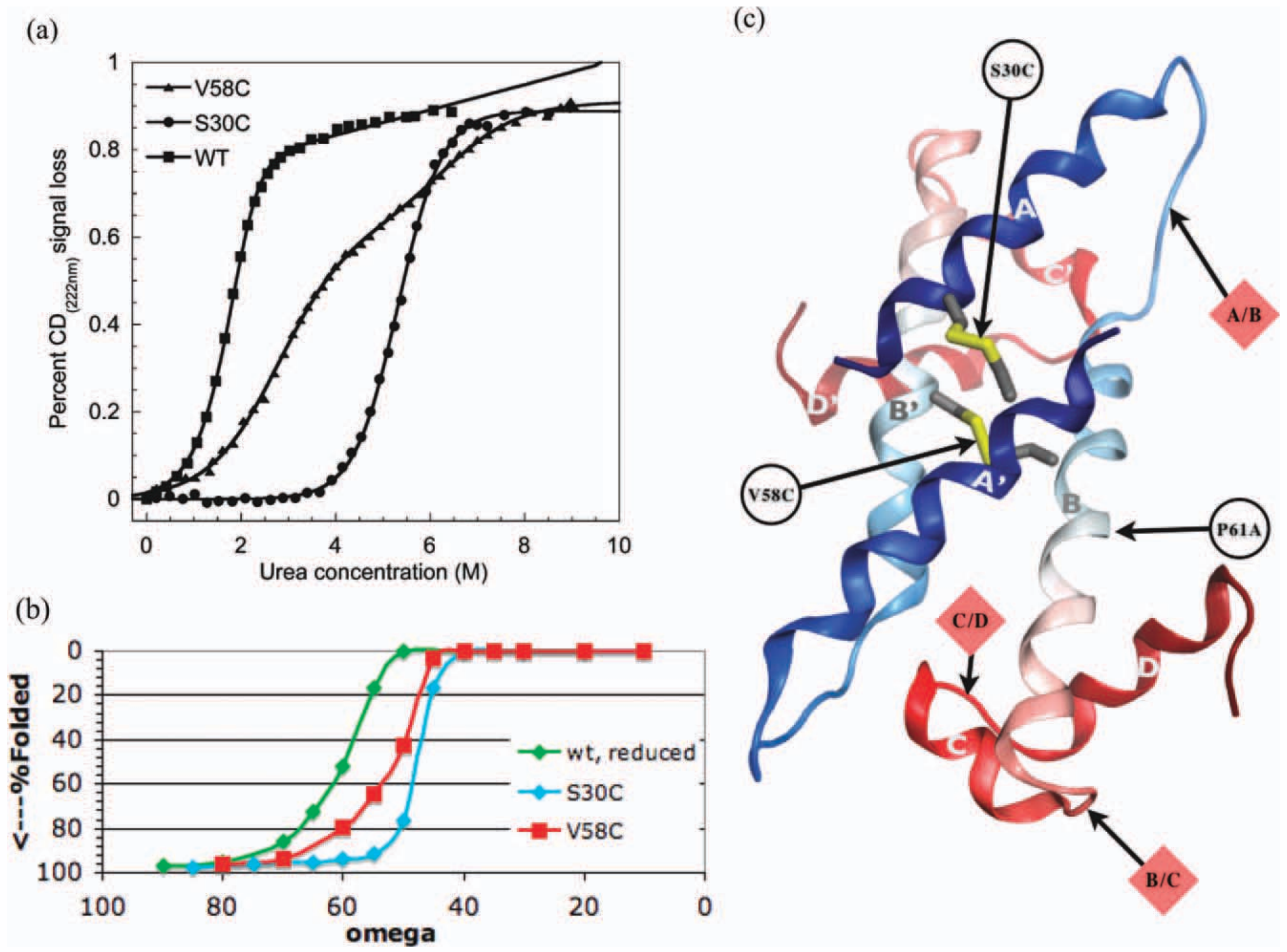


Figure 5

FIS. (**a**) Figure 4 from Meinhold *et al.*, used by permission, showing equilibrium unfolding circular dichroism data for FIS disulfide variants. (**b**) Simulated equilibrium unfolding curves from GeoFold for the same variants. Note that axes are reversed to conform with (a). (**c**) Ribbon diagram for alpha helical part of FIS dimer, showing principle cleavage points and locations of mutations. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

bonds, but that is because these terms, not hydrogen bonds, explain the kinetic effects of topological changes in the chain. In barnase, unfolding internal helices requires hinge motions whose barrier heights depend on chain stiffness and the orientation of the hinge axis, and this may be more energetically unfavorable step than would be expected from the breaking of hydrogen bonds alone.

Our model uses rigid body motions to unfold proteins. This is clearly a convenient simplification of a more detailed process. An actual pivot most likely involves micro-steps in which single hydrogen bonds or hydrophobic contacts are broken, much like the model of Rader. Our simplification is justified because it models that way a set of contacts is often broken in a concerted and cooperative way with one large-scale motion, effectively separating relatively rigid substructures. The relative simplicity of the model and the fact that it is deterministic, not stochastic, has the advantage of allowing the pathways to be explored essentially exhaustively.

The effects of disulfide linkages on folding and unfolding have been previously explored using lattice simulations and theory. Shakhnovich[41] showed that even in a very simple model, the kinetic effect of tying together two sequence positions is a function of the topology of the native state and can either speed up or slow folding, depending on whether the linkage occurs in the folding nucleus or not. Although not discussed in that paper, the implication is that the unfolding rate would be slowed if the linkage occurs outside the folding nucleus, not inside. This would place the energy perturbation on the unfolding side of the energy landscape, increasing the height of the barrier to unfolding. Indeed this is what we find.

Compared with the subtle energetic perturbations of a point mutation, the basis of phi-value analysis of folding pathways,[42] the addition of a disulfide bond is a relatively blunt instrument, probing the pathway by changing its course. We do not expect the current method to be able to reproduce the results of phi-value experiments unless the finer points of the energy function are extensively refined and trained first. Nonetheless, a clearer understanding of the effects of topology on kinetic stability is immensely valuable. In combination with simulations such as those presented here, disulfide mutations can experimentally elucidate the first steps in the unfolding pathway, and conversely, predictions of the first steps in unfolding could help us to engineer stability by inserting disulfide linkages.

CONCLUSIONS

The mechanism of protein unfolding has been hypothesized in this work to be a directed acyclic graph of native substructures, is accordance with theoretical studies and views.[12] An element of this tree is a revolute joint or a translation, splitting a substructure into two. We show that experimentally determined energetic and kinetic effects of engineered disulfides in four different proteins are captured in the energy landscapes produced for these proteins by GeoFold, based on their respective crystal structures. The unfolding pathways explain variable stabilization in barnase, lysozyme, DHFR and FIS that could not be explained by Flory's equation for entropy loss in the unfolded state.[29] Disulfide links stabilize the protein relative to wild type if the linked positions dissociate early in the pathway of the wild type molecule.

Simulated disulfide mutations in DHFR and FIS both reproduced the experimentally observed increases in stability and the decreases in the cooperativity of folding. Simulated disulfide mutations in barnase and lysozyme reproduced the relative changes in the unfolding rate and in stability.

METHODS

A kinetic model

A kinetic simulation for a system of chemical equations simulates the changes in concentration of each chemical species with time. For example, given a system of two coupled equilibria, A$\leftrightarrows$ B$\leftrightarrows$ C, and starting concentrations [A], [B], and [C], the change in [B] over time is given as

$$\frac{d[\mathrm{B}]}{dt} = k_{\mathrm{AB}}[\mathrm{A}] + k_{\mathrm{CB}}[\mathrm{C}] - (k_{\mathrm{BA}} + k_{\mathrm{BC}})[\mathrm{B}] \qquad (1)$$

where the subscripts indicate the directions of the reactions. Equation (1) is multiplied by a time-step to get new concentrations and the process is repeated. The simulation eventually reaches equilibrium, in this example, when $k_{\mathrm{AB}}[\mathrm{A}] + k_{\mathrm{CB}}[\mathrm{C}] = (k_{\mathrm{BA}} + k_{\mathrm{BC}})[\mathrm{B}]$. An accurate time-course of concentrations is obtained if the rates are correct and the time-step is sufficiently small. Protein unfolding can be viewed as a system of coupled elemental unfolding steps [Fig. 1(a)].

Unfolding operators

Protein topology defines the allowable unfolding motions. Three geometric operators can be defined to describe all two part structural partitions on a chain [Fig. 1(b)]. As a rule, covalent linkages cannot be broken or stretched in an unfolding operation, and atoms cannot penetrate each other. If the chain crosses only once from u_1 and u_2, then the allowable motion is a pivot, or a single point revolute joint. Pivot rotations can be in any direction, regardless of the direction of the backbone. If the chain crosses twice, rotation around the two crossing points defines a hinge, or two-point revolute joint. If the chain crosses more than twice, then a simple nondistorting motion is impossible unless all of the points lie in a line, in which case it is still a hinge. If the chain does not cross from u_1 and u_2, then the model consists of multiple chains or disjoint segments of one chain. The motion in this case is a simple translation, called a break in this study. A break is assigned the highest entropy change, followed by pivots, followed by hinges.

The elemental unfolding subsystem (cut)

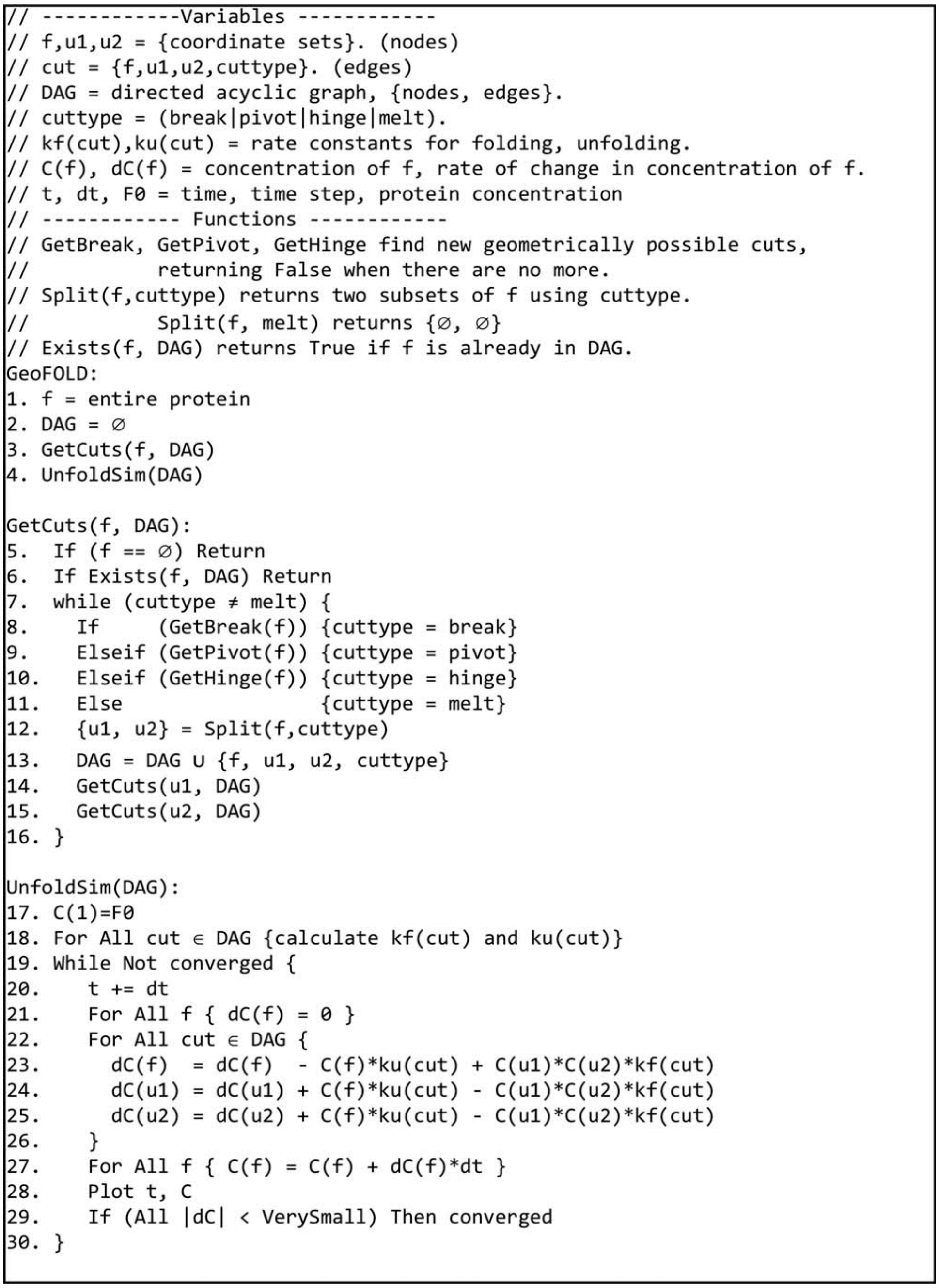
The directed graph consists of linked cuts. Starting from the native structure as the root of the graph, each folded species, *f*, is partitioned using pivots, hinges, and breaks into two smaller species, u_1 and u_2, at all possible locations as defined by the following conditions.

Pivots

(Figure 6, GetPivot) Residue *i* of *f* is the location of a pivot if the substructure N-terminal to *i* ($u_1 = f[:i]$) can rotate around *i* at least pivotcut = 30° in any direction without colliding with the substructure C-terminal to *i* ($u_2 = f[i+1:]$). If *f* is composed of multiple chains, then the coordinates of the additional chains are grouped with u_1 or u_2, in all combinations. If several adjacent positions qualify as pivots, a central representative position is chosen.

Hinges

(Figure 6, GetHinge) A single-axis rotation exists around an axis defined by residues *i* and *j* if the substructure represented by $u_1 = f[i{:}j]$ can rotate at least hingecut = 30° degrees in either direction about the axis *i*->*j* without colliding with the subset $u_2 = f[:i\text{-}1] * f[j+1:]$. If multiple chains are involved, then they are grouped with u_1 or u_2 in all possible combinations. If several consecutive hinge positions are possible, a central representative *ij*



```
// ------------Variables ------------
// f,u1,u2 = {coordinate sets}. (nodes)
// cut = {f,u1,u2,cuttype}. (edges)
// DAG = directed acyclic graph, {nodes, edges}.
// cuttype = (break|pivot|hinge|melt).
// kf(cut),ku(cut) = rate constants for folding, unfolding.
// C(f), dC(f) = concentration of f, rate of change in concentration of f.
// t, dt, F0 = time, time step, protein concentration
// ------------ Functions ------------
// GetBreak, GetPivot, GetHinge find new geometrically possible cuts,
//           returning False when there are no more.
// Split(f,cuttype) returns two subsets of f using cuttype.
//           Split(f, melt) returns {∅, ∅}
// Exists(f, DAG) returns True if f is already in DAG.
GeoFOLD:
f = entire protein
DAG = ∅
GetCuts(f, DAG)
UnfoldSim(DAG)

GetCuts(f, DAG):
If (f == ∅) Return
If Exists(f, DAG) Return
while (cuttype ≠ melt) {
  If     (GetBreak(f)) {cuttype = break}
  Elseif (GetPivot(f)) {cuttype = pivot}
  Elseif (GetHinge(f)) {cuttype = hinge}
  Else                 {cuttype = melt}
  {u1, u2} = Split(f,cuttype)
  DAG = DAG ∪ {f, u1, u2, cuttype}
  GetCuts(u1, DAG)
  GetCuts(u2, DAG)
}

UnfoldSim(DAG):
C(1)=F0
For All cut ∈ DAG {calculate kf(cut) and ku(cut)}
While Not converged {
   t += dt
   For All f { dC(f) = 0 }
   For All cut ∈ DAG {
     dC(f)  = dC(f)  - C(f)*ku(cut) + C(u1)*C(u2)*kf(cut)
     dC(u1) = dC(u1) + C(f)*ku(cut) - C(u1)*C(u2)*kf(cut)
     dC(u2) = dC(u2) + C(f)*ku(cut) - C(u1)*C(u2)*kf(cut)
   }
   For All f { C(f) = C(f) + dC(f)*dt }
   Plot t, C
   If (All |dC| < VerySmall) Then converged
}
```

Figure 6

GeoFOLD algorithm.

pair is chosen. The two parts of u_2 are labeled as different chain segments, allowing break moves.

Breaks

(Figure 6, GetBreak) If a substructure contains two chain segments, either because the protein is oligomeric or because a hinge operation has created two chain segments, and these segments can be separated by a simple translation without collisions in at least breakcut = 0.05 of all possible directions, then a break exists and the two chain segments are labeled u_1 and $u_{2.}$ If more than two segments are present in *f*, then all combinations of segments are tried.

The unfolding graph

(Figure 6, GetCuts) Starting with the native structure as the substructure *f* of the first cut, we find all geometrically possible cuts, giving preference to breaks, then pivots, then hinges. Each cut generates two substructures, u_1 and u_2. We then apply the same method to each of the substructures u_1 and u_2, recursively until the substructures are unfolded (defined below). The result is a directed acyclic graph (DAG), where the nodes are substructures and the bifurcating edges represent transition states of binary partitionings. Figure 1(c) shows a partial DAG for DHFR, showing geometrically possible unfolding steps with the energetically favored unfolding pathway highlighted in green.

Kinetic simulations

(Figure 6, UnfoldSim) A simulation of concentration changes over time can be produced by considering a single cut containing a folding intermediate *f* and the products u_1 and u_2 of a cut type, a pivot, hinge, or break. The amount of *f* lost is proportional to its concentration [*f*] and the unfolding rate, which can be calculated using transition state theory[43]:

$$\kappa_u = -\frac{d[f]}{dt} = \gamma[f]e^{(-\Delta G_u^{\ddagger}/RT)} \tag{2}$$

κ_u may be called the elemental unfolding rate. The barrier to unfolding for one cut, $\Delta G^{\ddagger}{}_{u}$, is a function of the energies of the *f*, $u_{1,}$ and u_2, and of the cut type. The subscript "*u*" indicates that the barrier height is measured in the unfolding direction. The transmission coefficient $\gamma$is equal to the rate of decomposition of the transition state. For a normal chemical reaction, this is $\gamma = k_BT/h$, or about 10^{13} s^{-1}. But a diffusion-controlled folding reaction is much slower, estimated by Fersht[43] to be about 10^6 s^{-1}, since compared to a chemical reaction, the protein folding reaction has a longer and flatter energy landscape with respect to a bond vibration.

The amount of *f* gained is proportional to the $[u_1]$ and $[u_2]$ and the folding rate term

$$\kappa_f = \frac{d[f]}{dt} = \gamma[u_1][u_2]e^{(-\Delta G_f^{\ddagger}/RT)} \tag{3}$$

κ_f may be called the elemental folding rate. The finite element simulations as described in Figure 6, UnfoldSim, are carried out on the set of all κ_f and κ_u to produce a set of all concentrations for each timestep *dt*. By summing over all cuts, *q* that involve *f*, and then multiplying by the timestep, we obtain the concentration change.

$$d[f] = \sum_{q \ni f \in q} \frac{d[f]_q}{dt} dt \tag{4}$$

To simulate unfolding, we initialize the concentration of the native state to F_0. Concentrations of all nodes are recalculated until equilibrium is established. To simulate folding, the leaf nodes are initialized to F_0. Equilibrium is assumed if there was no net change in concentration of the whole system. Note that the sum of the concentrations of all intermediates *f* that contain a given residue *i*, is equal to a constant, F_0, throughout the simulation. That is, the total concentration of residue *i* is conserved.

Table I
Settings of Specific Parameters of the Energy Function

Term	Value	Equation
R	8.314472 J°K^{-1} mol^{-1}	2, 3
d*t*	10^{-7} − 0.01 s	4
F_0	1.0M	5
U_v	0 J mol^{-1} K^{-1}	10
ω	0−100 J mol^{-1}Å^{-2}	8
λ_m	1.0	9
H_h	100 J mol^{-1}	11
ΔS_{break}	90 J mol^{-1}°K^{-1}	12
ΔS_{hinge}	30 J mol^{-1}°K^{-1}	12
ΔS_{pivot}	60 J mol^{-1}°K^{-1}	12
$m_c^{\ddagger}$	10^{-6} J mol^{-1}	14
$\Delta G_{hinge}^{\ddagger}$	50 J mol^{-1}	15, 17, 18
$\Delta G_{pivot}^{\ddagger}$	10 J mol^{-1}	15, 17, 18
$\Delta G_{break}^{\ddagger}$	0 J mol^{-1}	15, 17, 18
σ*	0.5	17, 18
T	300 K	17, 18

$$[i]_{\text{total}} = \sum_{f \ni i \subset f} [f] = F_0 \tag{5}$$

In other words, mass is conserved.

Components of the energy function

The folding and unfolding rates for a cut, $\kappa_{u,}$ and κ_f, are calculated directly from the substructures *f*, $u_{1,}$ and u_2. Free energies are composed of two parts: the dissociation energy ΔE_d, and the backbone configurational entropy $\Delta S_{\mathbf{q}}$. The dissociation energy is composed of four terms: solvation energy ΔE_ω, hydrogen bonds ΔH_h, side chain entropy ΔS_λ, and buried void entropy ΔS_v. Disulfide linkages are treated as constraints rather than energies. The parameter settings used in this study are given in Table 1.

Dissociation energy, ΔE_d

The energy of dissociation of two substructures is modeled using the increased solvation, increased side-chain entropy, loss of hydrogen bonds, and loss of buried void spaces.

$$\Delta E_d = \Delta E_\omega + \Delta H_h - T(\Delta S_v + \Delta S_\lambda) \tag{6}$$

where each term is defined below. Throughout this discussion, ΔE is used for free energies with unspecified entropic and enthalpic components, ΔG for free energies with specified enthalpic and entropic parts, ΔH for purely enthalpic terms and ΔS for purely entropic terms.

Solvation free energy, ΔE_{ω}

For simplicity we assume that the hydrophobic effect, coulombic interactions, and the van der Waals attractive force, are all roughly proportional to the change in solvent exposed surface area, and we therefore combine them in one term, called the solvation free energy. Note that the van der Waals repulsive term is assumed to play no part in unfolding since native structures are assumed to have no collisions. Changes in solvent accessible surface area are computed using MASKER.[44] The buried solvent accessible surface (SAS) exposed upon splitting one substructure, *f*, into two, u_1 and u_2, is approximated as the sum of pairwise residue SAS terms, where each term SAS_{jk} is the burial of SAS upon contact of residues *j* and *k*. Thus for a given cut, summing over residues separated by the cut, we get

$$\Delta\mathrm{SAS} = \sum_{j\in u_1}\sum_{k\in u_2} \mathrm{SAS}_j + \mathrm{SAS}_k - \mathrm{SAS}_{j,k} \qquad (7)$$

The buried surface is a good measure of the amount of water displaced by the folding step, and also is a rough estimate of the scale of the VDW attractive force. Desolvation of hydrophobic groups and hydrogen bonding groups is the primary force driving protein folding.

The solvation free energy in units of kJ mol^{-1} is simply

$$\Delta \mathrm{E}_{\omega} = \omega \Delta\mathrm{SAS} \qquad (8)$$

where ω is the surface tension in J mol^{-1} $\mathrm{\AA}^{-2}$, a value that may be thought of as a modeling the effect of urea at different concentrations. A negative or low value favors solvation and unfolding (high urea), while a high value favors desolvation and folding (low urea). The value of ω corresponding to pure water may be chosen empirically. Theoretical values for hydration of buried protein surfaces[45,46] range from 30 to 80 J mol^{-1} $\mathrm{\AA}^{-2}$. In this study, we did not attempt to break down ΔE_{ω} into its component parts.

Side-chain entropy increase, ΔS_{λ}

Upon unfolding, buried sidechains are exposed to the solvent and gain flexibility, each to a different extent. To calculate the change in sidechain entropy ΔS_{λ}, we multiply the relative change in the sidechain exposure with published values[47,48] for intrinsic sidechain entropy, ω, summing over all residues in *f*.

$$\Delta S_{\lambda} = m_{\lambda} \sum_{i\in f} \frac{\Delta \mathrm{SAS}_i}{\mathrm{SAS}_i^0} \lambda_i \qquad (9)$$

$\mathrm{SAS}^0{}_i$ is the total surface area of residue *i* in the unfolded state, and ω_i is its intrinsic sidechain entropy.

Buried void entropy, ΔS_v

All internal spaces large enough to hold one spherical probe of radius 1.2 Å, but not large enough to hold a water molecule (radius 1.4 Å), were found using MASKER.[44] A typical high resolution crystal structure contains dozens of such cavities, which are entropically unfavorable.[49] Surrounding each void are neighbor residues with atoms less than 7 Å from the void center. The void *v* is said to exist in substructure *f* if *f* contains all of *v*'s neighbors.

$$\Delta S_{\mathrm{v}} = U_{\mathrm{v}}(N_{\mathrm{v}}(f) - N_{\mathrm{v}}(u_1) - N_{\mathrm{v}}(u_2)) \qquad (10)$$

is the difference in the number of voids N_v, times the void cost U_v, an entropic term.

Hydrogen bonds, ΔH_h

Backbone hydrogen bonds were identified by adding hydrogens onto backbone amide nitrogens and finding backbone oxygens within a distance of 2.5 Å. An H-bond exists within *f* if both donor nitrogen and acceptor oxygen are present in *f*. Each H-bond was assigned an enthalpic value H_h, yielding,

$$\Delta H_{\mathrm{h}} = H_{\mathrm{h}}(N_{\mathrm{h}}(f) - N_{\mathrm{h}}(u_1) - N_{\mathrm{h}}(u_2)) \qquad (11)$$

where $N_{\mathrm{h}}(f)$ is the number of H-bonds present in substructure *f*. For simplicity, all H-bonds were assigned the same energy. Sidechain H-bonds were ignored.

Disulfide linkages

Disulfide bonds are treated as inseparable residues, but otherwise contribute nothing to the interaction energy. Any unfolding motion that would separate two disulfide-linked cysteines is disallowed.

Configurational entropy, ΔS_q

Our model assumes that configurational entropy depends only on the number of degrees of conformational freedom gained, and is independent of the size of the substructure. For example, partitioning a large subset of the protein was rewarded with the same entropy gain as partitioning a small piece. Rough entropic values were assigned to each of the three cut types, break, pivot, and hinge, reflecting the approximate number of added degrees of freedom. A hinge adds a single angular degree of freedom, a pivot adds two, and a break adds all three plus some degree of translational freedom. For the two entropies,

$$\Delta \mathrm{S}_{\mathrm{hinge}}, << \Delta \mathrm{S}_{\mathrm{break}} \qquad (12)$$

is enforced, and specific values were set empirically. $\Delta S_{\mathrm{pivot}}$ was necessarily set to the average of the two so that alternative pathways to the same state would always have the same entropy change, a requirement of any state function.

Transition state free energy

Interactions must be broken before full configurational entropy increase is possible during an unfolding step, therefore we spread the configurational entropy, ΔS_q, unevenly along the reaction coordinate, apportioning more than half of the entropy to the products side, after the transition state of unfolding. A term, $0.5 \geq \sigma^\ddagger \geq 0.0$, is used to set the fraction of ΔS_q expressed before the transition state. $\sigma^\ddagger$ was set to 0.25 for this work.

Another term, $0.2 \leq \theta^\ddagger \leq 0.8$, sets the fraction of ΔE_d expressed before the transition state. $\theta^\ddagger$ is calculated using the Hammond postulate, which states that the transition state most resembles the higher energy ground state. To quantify the Hammond behavior we adopted a reasonable simplifying assumption, that the slope of the energy with respect to the reaction coordinate is the same on both sides of the transition state. Using only the ground state energies and this assumption, the solution for the position of the transition state $\theta^\ddagger$ is found using similar triangles.

$$\theta^\ddagger = \frac{(2\Delta E_d - \Delta S_q)}{2\Delta E_d} \tag{13}$$

Note that $\theta^\ddagger$ goes to zero as ΔS_q approaches twice the value ΔE_d, which means that there would be no barrier (diffusion controlled) for weakly connected substructures. To maintain physical realism, $\theta^\ddagger$ is constrained to be in the range $0.2 \leq \theta^\ddagger \leq 0.8$.

Cavitation, $\Delta G_c^\ddagger$

Theoretical studies done independently by Scheraga[50] and Baker[51] have shown a barrier to hydrophobic collapse (or its inverse) due to the atomic size of solvent. The free energy of cavitation, $\Delta G_c^\ddagger$, is expressed only in the transition state of the cut, reflecting the cavity formation that must precede the inward diffusion of water. Based on the cavitation studies of Hummer *et al.*[49] we assume a quadratic relationship between ΔSAS and $\Delta G_c^\ddagger$.

$$\Delta G_c^\ddagger = m_c^\ddagger \Delta \mathrm{SAS}^2 \tag{14}$$

Configurational entropy barriers, $\Delta G_q^\ddagger$

A hinge motion may require the concerted motion of several backbone torsion angles, a pivot motion only one or two angles, and a break motion requires no angular shifts. Strain due to steric interactions may be greater in a hinge motion, than in a pivot motion. $\Delta G_q^\ddagger$ serves to model the barriers to rotation that occur only in the transition state and are dependent on the type of motion. Allowed empirical settings are

$$\Delta G_{hinge}{}^\ddagger \geq \Delta G_{pivot}{}^\ddagger \geq \Delta G_{break}{}^\ddagger = 0. \tag{15}$$

Equilibrium and transition state free energies, $\Delta G_{u\text{-}f}$, $\Delta G^\ddagger{}_f$, $\Delta G^\ddagger{}_u$

Using the transition state placement variables $\sigma^\ddagger$ and $\theta^\ddagger$, the free energy barriers for folding and unfolding are,

$$\Delta G_{u-f} = \Delta E_\omega + \Delta H_h - T(\Delta S_\nu + \Delta S_\lambda) \tag{16}$$

$$\Delta G_u^\ddagger = \theta^\ddagger[\Delta E_\omega + \Delta H_h - T(\Delta S_\nu + \Delta S_\lambda)] - \sigma^\ddagger T\Delta S_q + \Delta G_c^\ddagger + \Delta G_q^\ddagger \tag{17}$$

$$\Delta G_f^\ddagger = (1-\theta^\ddagger)[-E_\omega - \Delta H_h + T(\Delta S_\nu + \Delta S_\lambda)] + (1-\sigma^\ddagger)T\Delta S_q + \Delta G_c^\ddagger + \Delta G_q^\ddagger \tag{18}$$

Note that $\Delta G_{u\text{-}f,} = \Delta G^\ddagger{}_u - \Delta G^\ddagger{}_f$, as required. Values from Eqs. (17) and (18) for each elemental subsystem are used in Eqs. (2) and (3) to define the elemental rates, and the whole system is simulated using the finite difference method (Fig. 6, UnfoldSim).

Folded/unfolded states

For purposes of calculating the unfolding rate from a simulated unfolding trajectory, the folded state is defined as the set of all intermediate substructures in the folding pathway that retain 90% or more of the buried SAS of the native state. In unfolding trajectories, the concentration of the folded state (*F*) is the sum over all folded states.

The unfolded state is defined as all intermediates substructures that have less than 1000 Å^2 of buried SAS. This corresponds to an extended 10-residue fragment or smaller. The concentration of the unfolded state (*U*) is the average, over all sequence positions *i*, of the sum of the concentrations of all unfolded states that contain residue *i*.

Simulated unfolding kinetics

The empirical unfolding rate k_u was defined as $\ln(2)/t_{1/2}$, where $t_{1/2}$ is the time at which (*F*) first reaches 1/2 of its initial value.

ACKNOWLEDGMENTS

The authors thank Patrick Buck, Ke Xia and Yao-ming Huang for useful discussions.

REFERENCES

1. Plaxco KWS, Baker KT. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
2. Gromiha MM, Selvaraj S. FOLD-RATE: Prediction of protein folding rates from amino acid sequence. Nucleic Acids Research 2006; 34:W70–W74.
3. Beck DA, Daggett V. Methods for molecular dynamics simulations of protein folding/unfolding in solution. Methods (San Diego, Calif) 2004;34:112–120.

4. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282;740–744.
5. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
6. Snow CD, Sorin EJ, Rhee YM, Pande VS. How well can simulation predict protein folding kinetics and thermodynamics? Ann Rev Biophys Biomol Struct 2005;34:43–69.
7. Paschek D, Garcia AE. Reversible temperature and pressure denaturation of a protein fragment: a replica exchange molecular dynamics simulation study. Phys Rev Lett 2004;93:238105.
8. Yang WY, Gruebele M. Folding [lambda]-repressor at its speed limit. Biophys J 2004;87:596–608.
9. Chavez LL, Onuchic JN, Clementi C. Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. Chem Phys 1999;111:10375–10380.
10. Buck PM, Bystroff C. Constraining local structure can speed up folding by promoting structural polarization of the folding pathway. Protein Sci 2011;20:959–969.
11. Xia K, Manning M, Hesham H, Lin Q, Bystroff C, Colon W. Identifying the subproteome of kinetically stable proteins via diagonal 2D SDS/PAGE. Proc Natl Acad Sci USA 2007;104:17329–17334.
12. Hockenmaier JJ, Dill KA. Routes are trees: the parsing perspective on protein folding. Proteins 2007;66:1–15.
13. Tsai CJ, Nussinov R. The building block folding model and the kinetics of protein folding. Protein Eng 2001;14:723–733.
14. Galzitskaya OV, Finkelstein AV. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. Proc Natl Acad Sci USA 1999;96:11299–11304.
15. Garbuzynskiy SO, Finkelstein AV, Galzitskaya OV. Outlining folding nuclei in globular proteins. J Mol Biol 2004;336:509–525.
16. Maity H, Maity M, Krishna MM, Mayne L, Englander SW. Protein folding: the stepwise assembly of foldon units. Proc Natl Acad Sci USA 2005;102:4741–4746.
17. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
18. Zhang Y, Arakaki AK, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins 2005;61(Suppl 7):91–98.
19. Zaki MJ, Nadimpally V, Bardhan D, Bystroff C. Predicting protein folding pathways. Bioinformatics 2004;20:i386–i393.
20. Goldberg AL. Protein degradation and protection against misfolded or damaged proteins. Nature 2003;426:895–899.
21. Bogatyreva NS, Osypov AA, Ivankov DN. KineticDB: a database of protein folding kinetics. Nucleic Acids Res 2009;37(Suppl 1):D342–D346.
22. Clarke J, Fersht AR. Engineered disulfide bonds as probes of the folding pathway of barnase: increasing the stability of proteins against the rate of denaturation. Biochemistry 1993;32:4322–4329.
23. Clarke J, Henrick K, Fersht AR. Disulfide mutants of barnase I: changes in stability and structure assessed by biophysical methods and X-ray crystallography. J Mol Biol 1995;253:493–504.
24. Clarke J, Hounslow AM, Fersht AR. Disulfide mutants of barnase II: changes in structure and local stability identified by hydrogen exchange. J Mol Biol 1995;253:505–513.
25. Matsumura M, Becktel WJ, Levitt M, Matthews BW. Stabilization of phage T4 lysozyme by engineered disulfide bonds. Proc Natl Acad Sci USA 1989;86:6562.
26. Meinhold D, Beach M, Shao Y, Osuna R, ColÜn W. The location of an engineered inter-subunit disulfide bond in factor for inversion stimulation (FIS) affects the denaturation pathway and cooperativity. Biochemistry 2006;45:9767–9777.
27. Villafranca JE, Howell EE, Oatley SJ, Xuong NH, Kraut J. An engineered disulfide bond in dihydrofolate reductase. Biochemistry 1987;26:2182–2189.
28. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 2000;301:173–190.
29. Flory P. Statistical mechanics of chain molecules. Carl Hanser: Verlag; 1989, p 432.
30. Bystroff C, Kraut J. Crystal structure of unliganded *Escherichia coli* dihydrofolate reductase. Ligand-induced conformational changes and cooperativity in binding. Biochemistry 1991;30:2227–2239.
31. Jennings PA, Finn BE, Jones BE, Matthews CR. A reexamination of the folding mechanism of dihydrofolate reductase from *Escherichia coli*: verification and refinement of a four-channel model. Biochemistry 1993;32:3783–3789.
32. Kuwajima K, Garvey EP, Finn BE, Matthews CR, Sugai S. Transient intermediates in the folding of dihydrofolate reductase as detected by far-ultraviolet circular dichroism spectroscopy. Biochemistry 1991;30:7693–7703.
33. Touchette NA, Perry KM, Matthews CR. Folding of dihydrofolate reductase from *Escherichia coli*. Biochemistry 1986;25:5445–5452.
34. Clementi C, Jennings PA, Onuchic JN. How native-state topology affects the folding of dihydrofolate reductase and interleukin-1. Proc Natl Acad Sci USA 2000;97:5871.
35. Jones BE, Matthews CR. Early intermediates in the folding of dihydrofolate reductase from *Escherichia coli* detected by hydrogen exchange and NMR. Protein Sci 1995;4:167–177.
36. Arai M, Maki K, Takahashi H, Iwakura M. Testing the relationship between foldability and the early folding events of dihydrofolate reductase from *Escherichia coli*. J Mol Biol 2003;328:273–288.
37. Topping TB, Hoch DA, Gloss LM. Folding mechanism of FIS, the intertwined, dimeric factor for inversion stimulation. J Mol Biol 2004;335:1065–1081.
38. Hobart SA, Ilin S, Moriarty DF, Osuna R, ColÛn W. Equilibrium denaturation studies of the *Escherichia coli* factor for inversion stimulation: implications for in vivo function. Protein Sci 2002;11:1671–1680.
39. De Sancho D, Munoz V. Integrated prediction of protein folding and unfolding rates from only size and structural class. Phys Chem Chem Phys 2011;13:17030–17043.
40. Rader A, Hespenheide BM, Kuhn LA, Thorpe MF. Protein unfolding: rigidity lost. Proc Natl Acad Sci USA 2002;99:3540.
41. Abkevich VI, Shakhnovich EI. What can disulfide bonds tell us about protein energetics, function and folding: simulations and bioninformatics analysis. J Mol Biol 2000;300:975–985.
42. Matouschek A, Kellis JT, Jr, Serrano L, Bycroft M, Fersht AR. Transient folding intermediates characterized by protein engineering. Nature 1990;346:440–445.
43. Fersht A. Structure and mechanism in protein science. W.H. Freeman: New York; 1999; p 614.
44. Bystroff C. MASKER: improved solvent-excluded molecular surface area estimations using Boolean masks. Protein Eng 2002;15:959–966.
45. Wang Y, Zhang H, Scott RA. A new computational model for protein folding based on atomic solvation. Protein Sci 1995;4:1402–1411.
46. Still W, Tempczyk A, Hawley R, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
47. Doig AJ, Sternberg MJE. Side chain conformational entropy in protein folding. Protein Sci 1995;4:2247–2251.
48. Pickett SD, Sternberg MJE. Empirical scale of side-chain conformational entropy in protein folding. J Mol Biol 1993;231:825–839.
49. Hummer G, Garde S, Garcia AE, Pohorille A, Pratt LR. An information theory model of hydrophobic interactions. Natl Acad Sci 1996;93:8951–8955.
50. Czaplewski C, Rodziewicz Motowidlo S, Liwo A, Ripoll DR, Wawak RJ, Scheraga HA. Molecular simulation study of cooperativity in hydrophobic association. Protein Sci 2000;9:1235–1245.
51. Rank JA, Baker D. A desolvation barrier to hydrophobic cluster formation may contribute to the rate limiting step in protein folding. Protein Sci 1997;6:347–354.