SHAVING LEVINTHAL WITH OCCAM'S RAZOR: UNDERSTANDING THE RATE LIMITING STEP IN PROTEIN FOLDING

Thesis by

Derek A. Debe

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, CA

2001

(Defended April 6, 2001)

ACKNOWLEDGEMENTS

Caltech. What an awesome place. My five years here have been and will probably remain some of the finest of my adult life. Thus, my first acknowledgement goes to all of the students, faculty, and administrators who have worked hard to make Caltech the tremendously challenging and rewarding place that it is.

Secondly, and most importantly, I thank my primary graduate advisor, Prof. William A. Goddard III. I first met Bill during my prospective visit to Caltech. In less than an hour, I developed a hunch that this was a man who truly loved science, people, and life. My hunch was correct. I could not have asked for finer direction during my graduate career. Bill's unbridled enthusiasm and broad interests catalyzed my own desire to attack interesting and difficult problems, and his limitless supply of scientific insight is something I will always appreciate and admire.

Many others had a very positive impact on my studies at Caltech. First there is my secondary advisor, Prof. Sunney I. Chan, who was very supportive of my decision to do more theory than experiment (o.k. all theory!). Second, there's my committee, Prof. Rees, Prof. Dougherty, and in particular, Prof. Harry "Knob Creek" Gray. They all graciously took time out of their pressing schedules to help me with my scientific development. Finally, I would especially like to thank fellow Goddard group students Dr. Matt Carlson and Joe "who's your daddy" Danzer. Their hard work and insight helped me create a lot of the work discussed in this thesis.

Next I would like to thank all of my friends at Caltech with whom I have shared the ups and downs of graduate school. First, there are the guys from the old 930 E. Leslie House: Akif "I'm so L.A." Tezcan, Ivan "I drank all your juice" Dmochowski, Kevin "where's my butt" McDonnel, and Bob "green thumb" Beach. Then, there's my other good pals, like Lou "love the hurt" Madsen, Todd "teletubby" Younkin, Ken "Cooter" Brameld, Jim "Mein" Kempf, and Rob "can I buy you a drink" Weber. These are the folks that made Caltech fun after the science was done.

My best friend through all of this has of course been the lovely and talented Grace "Schmee of Pedley" Yang. Her love and support during good times and bad, for sicker, and for poorer, has been amazing. Maybe we'll have to get married.

Finally, a very special thanks goes to my parents. As I grow older and more responsible, I realize that the guidance that Mom inflicted on me as a child and teen is the most important gift anyone has ever given me. Without Mom's gentle tutelage, I never would have been Caltech material. Finally, I would like to thank my first scientific mentor, Dr. Mark K. Debe (a.k.a. Dad). Dad is one of those old school physicists that just seems to know how everything works. Dad, I'm not there yet, but I'm working on it.

ABSTRACT

How do proteins fold? This thesis addresses this simple yet important question by developing a first principles theoretical framework that accurately describes the experimentally observed protein folding rate data. The success of the new theory suggests that single domain proteins fold according two a two-state mechanism consisting of

(i) a random, diffusive search for the native topology, followed by

(ii) non-random, local conformation changes within the native topology to find the unique native state.

In chapter 1, a popular analogy between protein folding and the game of golf is used to qualitatively illustrate the most important aspects of the new theory. In chapter 2, mean-field computational methods are developed that allow the time involved in the rate limiting diffusive search for the native state to be calculated. Chapters 3 and 4 remove the mean-field restriction from the methods of chapter 2, allowing the folding rate for an arbitrary two-state folding protein to be calculated. Chapter 5 then explores how real proteins deviate from this ideal model by examining the roles that non-random mechanisms such as helix, hydrophobic core, and β-turn formation play in the early folding process. Finally, chapter 6 develops an empirical model that also capably predicts protein folding rates, adding further support to the proposed folding mechanism.

TABLE OF CONTENTS

CHAPTER 1

PROTEINS: GREAT SHORT GAME, NO DRIVE

In 1967, an MIT professor named Cyrus Levinthal used a simple, back of the envelope calculation to demonstrate that an average protein has far too many torsion states to be sampled in a biologically relevant period of time [1]. For nearly 25 years following this simple calculation, it was widely believed that Levinthal's Paradox implied that proteins must follow a directed pathway [2] in order to fold quickly, and that the intermediates along this pathway should be experimentally observable. More recently, experimental and theoretical work has led to a "new view" [3] or energy landscape picture of protein folding that removes the restriction that protein folding must proceed along a single reaction.

Early in my graduate career, this shift in understanding was detailed by Martin Karplus in a review entitled: "The Levinthal Paradox: Yesterday and Today" [4]. The abstract of this review begins as follows:

> A change in the perception of the protein folding problem has taken place recently. The nature of the change is outlined and the reasons for it are presented. An essential element is the recognition that a bias toward the native state over much of the effective energy surface may govern the folding process.

In the body of the review, Karplus explains the implications of the emerging new view:

> This means that the difference in energy and free energy between the denatured and native state is reflected in some way not only in the neighborhood of the native state ('golf course' surface) but over a significant portion of the surface sampled during the folding process.

He concludes the review by outlining a challenge for future research:

> We are much more optimistic about being able to solve the folding problem because the Levinthal paradox is no longer a concern. However, we are now faced with the issue of how the energy bias toward the native state is made to extend over a sufficient portion of configuration space to make folding possible on the experimental timescale.

Karplus' comments on the new view's resolution of the Levinthal Paradox can be further understood using the familiar "drunken golfer" analogy. Obviously, a drunken golfer spraying golf balls in random directions has almost no chance of successfully getting the ball into the hole. The pathway model suggests that in order to succeed, the golfer must wait and sober up, and then go out and "direct" the ball into the hole. In contrast, the new view reasons that the drunken golfer can in fact succeed at golf, as long as he or she finds a much less challenging golf course. Given a funnel-shaped course with the hole at the funnel's bottom, any number of poor shots by our beleaguered golfer will be successful.

In terms of this analogy, this thesis demonstrates that there is another type of golf course that would allow the drunken golfer to succeed. A primarily flat golf course with appreciably sized, funnel-shaped putting greens, will also allow the drunken golfer to succeed and make it back to the watering hole by nightfall. The course with the funnel-shaped greens may require a few more strokes than the new view course, but even a drunkard's ball will eventually land on the green and slide into the hole.

So, we have a newer view for protein folding that seems plausible, at least in the context of the golf analogy. Now we must ask the all important "is this a testable theory" question:

> Is there an experimental observable that will allow us to determine if the newly proposed theory is better than the prevailing theory?

We will return to our golf analogy to find out why the answer to this critical question is yes. Both theories suggest that the drunken golfer can succeed in a finite amount of time. However, the parameters that define how long it will take to succeed in each model are very different. According to the new view, the time required to sink the ball is related to the average slope of the course's funnel shaped walls, and the number of sticks, twigs and traps that can hinder the ball's progress toward the hole. In the model proposed by this thesis, the time required to sink the ball is related to the ratio of the size of the green to the size of the golf course, since the rate-limiting step is the search for the green.

Thus, the golf analogy, while simple, clearly shows that the key to determining which folding theory best describes the folding process lies in the analysis of protein folding rate data. If the newer view is correct, the rate that proteins fold will be related to the ratio between the number of states that can quickly descend to the native state and the number of total states that are searched during the folding process. Thus, testing the applicability of the newer view does not require an understanding of the energy of each and every conformation state; it simply requires that the native state is the energy and free energy minimum.

The fact that we do not need to estimate the energy of each state makes this model very simple (satisfying Occam's Razor [5]), and most importantly, *testable*. The chapters in this thesis are devoted to developing the necessary theoretical and computational framework required to validate the new folding theory. The analysis of the folding kinetic data herein represents very strong evidence that the rate-limiting step in protein folding is predominantly a diffusive search for the native topology.

REFERENCES

1. C. Levinthal, in *Mossbauer Spectroscopy in Biological Systems*, eds. P. J. Debrunner, J. C. M. Tsibris, and Münck, E. (Univ. Illinois Press, Urbana), 21-24 (1969).

2. C. Levinthal, *J. Chim. Phys.* **65**, 44-45 (1968).

3. K. A. Dill and H. S. Chan, *Nat. Struct. Biol.* **4**, 10-19 (1997).

4. M. Karplus, *Folding and Design* **2**, S69-S75 (1997).

5. William of Occam (1300-1349). *Frustra fit perplura, quod fieri per paiciora.* It is vain to do with more what can be done with less.

CHAPTER 2

THE TOPOMER SAMPLING MODEL OF PROTEIN FOLDING

ABSTRACT

Clearly a protein cannot sample all of its conformations (e.g., ~$3^{100} \approx 10^{48}$ for a 100 residue protein) on an *in vivo* folding timescale (<1 second). To investigate how the conformational dynamics of a protein can accommodate sub-second folding time scales, we introduce the concept of the *native topomer*, which is the set of all structures similar to the native structure [obtainable from the native structure through local backbone coordinate transformations that do not disrupt the covalent bonding of the peptide backbone]. We have developed a computational procedure for estimating the number of distinct topomers required to span all conformations (compact and semi-compact) for a polypeptide of a given length. For 100 residues, we find ~3×10^{7} distinct topomers. Based on the distance calculated between different topomers, we estimate that a 100-residue polypeptide diffusively samples one topomer every ~3 nanoseconds. Hence, a 100-residue protein can find its native topomer by *random* sampling in just ~100 milliseconds. These results suggest that sub-second folding of modest sized, single domain proteins can be accomplished by a two-stage process of:

- Topomer diffusion: random, diffusive sampling of the 3×10^{7} distinct topomers to find the native topomer (~0.1 sec) followed by
- Intra-topomer ordering: non-random, local conformational rearrangements within the native topomer to settle into the precise native state.

INTRODUCTION

The question, "How do proteins fold?" [1] has puzzled researchers for decades. Based on a very simple calculation, Levinthal [2] estimated that an average sized protein would require longer than the age of the universe to sample every state (for example, if there are three possible conformations for each residue [3], a 100-residue protein would have $\sim 3^{100} \approx 10^{48}$ distinct backbone conformations, which would require $\sim 10^{30}$ years to sample every state). Since proteins of this length can fold on a millisecond timescale, they clearly sample only an infinitesimal fraction of their possible conformations. It was originally assumed that proteins overcome this Levinthal Paradox by following a directed folding pathway [4] that drastically reduces the number of structures that must be sampled. Currently, however, it is generally acknowledged that proteins need not follow a single pathway to fold on a millisecond timescale. Just as a water droplet can follow many different trajectories while descending from the top of a ceramic funnel, a folding energy landscape shaped like a funnel [5] can have numerous folding pathways leading to a properly folded state at the base of the funnel. This suggests that proteins fold along an ensemble of pathways with the folding time scale determined by the ruggedness (kinetic barriers) and slope of the folding energy landscape (see [6] for an excellent review of the "new view" of protein folding [7, 8]).

In considering the nature of the dynamics of an ensemble of folding protein conformations, we find it useful to introduce the concept of a *topomer*. A *topomer* is the set of structures that are obtainable from a specific structure through local backbone coordinate transformations that do not disrupt the covalent bonding of the peptide backbone. Thus, the native topomer is the set of near-native structures for a protein. In this paper, we present the GP computational procedure to estimate the number of disjoint topomers required to span all possible compact and semi-compact conformations for an N-residue polypeptide. For 100 residues, we find ~3×10^7 disjoint topomers. This procedure also leads to an estimate of the distance between neighboring topomers. By combining this distance with an experimental determined protein intra-chain diffusion constant, we estimate that a 100-residue polypeptide undergoing random, diffusive motion samples one topomer every ~3 nanoseconds. This suggests that a 100-residue protein can find its native topomer (the topomer containing the native conformation) by *random* sampling in ~100 milliseconds. This is comparable to the experimentally observed timescale required for a denatured protein domain to reestablish its native structure. These results suggest that for a 100-residue protein (an average sized protein domain), the folding from a denatured form *can* proceed in a two stage folding process consisting of:

- **Topomer diffusion**: random, diffusive sampling to find the native topomer, followed by
- **Intra-topomer ordering**: non-random, local conformational changes within the native topomer to find the unique native state.

Our results suggest that the topomer diffusion step requires ~100ms for a 100-residue protein. We expect that the time required for intra-topomer ordering may be more rapid than the topomer diffusion stage, leading to a cooperative, two-state folding mechanism [9, 10], or comparable to the topomer diffusion stage, leading to multi-state folding kinetics.

METHODS

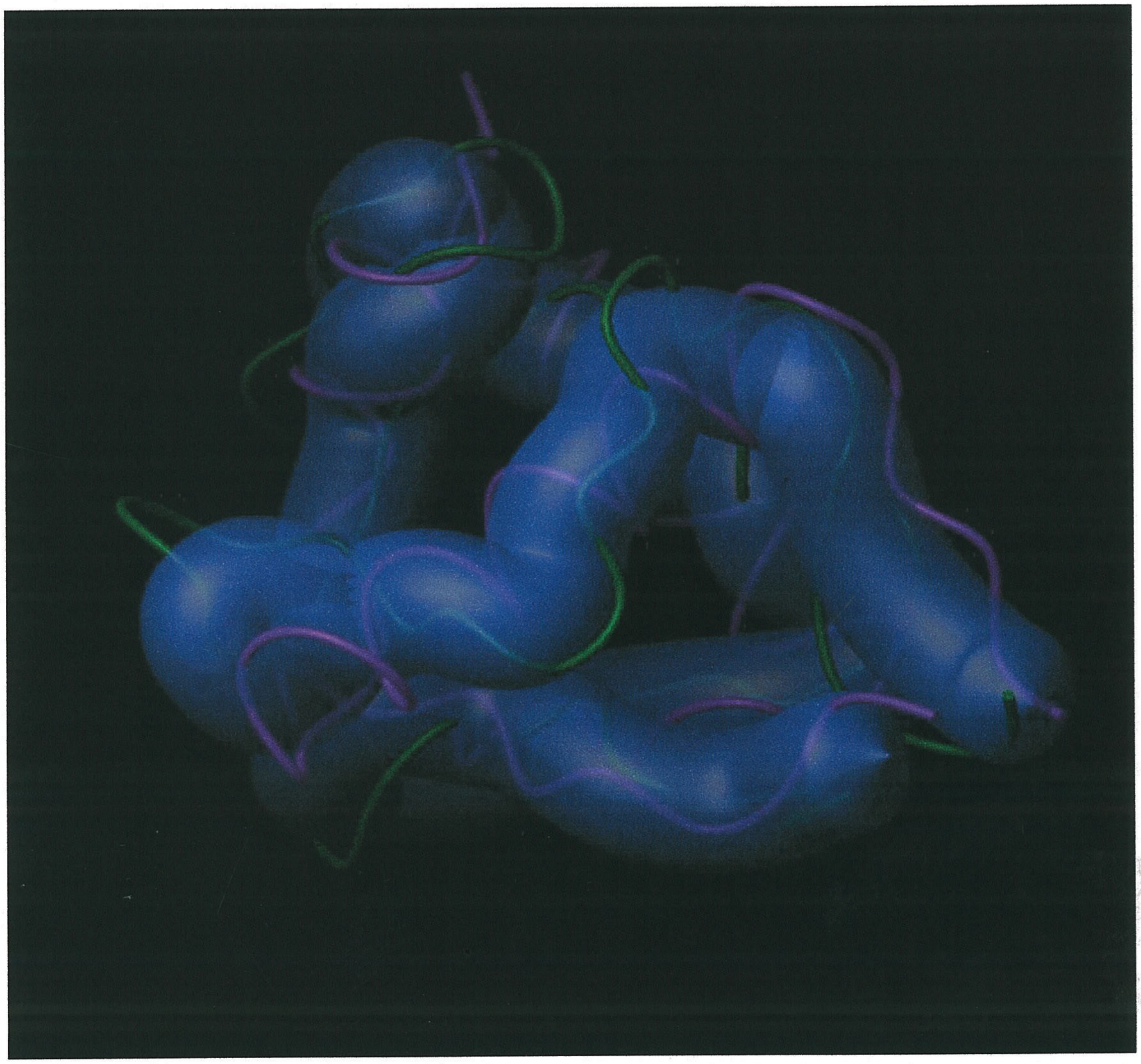
We want to estimate the number of disjoint topomers required to span all possible compact and semi-compact conformations for a polypeptide of length N. To do this we use the Generic Protein Direct Monte Carlo procedure (GP) described below to generate large ensembles of self-avoiding protein conformations. We compare each conformation to a test set of ~20 dissimilar native protein structures of length N and determine if it is topomeric to any of the test proteins. This process is continued until we have generated at least one topomeric match to each and every one of the ~20 test proteins. The number of conformations generated at this point is a measure of the total number of disjoint topomers for an N-residue polypeptide.

Definition of a topomer. We define two protein conformations to be *topomeric* if they have the same backbone topology [11], that is, if one conformation is obtainable from the other through local backbone coordinate transformations that:

- do not require cooperative movements between non-local residues and
- do not disrupt the overall compactness of the structure or covalent bonding of the peptide backbone.

We define a *topomer* as the set of all conformations topomeric to a particular conformation. Thus, a topomer is a bundle of conformations sharing the same backbone topology. The *native topomer* for a protein consists of all conformations topomeric to the native conformation. Figure 2.1 shows an example of two topomeric structures. We present below a simple algorithm to test whether two conformations are topomeric.

Figure 2.1 Example of two topomeric protein structures, one shown in purple, the other in green. Any structure lying within the larger blue tube would be considered topomeric to each of these structures as well.



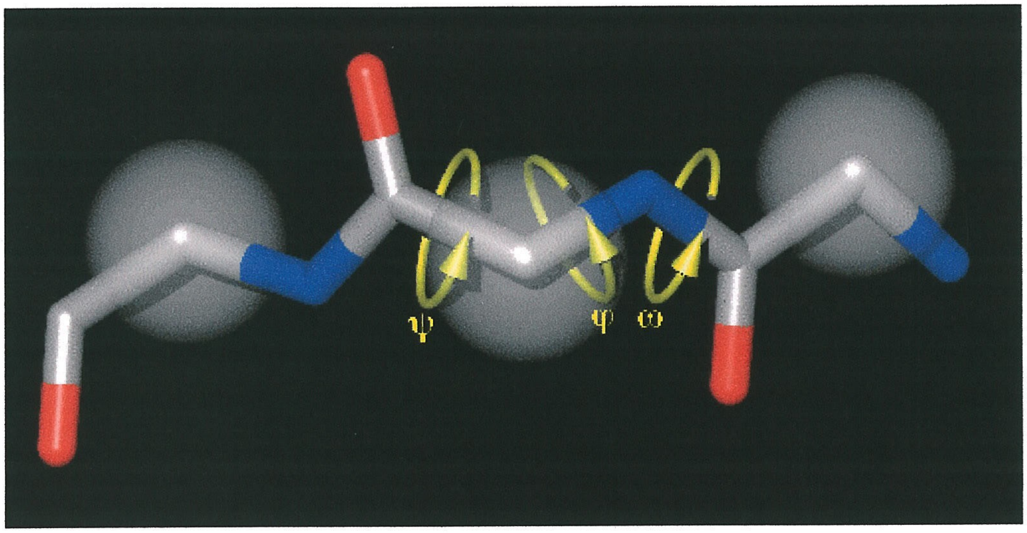
The Native Protein Test Sets. The native test proteins were compiled from the CATH protein domain database (http://www.biochem.ucl.ac.uk/bsm/cath) [12]. In order to have at least 20 test structures for each protein length N, we included longer structures truncated at the carboxyl terminus. For example, our test set for N = 45 consists of residues 1-45 from available protein structures with lengths of 45-49. In instances where the coordinate file contained more than one set of coordinates for a given structure, we used the first set. The 22 proteins in the test set for N = 100 are listed here by their Brookhaven Databank or CATH domain classification name: 1aaj, 1ab2, 1acx, 1bet, 1cmbA, 1etc, 1fd2, 1fkb, 1fus, 1hks, 1hrc, 1ltsD, 1onc, 1pal, 1put, 1thx, 1tlk, 1ycc, 2atcB, 2cdv, 2imn, and 2pna. The complete list for each N is available at http://www.wag.caltech.edu/home/derek/gp.

Generic Protein (GP) Direct Monte Carlo Method. The GP direct Monte Carlo method employs the CCBB Direct Monte Carlo [13] procedure in conjunction with a protein representation where:

- six (ϕ,ψ) backbone torsion pair choices [14] are allowed for each residue (the torsion about the peptide bond is fixed at 180°, and all bonds and angles have fixed standard values [15]), and
- a simple 12-6 Lennard-Jones potential is used to account for both the excluded volume and the cohesion of each residue (*identical for all amino acids*).

This representation is shown in Figure 2.2.

Figure 2.2 The generic protein representation.


ψ
φ
ω

A GP conformation is constructed by adding residues one by one (alternating right and left) to a single residue-starting fragment located at the center of the protein sequence. During buildup, the probability of selecting one of the six (ϕ,ψ) candidates is given by

$$P_j = \frac{\exp\left({}^{-E_j}\!/_{kT} \right)}{\sum_{i=1}^{6} \exp\left({}^{-E_i}\!/_{kT} \right)}. \tag{1}$$

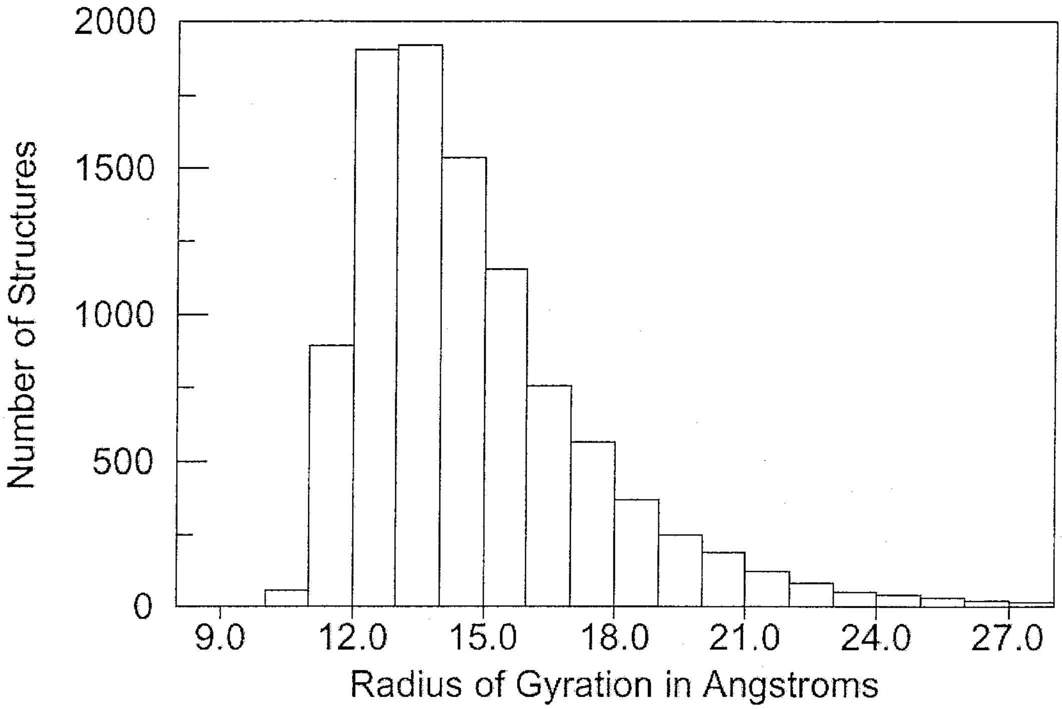
The addition energy, E_i, of a single residue is given by the summation of its pair-wise interaction energies with each residue in the polypeptide fragment. For all amino acids, the energy of a residue pair is

$$E_{ij}(R) = E_0\left[\left(\frac{R}{R_0}\right)^{12} - 2\left(\frac{R}{R_0}\right)^{6}\right], \tag{2}$$

where $R_0 = 5.5$Å, $E_0 = 0.15$ kcal/mol, and R is the distance between the α-carbon of each residue. Here i and j include all pairs within a cutoff of 10Å but excluding nearest and next-nearest neighbors in the sequence. Energetically favorable addition steps are replicated by a factor $m = \mathrm{int}[(z_i/\langle z_i\rangle)/(z_{i-1}/\langle z_{i-1}\rangle)]$, where $z_i = \exp(-E_i/kT)$ and $\langle z_i\rangle$ denotes the average value of z at residue i over all generated chains, according to the CCBB [13] procedure.

The parameter values $R_0 = 5.5$Å and $E_0 = 0.15$ kcal/mol were selected because they yield an ensemble of generic folds with about the same distribution for the radius of gyration found in the protein databank. For the GP ensemble of 100-residue conformations, half have a radius of gyration between 12Å and 15Å (Figure 2.3), the observed range for the radius of gyration for 100-residue globular proteins [16]. The GP ensemble is 10% more compact than 12Å, while 40% of the structures are less compact than 15Å. Thus, the GP procedure rapidly generates a diverse ensemble of compact and semi-compact protein chains with realistic peptide backbone geometries [>10^6 conformations for a 50 residue protein are generated in one day on a single processor Silicon Graphics Inc. (SGI) R10000 workstation]. Since no information about sequence identity is included in the GP energy expression, the GP ensemble is a generic, sequence independent set of self-avoiding polypeptide conformations.

Figure 2.3 Radius of gyration histogram for ten thousand 100-residue structures generated by the GP method. Compact globular protein structures 100 residues in length typically have a radius of gyration between 12Å and 15Å [16]. One half of the GP structures are within this range, with only 10% of the GP structures more compact.


2000
1500
1000
500
0
Number of Structures
9.0
12.0
15.0
18.0
21.0
24.0
27.0
Radius of Gyration in Angstroms

Determining the number of distinct topologies for an N-residue polypeptide. We determined the number of distinct topologies for an N-residue polypeptide by calculating how many GP structures must be generated in order to obtain a topomeric match to each of ~20 dissimilar native test proteins of length N. As each GP structure was generated, we calculated its CRMS (α-carbon root mean square deviation [17]) distance from each structure in the native protein test set. Every GP structure with a relatively low CRMS to any of the test structures was saved along with the point at which it was generated. Thus after generating a large ensemble of GP structures, we retained a small subset of structures (typically 100) with a low CRMS difference to each native test structure. [It was necessary to save many structures for subsequent analysis, since a low CRMS difference does not necessarily imply that two structures are topomeric.]

From the retained sets of structures, we used the Native Topomer Test Procedure to verify which structures (if any) were topomeric to each native test structure. First, each candidate GP backbone was optimally superimposed onto the corresponding native test structure. Next, each α-carbon in the candidate GP backbone was tethered with a harmonic constraint [using a force constant of 5 (kcal/mol)/A^2] to the coordinates of the same α-carbon in the native test structure. Conjugate gradient minimization (200 steps) was then performed on the constrained GP backbone (using Dreiding [15] force-field parameters). During minimization, each α-carbon in the GP structure attempts to follow a direct, non-cooperative trajectory toward the corresponding native α-carbon. Topology differences are easily observed by the inability of the GP structure to minimize to the native coordinates, since the force-field parameters do not permit covalent bond breakage in the peptide backbone. Using this automated method, it is possible to determine quite quickly whether a retained GP structure is topomeric to the corresponding native test structure. Note that the Native Topomer Test Procedure is simply a computational test to determine if two structures are topomeric. This Procedure does not accurately simulate how a protein finds its precise native state once it has found its native topomer. However, the Test Procedure minimization trajectories followed by the GP structures to their corresponding native states are useful for visualizing the conformational differences that two topomeric structures may possess. QuickTime movies of the minimization trajectories for all 277 native test structures are available at http://www.wag.caltech.edu/home/derek/gp.

The GP algorithm does not include any mechanism to prevent the generation of more than one structure for each topology. Thus, by the point at which all 22 test proteins had been matched for the N=100 calculation, we had found an average of about 5 matches for each test protein. This suggests that our measurement slightly overestimates the number of distinct topologies. On the other hand, the use of a finite number (~20) of test systems may underestimate the number of GP structures required to generate a topomeric match to topologies more complex than any of the test proteins. We expect that these factors balance each other. The calculated number of topomers (Figure 2.4) increases monotonically with the number of residues despite completely independent choices of the native protein test sets. This suggests that the estimate has systematic inaccuracies well less than an order of magnitude.

Figure 2.4 The number of disjoint topomers estimated for an N-residue polypeptide. Beyond N = 50, the number of topomers, S_N, scales as $S_N = (83936)\mathbf{x}(1.0624)^N$. For N=100, the number of topomers is ~$(1.19)^N$.

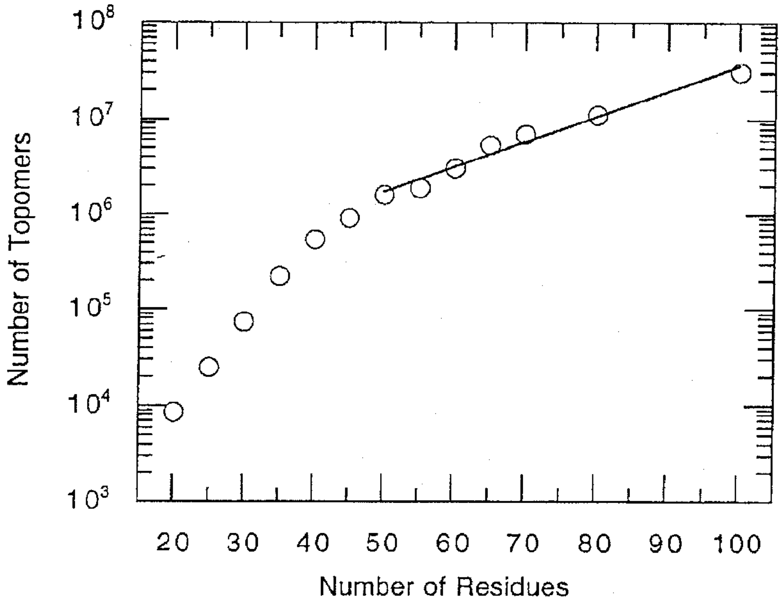

10^8
10^7
10^6
10^5
10^4
10^3
Number of Topomers
20
30
40
50
60
70
80
90
100
Number of Residues

Figure 2.5 Comparisons of the native conformations (purple) with their topomeric counterparts from the generic structure sets (yellow). In order to facilitate viewing, the local geometry of each generic conformation has been refined to incorporate native helix and β-strand segments while preserving the tertiary fold topology. This refinement is demonstrated in (a), where the generic structure (left, in yellow) is refined using the native helix assignment (right, in yellow).

(a) The 65-residue segment from the NMR determined structure of the proteolytic fragment from Bacteriorhodopsin [18] (1bct). This example is one of many semi-compact test folds that was topomerically matched by a GP structure. Thus, our estimate considers semi-compact as well as compact topomers.

(b) 65-residue Porcine C5a$_{\mathrm{desArg}}$ (1c5a) [19].

(c) 80-residue fragment from acyl-coenzyme A binding protein (1aca) [20].

(d) 80-residue segment from domain four of the N-terminal domain of 70kD Heat-Shock Cognate protein (1hpm04) [21].

(e) 100-residue segment from Heat Shock Transcription Factor (1hks) [22].

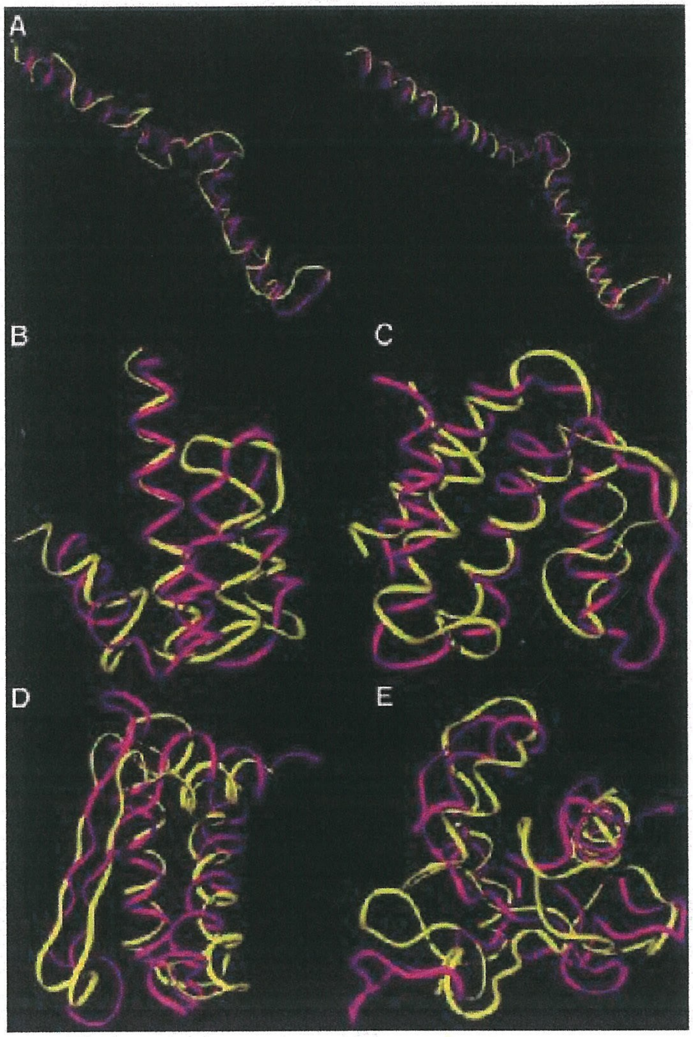
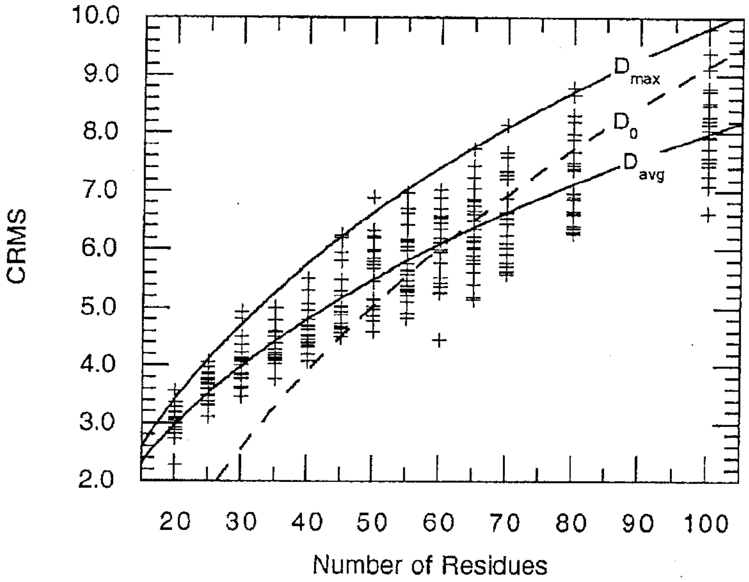

A
B
C
D
E

Figure 2.6 The CRMS between each of the 277 native test conformations and their topomeric matches from the generic structure sets. The dashed line in the figure represents a previously developed average threshold for topological similarity developed by Maiorov and Crippen [11]. They found that two N-residue structures are topologically similar when their CRMS is below the threshold, $D_0 = a + b\ (N)^{1/3}$, where a = -10.82 ± 0.37 and b = 4.31 ± 0.08. For $N \geq 50$, the CRMS values we obtained from topomeric matches correlate well with the Maiorov-Crippen D_0 threshold for topological similarity. Fitting a similar functional form to the average and maximum of our CRMS data for topomeric conformations yields D_{avg} (a = -4.12 ± 0.24; b = 2.61 ± 0.06) and D_{max} (a = -5.62 ± 0.40; b = 3.33 ± 0.11), respectively.


10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
CRMS
20
30
40
50
60
70
80
90
100
Number of Residues
D_{max}
D_0
D_{avg}

RESULTS AND DISCUSSION

Total Number of Topomers. Figure 2.4 shows the number of topomers estimated for polypeptides of length 20-100. For N = 55 to 100, the number of topomers scales as $(1.06)^N$, even though the number of distinct conformation states scales at least as fast as 3^N. For N = 100, we find ~3×10^7 topomers, a large number, but vastly smaller than $3^{100}\approx10^{48}$. Visual comparisons between some of the test structures and the topomeric GP structures are shown in Figure 2.5.

Estimates of Folding Times. Next, we estimated how long it would take a protein to randomly sample all of its compact and semi-compact topomers. Figure 2.6 shows the CRMS between each of the 277 conformations in the native protein test sets and its topomeric match in the ensemble of GP structures. For 100-residues there is a maximal CRMS distance of 9.8Å between each native test protein and its topomeric conformation in the GP set. This indicates that the greatest distance between any two conformations in the same topomer is ~9.8Å CRMS. Thus, any two conformation more than ~9.8Å CRMS from each other are necessarily members of different topomers. Hence, the maximum distance between neighboring, yet disjoint topomer is ~9.8Å CRMS. To estimate the sampling timescale we use the three-dimensional Einstein diffusion equation,

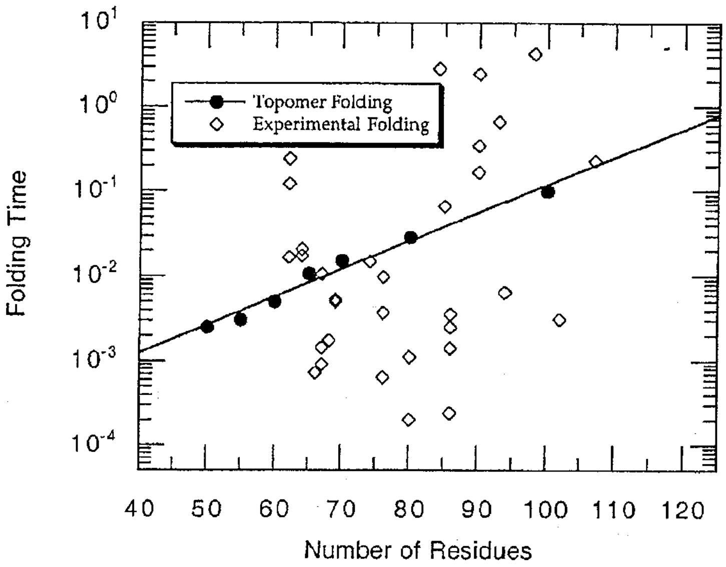
$$\tau = \bar{x}^2 / 6D \,, \tag{3}$$

where $\bar{x}$ is the CRMS between neighboring, disjoint topomers, D is the diffusion coefficient, and τ is the topomer-sampling time. Eaton and coworkers [23] determined that $D\approx5\times10^{-7}$ cm^2/sec for extensive intra-chain protein motion in cytochrome c folding. Using this value for D in equation (3), with $\bar{x}$ = 9.8Å, suggests that the topomer-sampling time for N=100 is $\tau \approx 3.2$ ns. Given ~3×10^7 topomers and an average topomer-sampling rate of one topomer every ~3.2ns, we estimate that a 100-residue protein can *randomly* sample all compact and semi-compact topomers in ~100 milliseconds.

Similar estimates for other N (using the maximum CRMS for each N from Figure 2.6 and the number of topomers for each N from Figure 2.4) lead to the plot in Figure 2.7. In this plot, the solid circles represent the time estimated for a polypeptide to randomly sample all of its topomers (for N= 50, 55, 60, 65, 70, 80, and 100), and the solid line is the exponential fit through these points.

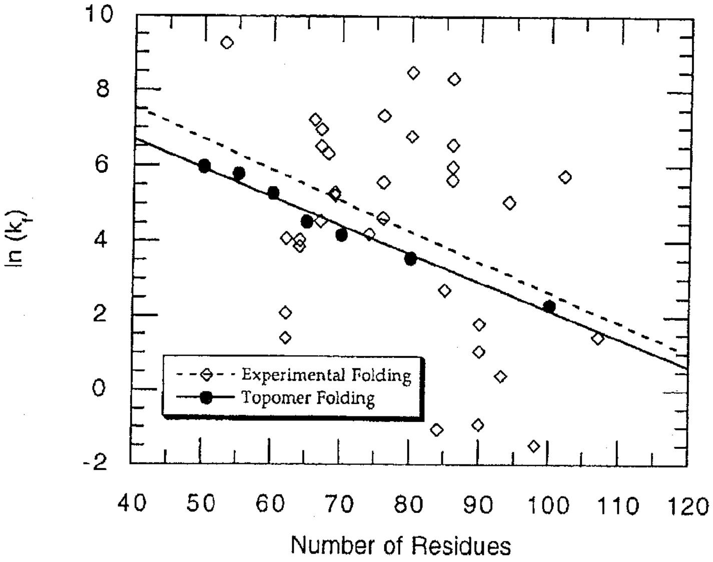
It is interesting to compare the folding timescales predicted by the topomer-sampling model with experimentally determined folding times. The open diamond points in Figure 2.7 represent 32 experimentally determined folding times (time = $1/k_f$) for single domain, two-state folding proteins compiled in Table 1 of a recent review by S. E. Jackson [24]. The predicted topomer-sampling model timescale (10^{-3}-10^{0} seconds) correlates well with the experimentally determined folding times. Note that the correct folding timescale is achieved in our model without using any tunable parameters (the topomer folding timescale is determined directly from the number of topomers, the distance between topomers, and an experimentally determined intra-chain diffusion constant). [Table 1 in reference [24] contains 38 folding rates for small, monomeric proteins that fold with two-state kinetics. Six of these rates were considered unsuitable for this plot and were excluded: λ-repressor (native helix stabilizing mutations), Arc repressor (two domains connected by a linker), Villin 14T (greater than 120 residues), and the three cytochrome c variants (heme-containing).]

Figure 2.7 The dark circles represent the estimated time in seconds for a polypeptide of length N to randomly sample all of its topomers. This is based on the results in Figures 2.4 and 2.6 combined with equation (3) using the experimentally derived diffusion constant, $D=5\times10^{-7}$ cm^2/sec. The solid line is the best fit to these first principles predicted topomer sampling times. It leads to a topomer sampling folding time, t_{fold}(seconds) = $(5.98\times10^{-5})\times(1.079)^N$. The open diamond points are 32 experimentally determined folding timescales (time = $1/k_f$) for single domain proteins less than 120 residues in length compiled in Table 1 of a recent review by S. E. Jackson [24]. The predicted topomer-sampling model timescale (10^{-3}-10^0 seconds) correlates well with the experimentally determined folding times.



Folding Time
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
Topomer Folding
Experimental Folding
40
50
60
70
80
90
100
110
120
Number of Residues

Figure 2.8 The timescale data in (a) replotted as the natural log of the intrinsic folding rate, $\ln(k_f)$. The dashed line is the best exponential fit through the experimental folding rate points. The p-value for this fit is p=0.082, suggesting that there is only a 1 in 12 chance that a correlation with this significant a slope would appear by chance. Thus, the topomer sampling model (solid line) predicts the correct magnitude and length dependence (slope), for the folding rates of two-state folding proteins without using any adjustable parameters.


10
8
6
4
2
0
-2
ln (k_f)
40
50
60
70
80
90
100
110
120
Number of Residues
- -◇- - Experimental Folding
—●— Topomer Folding

In Figure 2.8, we replot the timescale data in Figure 2.7 as the natural log of the intrinsic folding rate, $\ln(k_f)$. Experimental folding times can vary by 3 orders of magnitude for proteins of similar length (even for homologous sequences [25]), suggesting that factors independent of protein length (such as topological complexity [26] and sequence mutation) drastically affect the rate of protein folding. However, we expect that these factors average out over the different proteins in the experimental data set. Hence, the best exponential fit through these experimental points (the dashed line in Figure 2.8) is a reasonable estimate of the length dependent part of the protein folding timescale. The p-value for this fit is p=0.082, implying that there is only a 1 in 12 chance that a correlation with this significant a slope would appear by chance (see [26] for a detailed explanation of p-values in this context). Remarkably, the predicted topomer-sampling timescale (solid-line) and the apparent length dependent part of the experimental folding timescale (dashed line) are in excellent agreement. Thus, the topomer sampling model (solid line) predicts the correct magnitude and the correct length dependence (slope), for the folding rates of two-state folding proteins without using any adjustable parameters.

Folding Mechanisms. Our results suggest that an average sized protein domain can find its native topology without any mechanisms to simplify the conformational search [27, 28]. Thus, the topomer-sampling model is fundamentally different from folding models that insist that regions of correctly folded structure form during the early stages of protein folding, before a structure with the native topology has been sampled. The topomer-sampling model suggests that the condensation of specific native contacts [29] is not required to simplify the search for the native topomer. Furthermore, the topomer-sampling model suggests that early nucleation of native secondary structure [30, 31] is not essential for an average sized domain to fold. Indeed, the 86 amino acid reduced HIV-1 Tat (*trans*-activator) protein [32] folds on a biologically relevant time frame to a structure with a well-defined core, yet possesses no secondary structure or disulfide bonds.

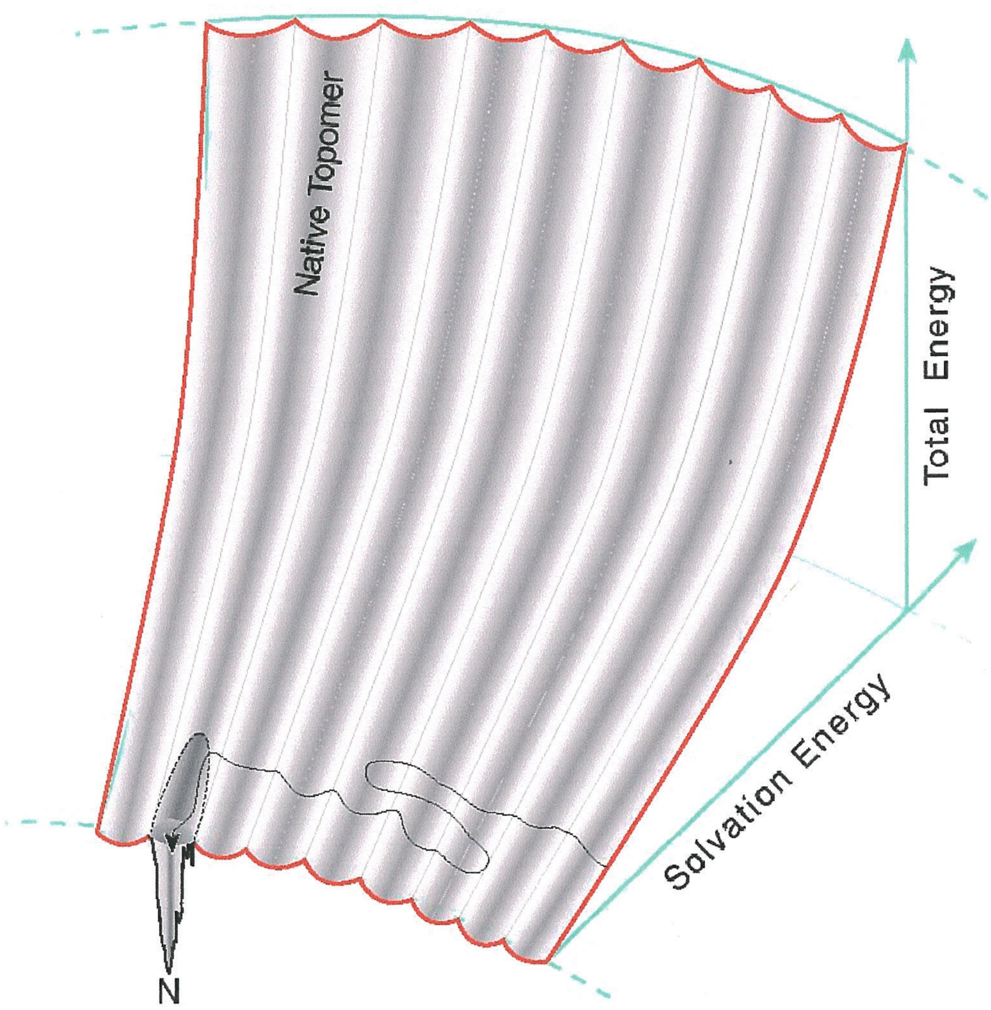
For large protein domains (longer than ~120 residues), our results imply that some type of early nucleation or condensation mechanism is required for the native topomer to be found in less than a second (Figure 2.7). Indeed, we expect that for many large proteins (especially those with high helical content), such mechanisms greatly expedite the search for the native topology and lead to folding rates that are faster than those found in small proteins (because small proteins may not require early nucleation or condensation mechanisms to fold, such mechanisms may not have evolved in short sequences to the degree that they have in long ones). Experiments have shown that native-like secondary structure is found in the kinetic folding intermediates of many larger proteins [33] and in fragments excised from proteins [34, 35]. Such moderate local structural biases probably help large domains find the native topology by reducing the complexity of the search for the native topomer. These biases certainly help proteins of all sizes find their precise native conformation once they have found the native topomer.

The Folding Landscape. To this point, we have treated the energy landscape outside the native topomer as flat, yet rugged, like a golf course [36]. However, calorimetric studies [37] and experiments using the hydrophobic fluorescent probe ANS [38] show that a significant portion of the nonpolar surface area that is buried in the native state is also buried in partially folded structures. Thus, the hydrophobic effect operates on the protein long before the protein has found its native topology, and conformations with poor solvation energies [39] are not sampled during the search for the native topomer.

However, the fact that a protein only samples conformations with favorable solvation energies need not drastically limit the number of topologies searched. Two structures within the same topomer can have very different solvation energies, since small perturbations in the backbone conformation can drastically affect the orientation of the side chains with respect to the interior of the overall fold. Thus, one can easily construct a conformation that is topomeric to the native structure such that the nonpolar sidechains are directed away from the core and the polar sidechains are buried in the interior. Conversely, most compact and semi-compact topomers contain conformations such that the nonpolar sidechains are properly directed into the interior, and the polar sidechains extend into the solvent. A protein will tend to sample good solvation energy structures within each topomer.

Figure 2.9 presents a diagram for the folding energy landscape that simultaneously illustrates these ideas about the variability of solvation energies and the similarity of conformational states within a single topomer. The folding energy landscape is shaped like the seating in the Rose Bowl. The total energy is given by the height of the stadium. Conformations with poor solvation energy are situated far away from the playing field, while conformations with favorable solvation energies are situated close to the field. The conformations within one topomer are distributed in a single, columnar section in the stadium (the complete energy landscape for a 100-residue polypeptide contains 3×10^7 topomer columns). Thus, each topomer contains conformations with both very poor and very favorable solvation energies. As a protein folds, it samples different topomers by randomly sampling the favorable solvation energy states. When the protein samples a conformation in the native topomer, the native funnel directs the protein to its unique native structure.

Figure 2.9 A representation of the folding energy landscape suggested by the topomer-sampling model. This diagram indicates that structures within the same topomer have a variety of solvation energies (shown along the radial axis). The landscape is shaped like the seating in the Rose Bowl. The total energy is given by the height in the stadium. Conformations with poor solvation energy are situated far from the playing field, while conformations with favorable solvation energies are situated close to the field. The conformations within a single topomer are distributed in a single, columnar section of the stadium. For a 100-residue polypeptide, the complete folding energy landscape contains 3×10^7 such topomer columns. On this topomer folding diagram, the topomer-sampling model of protein folding is a meandering trajectory (black line with arrowhead) that travels from topomer to topomer, sampling only favorable solvation energy conformations within each topomer. When the protein samples a conformation within its native topomer, specific favorable hydrogen bonding and core packing interactions (represented by a funnel within the native topomer) direct the protein to its unique native structure (N). We show this funnel connected to only a part of the space spanned by the native topomer to indicate that only the favorable solvation energy structures in the native topomer are near the native funnel. Thus, mutations which affect the solvation properties of the protein can drastically affect the time required for a protein to find its native funnel (see text). On this diagram, an early folding nucleation event decreases the number of topomer columns that must be sampled, thereby decreasing the folding rate (by whatever fraction of the total number of topomers is eliminated).


Native Topomer
Total Energy
Solvation Energy
N

In the topomer-sampling model, even though an average sized protein is assured of randomly sampling *some* conformation in the native topomer, there is no guarantee that this conformation will be within the clutches of the native folding funnel. We believe that the hydrophobic effect plays a key role in ensuring that when a protein samples a conformation in the native topomer, its sidechain and hydrogen bond donor orientations will be appropriate for a cooperative collapse to the native state.

In the complete absence of a hydrophobic effect, the solvation energy dimension of the folding energy landscape collapses (Figure 2.9), so that the folding energy landscape becomes a flat, rugged surface. In such a scenario, the line representing the protein folding trajectory is not confined to the lower levels of a stadium-like surface but is allowed to wander over an entire flat landscape, precluding the protein from finding the native folding funnel on a tractable timescale. In this manner, we expect that disruptions in the solvation properties of a protein (by changing the solvent or making sequence mutations) will drastically influence the time it takes to find the native funnel and consequently have a large effect on the overall folding rate. Consistent with this, numerous experiments have demonstrated that there is a strong correlation between protein folding rates and protein stability across differing solvent conditions [40], and that stability is a significant determinant of the relative kinetics of homologous proteins [25, 41, 42].

Our estimate of the folding timescale as the time it takes to randomly sample all compact and semi-compact topomers assumes that each topomer contains one or more conformations of favorable solvation energy and that each topomer is sampled as the protein moves between favorable solvation energy conformations. Barron and coworkers

[43, 44] have recently used Raman optical activity experiments to show that residues in disordered regions in molten globule states "flicker" between the allowed regions of the Ramachandran plot at rates of ~$10^{12}s^{-1}$. This suggests that local polypeptide chain dynamics can accommodate very fast equilibration to low solvation energy conformations without disturbing the tertiary topology. We have not yet evaluated the solvation energy for all possible conformations of a 100-residue polypeptide. Hence, we do not yet know how many topomers do not contain any conformations with favorable solvation energies. However, we believe that it is not a significant fraction (probably less than a factor of 100), because our assumption that all semi-compact and compact topomers are sampled correlates well with experimental folding rate data.

CONCLUSION

We find that partitioning conformation space into sets of topologically equivalent conformations (topomers) allows us to understand how proteins can fold to native structures on a sub-second timescale. Our results suggest that average sized protein domains (<120 residues) *can* fold by a two-step process:

- **Topomer diffusion**: a random, diffusive search for a conformation with the native topology (~ 0.1 sec for 100 residues) followed by
- **Intra-topomer ordering**: a non-random, "funneled" local conformational search for the precise native state.

Thus early protein folding *can* be a highly dynamic, diffusive process. This highly dynamic mechanism for folding is consistent with recent experiments showing that the rate of protein folding is strongly dependent on the viscosity of the solvent [45], [46], [47]. Resolving the exact details of these early folding processes requires monitoring protein folding in the microsecond time regime.

This dynamic picture of early folding is also consistent with the phenomenon of prions [48], proteins that apparently have more than one stable conformation. The topomer-sampling model suggests that numerous non-native topologies are explored before the native topology is sampled. It is quite conceivable that there could be more than one topology containing a funnel with the correct properties to yield a kinetically trapped folded state. Evidently, evolution has selected for protein sequences that have only one such funnel and hence fold to a singular native state at biological temperatures.

REFERENCES

1. Šali, A., Shakhnovich, E., & Karplus, M. (1994), *Nature* **369**, 248-251.

2. Levinthal, C. (1969) in *Mossbauer Spectroscopy in Biological Systems*, eds. Debrunner, P., Tsibris, J. C. M., & Münck, E. (Univ. Illinois Press, Urbana), pp. 21-24.

3. Zwanzig, R., Szabo, A., & Bagchi, B. (1992), *Proc. Natl. Acad. Sci.* **89**, 20-22.

4. Levinthal, C. (1968) *J. Chim. Phys.* **65**, 44-45.

5. Wolynes, P. G., Onuchic, J. N., & Thirmulai, D. (1995), *Science* **267**, 1619-1620.

6. Dill, K. A. & Chan, H. S. (1997), *Nat. Struct. Biol.* **4**, 10-19.

7. Baldwin, R. L. (1994), *Nature* **369**, 183-184.

8. Karplus, M. (1997), *Folding and Design* **2**, S69-S75.

9. Jackson, S. E. & Fersht A. R. (1991), *Biochem.* **30**, 10436-10443.

10. Creighton, T. E. (1993), in *Proteins*, (W. H. Freeman and Company, New York), pp. 290-291.

11. Maiorov, V. N. & Crippen, G. M. (1994), *J. Mol. Biol.* **235**, 625-634.

12. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., & Thornton, J.M. (1997), *Structure* **5**, 1093-1108.

13. Sadanobu, J. & Goddard III, W. A. (1997), *J. Chem. Phys.* **106**, 6722-6729.

14. Rooman, M. J., Kocher, J. P., & Wodak, S. J. (1992), *J. Mol. Biol.* **221**, 961-979.

15. Mayo, S. L., Olafson, B. D., & Goddard III., W. A. (1990), *J. Phys. Chem.* **94**, 8897-8909.

16. Maiorov, V. N. & Crippen, G. M. (1995), *Prot. Struct. Func. Gen.* **22**, 273-283.

17. Kabsch, W. & Sander, C. (1978), *Acta Crystallog. Sect. A* **34**, 827-828.

18. Barsukov, I. L., Nolde, D. E., Lomize, A. L., & Arseniev, A. S. (1992), *Eur. J. Biochem.* **206**, 665-672.

19. Williamson, M. P. & Madison, V. S. (1990), *Biochem.* **29**, 2895-2905.

20. Kragelun, B. B., Anderson, K. V., Madsen, J. C., Knudsen, J. & Poulsen, F. M. (1993), *J. Mol. Biol.* **230**, 1260-1277.

21. Wilbanks, S. M. & McKay, D. B. (1995), *J. Biol. Chem.* **270**, 2251-2257.

22. Vuister, G. W., Kim, S. J., Orosz, A., Marquardt, J., & Bax, A. (1994), *Nat. Stuct. Biol.* **1**, 605-614.

23. Hagen, S. J., Hofrichter, J., Szabo, A., & Eaton, W. A. (1996), *Proc. Natl. Acad. Sci.* **93**, 11615-11617.

24. Jackson, S. E. (1998), *Folding & Design* **3**, R81-R91.

25. Plaxco, K. W., Spitzfaden, C. Campbell, I. D., & Dobson, C. M. (1997), *J. Mol. Biol.* **270**, 763-770.

26. Plaxco, K. W., Simons, K. T., & Baker, D. (1998), *J. Mol. Biol.* **277**, 985-994.

27. Guijarro, J. I., Morton, C. J., Plaxco, K.W., Campbell, I. D., & Dobson, C. M. (1998), *J. Mol. Biol.* **276**, 657-667.

28. Dill, K. A. (1985), *Biochemistry* **24**, 1501-1509.

29. Fersht, A. R. (1997), *Curr. Opin. Struct. Biol.* **7**, 3-9.

30. Karplus, M. & Weaver, D. L. (1976) *Nature* **260**, 404-406.

31. Wetlaufer, D. B. (1973), *Proc. Natl. Acad. Sci.* **70**, 697-701.

32. Bayer, P., Kraft, M., Ejchart, A., Westendorp, M., Frank, R., & Rösch, P. (1995), *J. Mol. Biol.* **247**, 529-535.

33. Baldwin, R. L. (1993), *Curr. Opin. Struct. Biol.* **3**, 84-91.

34. Brown, J. E. & Klee, W. A. (1971), *Biochemistry* **10**, 470-476.

35. Bierzynski, A. P., Kim, S., & Baldwin, R. L. (1982), *Proc. Natl. Acad. Sci.* **79**, 2470-2474.

36. Bryngelson, J. D. & Wolynes, P. G. (1989), *J. Phys. Chem.* **93**, 6902-6915.

37. Parker, M. J., Lorch, M., Sessions, R. B., & Clarke, A. R. (1998), *Biochem.* **37**, 2538-2545.

38. Ptitsyn, O. B., Pain, R. H., Semisotnov, G. V., Zerovnik, E., & Razgulyaev, O. I. (1990), *FEBS Lett.* **262**, 20-24.

39. Eisenberg, D. & McLachlan, A. D. (1986), *Nature* **319**, 199-203.

40. Chen, B. L., Baase, W. A., Nicholson, H. & Schellman, J. A. (1992), *Biochem.* **31**, 1464-1476.

41. Mines, G. A., Pascher, T., Lee, S. C., Winkler, J. R., and Gray, H. B. (1996), *Chem. & Biol.* **3**, 491-497.

42. Plaxco, K. W., Guijarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D. & Dobson, C. M. (1998), *Biochem.* **37**, 2529-2537.

43. Wilson, G., Hecht, L. & Barron, L. D. (1996), *Biochem.* **35**, 12518-12525.

44. Barron, L. D., Hecht, L., & Wilson, G. (1997), *Biochem.* **36**, 13143-13147.

45. Plaxco, K. W. & Baker, D. (1998), *Proc. Natl. Acad. Sci.* **95**, 13591-13596.

46. Jacob, M., Schindler, T., Balbach, J., & Schmid, F. X., (1997), *Proc. Natl. Acad. Sci.* **94**, 5622-5627.

47. Creighton, T. E. (1997), *Curr. Opin. Struct. Biol.* **7**, R380-R383.

48. Prusiner, S. B. (1997), *Science* **278**, 245-251.

CHAPTER 3

PROTEIN FOLD DETERMINATION FROM SPARSE DISANCE RESTRAINTS: THE RESTRAINED GENERIC PROTEIN DIRECT MONTE CARLO METHOD

ABSTRACT

We present the *generate-and-select* hierarchy for tertiary protein structure prediction. The foundation of this hierarchy is the Restrained Generic Protein (RGP) Direct Monte Carlo method. The RGP method is a highly efficient off-lattice residue buildup procedure that can quickly generate the complete set of topologies that satisfy a very small number of inter-residue distance restraints. For 3 restraints uniformly distributed in a 72-residue protein, we demonstrate that the size of this set is ~10^4. The RGP method can generate this set of structures in less than one hour using a Silicon Graphics R10000 single processor workstation. Following structure generation, a simple criterion that measures the burial of hydrophobic and hydrophilic residues can reliably select a reduced set of ~10^2 structures that contains the native topology. A minimization of the structures in the reduced set typically ranks the native topology in the five lowest energy folds. Thus, using this hierarchical approach, we suggest that *de novo* prediction of moderate resolution globular protein structure can be achieved in just a few hours on a single processor workstation.

INTRODUCTION

Given the difficulty of protein structure prediction, it is important to simplify the problem using prediction approaches that incorporate predicted or experimentally determined structural information. For many prediction targets, distance restraints are available from labeling experiments, disulfide bond connectivity, or preliminary NMR data. Furthermore, methods exist for predicting local structural characteristics[1] such as residue contacts,[2, 3] secondary structure[4, 5] and accessible surface area,[6] and surface turns.[7]

Dewitte *et al.*[8] established the basic feasibility of obtaining fold predictions using a limited amount of distance information. They developed a method to exhaustively enumerate all walks on a diamond lattice consistent with a set of lattice pair restraint conditions. Their work demonstrated that as few as one restraint per residue could successfully limit the number of possible walks (conformations) to $\sim 10^3$. Unfortunately, the method was not computationally feasible when the number of restraints was small compared to the number of lattice steps (residues).

Since this original work, several different methods have been applied to the problem of structure prediction using a small number of distance restraints. Aszódi *et al.*[9] developed a distance-geometry-based approach that incorporated distance restraints and native secondary structure assignments as well as knowledge-based criteria such as backbone connectivity, hydrophobicity, and conservation data obtained from multiple alignments. This method efficiently generated structures with the correct topology using as few as N/10 restraints for very simple protein topologies, and ~N/4 restraints for more complex folds.

Another approach is a dynamic Monte Carlo (MC) method, MONSSTER,[10] that folds random coil conformations using an energy function incorporating secondary structure and distance restraint information. Recently, this method achieved low-resolution structures (typically 5-6Å CRMS) for a number of small proteins when used in conjunction with secondary structure and residue contact predictions.[11] However, since the algorithm is a dynamic procedure, generating a single structure that satisfies the restraints requires overnight simulation.

The results obtained by distance-geometry and dynamic MC suggest that knowing the correct secondary structure and ~N/4 distance restraints leaves a very small number of possibilities for the topology of a polypeptide. In both methods, coupling this distance information with simple energy criteria usually resulted in an unambiguous determination of the native fold topology. Thus each method capably finds the correct overall fold when the amount of distance knowledge specifies the correct topology with little ambiguity.

In this paper, we present a novel method that is useful in instances when very limited (sparse) structural information is available and the topology of the protein is far from uniquely specified. The method efficiently generates the complete set of topologies consistent with a set of inter-residue restraints, even when the number of restraints is very small. We will show that as few as N/24 inter-residue restraints reduce the number of topologies sufficiently so that a simple residue burial score can identify the native topology in a very small set of candidates (typically < 5). We expect that improved contact prediction approaches will be capable of obtaining reliable sparse restraint information (at a level of ~N/12-N/24) for a wide array of protein prediction targets.

Furthermore, for many protein sequences, knowledge of simple biochemical information such as disulfide bond connectivity provides enough information to successfully apply our prediction hierarchy. With this hierarchical approach, moderate resolution globular protein structure can be determined from sparse distance information in just a few hours on a single processor workstation.

METHODS

The Restrained Generic Protein (RGP) Direct MC method is an off-lattice residue buildup procedure for generating all polypeptide topologies that are consistent with a set of inter-residue distance restraints. The RGP method is the first step in the *generate-and-select* hierarchical structure prediction procedure shown in Figure 3.1. In the second step of the hierarchy, a static residue burial (S-RB) scoring function is used to select a small set of candidates from the RGP ensemble. In the third hierarchical step, an intact peptide backbone representation is constructed for each fold in the selected set (the RGP method produces an α–carbon trace of each conformation). Following the construction of the intact peptide backbone, each of the selected conformations is minimized with respect to the residue burial function used in step 2. This *dynamic* residue burial (D-RB) selection process further reduces the set of remaining fold candidates. The final stage of the prediction hierarchy uses predicted secondary structure information or additional distance restraints to further reduce and refine the surviving set from the previous step.

Figure 3.1 Flowchart diagram of the *generate-and-select* hierarchical method for predicting moderate resolution tertiary protein structure from sparse distance restraints.

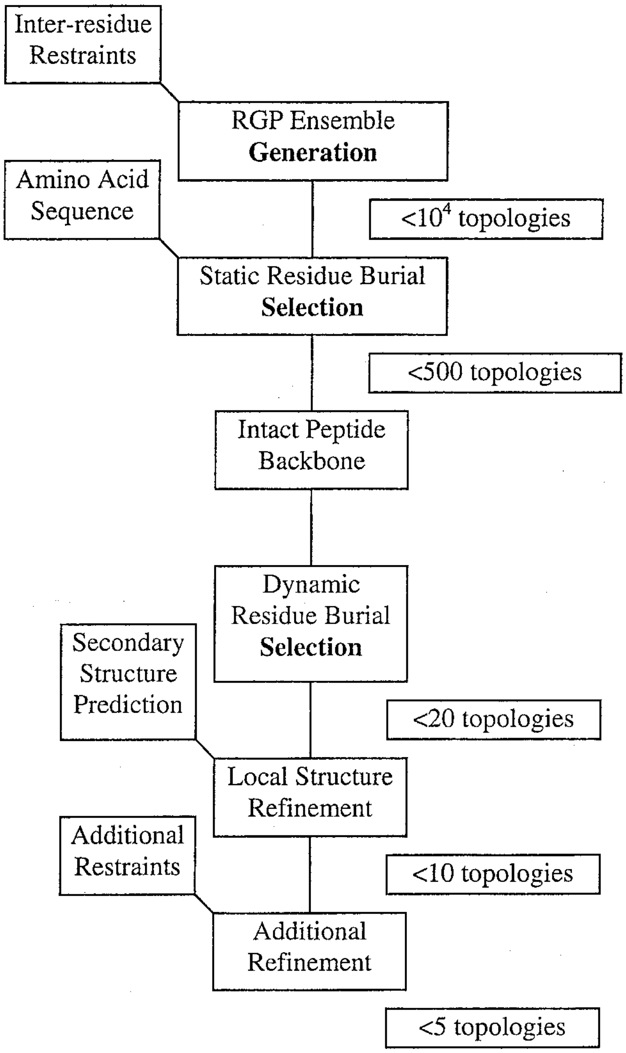
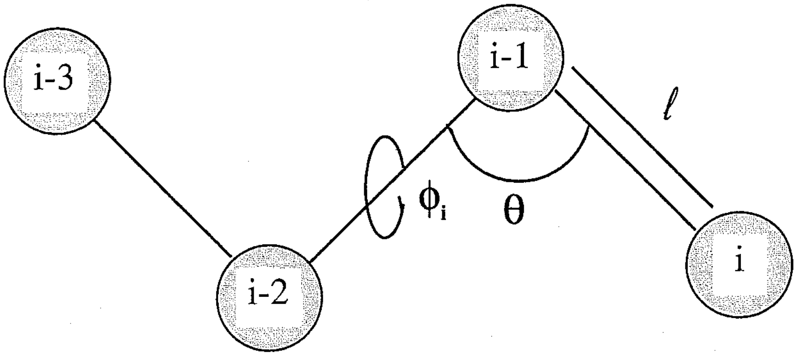

Inter-residue Restraints
RGP Ensemble **Generation**
Amino Acid Sequence
<10⁴ topologies
Static Residue Burial **Selection**
<500 topologies
Intact Peptide Backbone
Dynamic Residue Burial **Selection**
Secondary Structure Prediction
<20 topologies
Local Structure Refinement
Additional Restraints
<10 topologies
Additional Refinement
<5 topologies

Figure 3.2 Ball and stick peptide representation used in the RGP-DMC method. Each residue is connected to its neighboring residues by a fixed bond length, ℓ= 3.8Å, with fixed bond angle, θ =120°. The possible values of ϕ_i in an n-state per residue representation are $\phi_i = i\times(360°/n)$ for i=0,1,2,...n−1.


i-3
i-2
i-1
i
ϕ_i
θ
ℓ

A. Protein representation

The RGP method employs a "ball and stick" protein model.[12] Each residue is connected to its neighboring residues by a fixed bond length, ℓ= 3.8Å, with fixed bond angle, $\theta = 120°$. Thus, the coordinates of residue i are precisely determined from the coordinates of residues i−1, i−2, and i−3, given a single torsion, ϕ_i, about the central bond (Figure 3.2). The possible values of ϕ_i in an n-state per residue representation are $\phi_i = i\times(360°/n)$ for i=0,1,2…n−1. Hence, for a 6-state per residue representation, ϕ_i = 0°, 60°, 120°, 180°, 240°, or 300°.

B. Restraint implementation

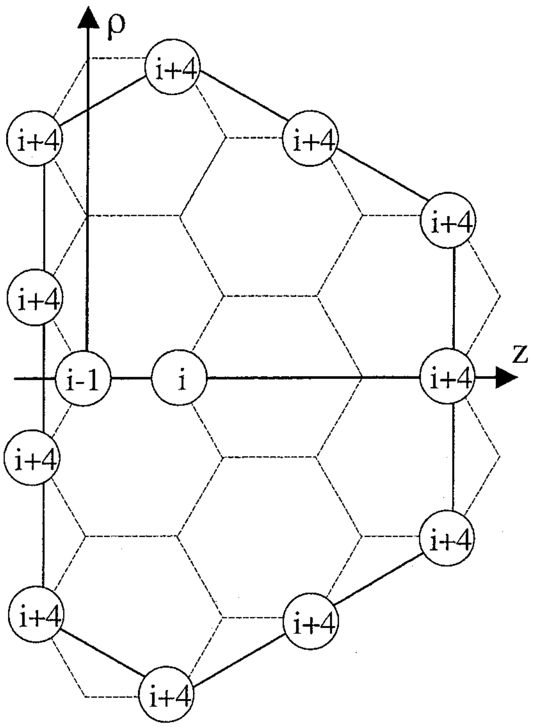
The RGP method is a residue buildup procedure. Residues are added one by one from the N- to C-terminus to construct a complete polypeptide. An efficient restraint technique assures that the polypeptide conformations are consistent with a set of user defined inter-residue distance restraints. Consider using a buildup procedure to construct a polypeptide where residue j and residue k are less than 6Å apart ($j < k$). The simplest approach is to randomly enumerate all possible conformations of residue j through residue k and discard the "dead end" conformations that do not satisfy this restraint.[8] Unfortunately, this approach becomes prohibitively expensive as the sequential distance between j and k increases, since the detection of a dead end occurs after the construction of residue k.

An algorithm that can determine if a conformation is a dead end prior to the addition of residue k yields a vast improvement in efficiency. The longest distance traversed by each residue addition step is a single bond length, ℓ= 3.8Å. Thus, it is impossible to place residue k within 6Å of residue j if residue i ($j < i < k$) is greater than

6+3.8(k-i) angstroms from residue j. Thus, it is possible to predict at step i if a conformation must eventually result in a dead end at step k.

The restraint method incorporated into the RGP method is slightly more complex in that it also considers the angle between residues j, i, and i-1. Figure 3.3 shows the possible positions for residue i+4 in our peptide model when ϕ_{i+2}, ϕ_{i+3}, and ϕ_{i+4}, = 0° or 180 °. Consider a cylindrical coordinate system where the z-axis travels through the bond between residue i–1 and residue i, and the z-axis origin is at residue i–1. The radial axis, ρ, represents the perpendicular distance to the z-axis. In the figure, the solid line around the perimeter traces the maximum radial distance that residue i+4 may be from the z-axis for a given value of z. Hence, this solid line represents the most extreme position in (z, ρ) space that residue (i-1+5) may be placed from residues i–1 and i in our polypeptide model.

Figure 3.3 The allowed positions of residue i+4 in relation to residue i-1 and residue i when residues i-1, i, i+1, i+2, i+3, and i+4 all lie in the same plane. For the cylindrical coordinate system (z, r), the maximum value of r for residue i+4 may be expressed as a function of z. This is used to derive equations (1)-(7).


ρ
z
i+4
i+4
i+4
i+4
i+4
i-1
i
i+4
i+4
i+4
i+4
i+4
i+4

Similar diagrams lead to a general expression for the maximum value of ρ, ρ_{max}, for an arbitrary residue (i-1)+n at a specific z coordinate. Defining

$$\rho_{peak} = (n\text{-}1)(\ell \sin 60°), \tag{1}$$

we find that

(a) If n is even, then z must lie between

$$\{z_{min}, z_{max}\} = \{(-3\ell/4)(n-4), (3\ell/4)(n)\}, \tag{2}$$

and two cases define ρ_{max}:

(a.1) for $z \geq 3\ell/2$,

$$\rho_{max} = \rho_{peak} - (\tan 30°)(z-(3\ell/2)), \tag{3}$$

(a.2) for $z < 3\ell/2$,

$$\rho_{max} = \rho_{peak} + (\tan 30°)(z-(3\ell/2)). \tag{4}$$

(b) If n is odd, then z must lie in the range

$$\{z_{min}, z_{max}\} = \{(-\ell/4)(2+3(n-5)), (\ell/4)(4+3(n-1))\}, \tag{5}$$

and two cases define ρ_{max}:

(b.1) for $z \geq \ell$

$$\rho_{max} = \rho_{peak} - (\tan 30°)(z-\ell), \tag{6}$$

(b.2) for $z < \ell$

$$\rho_{max} = \rho_{peak} + (\tan 30°)(z-\ell). \tag{7}$$

Thus, expressions (1)-(7) specify the greatest distance in (z, ρ) space that any residue (i-1+n) may be placed from residues i–1 and i.

Now we return to the example of constructing a polypeptide conformation with restrained residues j and k. Assume that the restraint limits the distance between j and k to a maximum of 6.58Å. This distance is equivalent to the maximum distance traveled in 2 residue addition steps, i.e., $2\ell(\sin\theta/2) = 6.58$Å. Thus placing residue k within 6.58Å of residue j is similar to requiring that residue j lies in allowed (z, ρ) space for residue k+2. Thus if a candidate torsion ϕ_i places residue i ($j < i < k$) in a location such that residue j lies outside allowed (z, ρ) space for $n= k+2-(i-1)$, the torsion will inevitably result in a dead end, and we can eliminate it.

In the above example, we assigned the distance restraint between residue j and residue k a *bond order* ($bo_{j,k}$), which represents the number of residue addition steps required to span the restraint distance. A single addition step of length ℓ spans 3.8Å; hence, $bo_{j,k}=1$ represents this distance. Two residue addition steps span distances up to $2\ell \times (\sin\theta/2) = 6.58$Å, hence for 3.8Å < d <6.58Å, $bo_{j,k}=2$.

The above discussion specifies how the RGP method satisfies a single inter-residue restraint. A single restraint between two residues is called a *first order* restraint. A first order restraint occurs when residues j and k are restrained, and we seek to add residue i ($j< i < k$) such that

$$i-j \geq k-i+bo_{j,k}-2. \qquad (8)$$

If two restraints (j, k) and (p, q) are specified where $j < k < p < q$, then the restraint on (j, k) is satisfied before residue p is added. Thus these restraints are separate. However, if

$p<k$, then when adding new residues i, where $p \leq i \leq k$, we must simultaneously consider both restraints. We refer to this as a *second order* restraint. Consider a polypeptide with a first order restraint between residues ($bo_{5,39}=2$) and a first order restraint between residues 17 and 39 ($bo_{17,39}=2$). Then there is effectively a second order restraint between residues 5 and 17, with $bo_{5,17} = bo_{5,39}+bo_{17,39}= 4$. Thus as we grow each residue i, such that $i-5 \geq 17-i+bo_{5,17}-2$ (i.e., residues 12, 13, 14, 15, 16, and 17), residue 5 must lie in allowed (z, ρ) space for $n = 17+bo_{5,17}-(i-1)$. Thus, depending on the configuration of the inter-residue restraints in the protein, there can be first and second order restraints that require attention at each growth step i.

C. Conformation sampling procedure

Now that we have described the restraint technique, it is possible to list the steps followed to construct a complete polypeptide by the RGP method.

1. The inputs required for the RGP method are the number of residues in the polypeptide (N), and a list of inter-residue distance restraints with restraint bond orders, $bo_{j,k}$. The first and second order restraints for each residue addition step i are determined.
2. A three-residue starting fragment corresponding to the first three residues in the polypeptide sequence is constructed. Residues are added one at a time in one of $p = 6$ different torsional states to construct the complete N-residue polypeptide. For each residue addition step, q, the restraint conditions are evaluated. If the candidate torsion does not satisfy the restraints, the probability of selecting this torsion is zero.

If a candidate torsion does satisfy the restraints, the probability of selecting this torsion is

$$P_q = \exp\left(-E_q/_{kT}\right) \Big/ \sum_{i=1}^{p} \exp\left(-E_i/_{kT}\right), \tag{9}$$

where p is the number of candidate torsions and E_q is the addition energy for a specific torsion candidate q, according to the CCB-DMC procedure.[13] The addition energy of a torsion candidate for residue i is given by the summation of the energy between residue i and each existing residue in the peptide fragment. For all residue types, the energy of a residue pair is taken as

$$E_{ij}(R) = E_0 \left[\left(\frac{R}{R_0} \right)^{12} - 2\left(\frac{R}{R_0} \right)^{6} \right], \tag{10}$$

where R_0 = 5.5Å, E_0 = 0.15 kcal/mol, R is the distance between the coordinates of each residue, and i and j are not nearest neighbors in the sequence. This sequence independent energy function accounts for the excluded volume of each residue.

3. At a given residue addition step i, if no candidate torsion satisfies the restraint conditions, the polypeptide is re-grown from residue i-4 in an attempt to satisfy this restraint. The current implementation allows one such *backtrack* before discarding the entire polypeptide and growing a new polypeptide from the starting fragment.

4. A *look-ahead* strategy may also be performed, where the placement of residue i+1 determines the probability of selecting the torsion angle for residue i. That is, for a particular torsion candidate for residue i, if there is no torsion candidate ϕ_{i+1} that

satisfies the restraints on residue i+1, the probability of selecting that particular torsion candidate for residue i is zero.

D. Static residue burial (S-RB) score

The RGP method generates the α–carbon coordinates for distance-restrained protein conformations without considering the identity of the amino acid sequence. In order to assign an energy to each of the RGP conformations, we developed a very simple, static residue burial (S-RB) score based on the observation that the α–carbon positions for the hydrophobic residues (Cys, Ile, Leu, Phe, and Val) lie closer to the protein center of mass than the hydrophilic residues (Arg, Asn, Asp, Gln, Glu, Lys, Pro, and Ser) (Figure 3.4). Once the RGP method generates a complete polypeptide, the center of mass is calculated from the α–carbon coordinates. The distance from each hydrophilic and hydrophobic residue to the center of mass is calculated and expressed as a factor of

$$R_g(N) = -1.26 + 2.79N^{1/3}, \tag{11}$$

where R_g represents the expected minimum radius of gyration for a globular protein of N residues.[14] Each hydrophobic and hydrophilic residue receives a residue burial score, W, that depends on its distance from the center of mass, $|R\text{-}R_{cm}|$.

For hydrophobic residues we take

$$W = -1 \text{ if } |R\text{-}R_{cm}| \leq D_{phob} \tag{12a}$$

$$W = 2 \text{ if } |R\text{-}R_{cm}| > D_{phob}$$

where

$$D_{phob} = 1.2R_g \text{ for Phe and Ile residues;} \tag{12b}$$

$$D_{phob} = 1.25R_g \text{ for Leu and Val residues; and}$$

$D_{phob} = 1.3R_g$ for Cys residues.

For hydrophilic residues we take

$$W = 2 \text{ if } |R\text{-}R_{cm}| \leq D_{phil} \tag{13a}$$

$$W = -1 \text{ if } |R\text{-}R_{cm}| > D_{phil}$$

where

$$D_{phil} = 0.85R_g \text{ for Asp residues;} \tag{13b}$$

$D_{phil} = 0.8R_g$ for Asn, Gln, Glu, Lys, Pro, and Ser residues; and

$D_{phil} = 0.75R_g$ for Arg residues.

The S-RB score for the polypeptide is the sum of the individual residue burial scores,

$$\text{S-RB} = \sum_{i=1}^{N} W_i \, . \tag{14}$$

E. Intact backbone construction

The RGP method generates the α–carbon trace of a polypeptide. Thus, an intact peptide backbone must be constructed for each structure in the selected set. Since the RGP structures are very low-resolution folds, it is not critical that the backbone preserves the original trace exactly. To this end, we find that an algorithm developed by Park and Levitt[15] works well for quickly producing an intact backbone highly similar to the original RGP trace. A six-state per residue backbone representation[16] generates a full atom backbone from the α–carbon coordinates that is typically less than 3Å CRMS from the original RGP conformation. A much better method is given by Milik et al.

Figure 3.4 Frequency histogram of the distance of hydrophobic (solid line) and hydrophilic (dotted line) residues to the center of mass for 61 non-homologous, single-domain proteins. The distance is normalized by the factor $R_g(N)$, the expected minimum radius of gyration for a globular protein structure of N residues (see equation (11) in text).

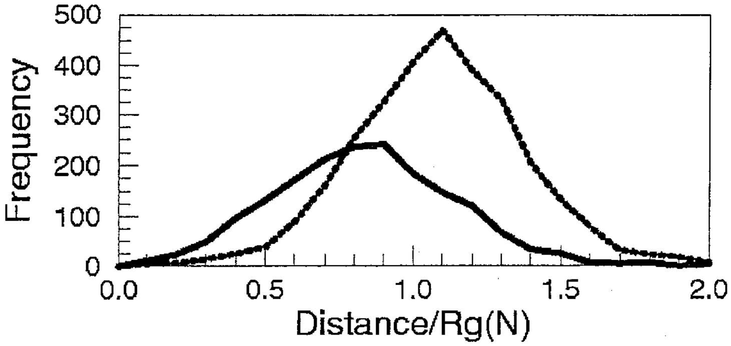
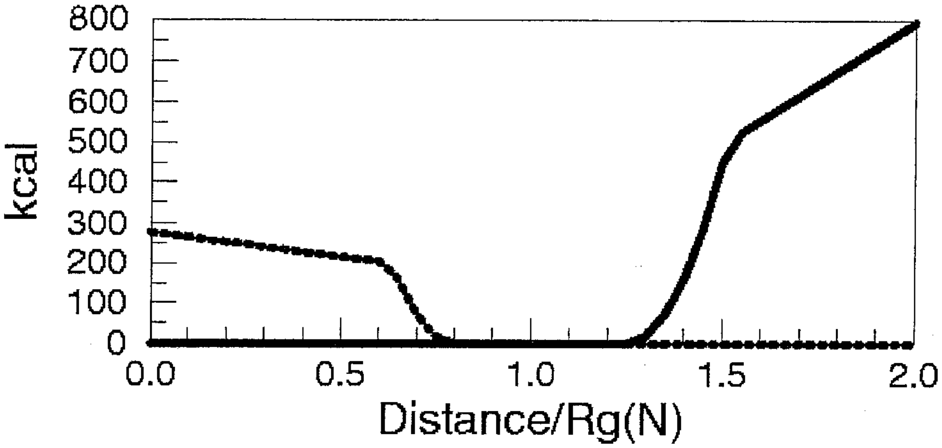

500
400
300
200
100
0
Frequency
0.0
0.5
1.0
1.5
2.0
Distance/Rg(N)

Figure 3.5 Burial bias potentials used for D-RB minimization. The solid line is the hydrophobic burial bias potential, $E_{b\text{-}phob}$, and the dotted line is the hydrophilic burial bias potential, $E_{b\text{-}phil}$ (R_g=12Å). Using these potentials, hydrophobic residues are drawn toward the protein interior, while hydrophilic residues are excluded from the protein core.


kcal
800
700
600
500
400
300
200
100
0
0.0
0.5
1.0
1.5
2.0
Distance/Rg(N)

F. Dynamic residue burial (D-RB) score

Once a peptide backbone has been constructed, each backbone is minimized with respect to the S-RB criteria using the simple burial bias potentials, $E_{b\text{-phob}}$ and $E_{b\text{-phil}}$, shown in Figure 3.5. Letting $x = |R-R_{cm}|$, these have the form

$$E_{b\text{-phob}}(x) = 0 \text{ kcal for } x \leq D_{phob}. \tag{15}$$

$$E_{b\text{-phob}}(x) = 50(x - D_{phob})^2 \text{ kcal/Å}^2 \text{ for } D_{phob} < x \leq D_{phob} + 3.16\text{Å}.$$

$$E_{b\text{-phob}}(x) = (500 + 50(x - D_{phob})/\text{Å}) \text{ kcal for } x > D_{phob} + 3.16\text{Å}.$$

$$E_{b\text{-phil}}(x) = (200 + 10(x - D_{phob})/\text{Å}) \text{ kcal for } x < D_{phil}\text{-}2\text{Å}. \tag{16}$$

$$E_{b\text{-phil}}(x) = 50(x - D_{phob})^2 \text{ kcal/Å}^2 \text{ for } D_{phil}\text{-}2\text{Å} \leq x \leq D_{phil}.$$

$$E_{b\text{-phil}}(x) = 0 \text{ kcal for } x > D_{phil}.$$

Distance restraint bias potentials are added to preserve the original inter-residue distance restraints between residue j and residue k. Letting $x = |R_j - R_k|$, these have the form

$$E_{j,k}(x) = (200 + 10(4\text{Å}-x)/\text{Å}) \text{ kcal for } x < 2\text{Å}.$$

$$(17)$$

$$E_{j,k}(x) = 50(4\text{Å}-x)^2 \text{kcal/Å}^2 \text{ for } 2\text{Å} \leq x < 4\text{Å}.$$

$$E_{j,k}(x) = 0 \text{ kcal for } 4\text{Å} \leq x \leq 7\text{Å}.$$

$$E_{j,k}(x) = 50(x-7\text{Å})^2 \text{kcal/Å}^2 \text{ for } 7\text{Å} < x \leq 7\text{Å} + 3.16\text{Å}.$$

$$E_{j,k}(x) = (500 + 50(x-7\text{Å})/\text{Å}) \text{ kcal for } x > 7\text{Å} + 3.16\text{Å}.$$

Once the hydrophobic, hydrophilic, and distance restraint potentials are in place, 500 steps of conjugate gradient minimization are performed, where standard force field terms represent the peptide backbone.[17] After minimization, the S-RB score is recalculated, yielding the D-RB score.

G. Additional restraints

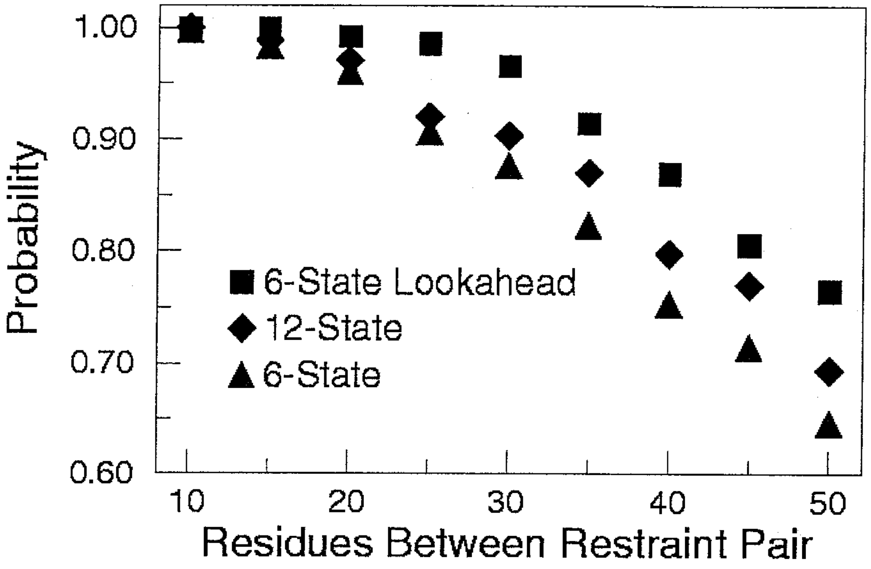
Additional inter-residue distance restraints can be incorporated into a structure by minimization using the restraint bias potential described in equation (17) between the newly restrained residues. Local structure can be refined to incorporate a secondary structure prediction using a restrained minimization procedure that refines the α–carbon coordinates in the original structure to comply with an optimally superimposed secondary structural unit.

COMPUTATIONAL PROTOCOL AND EFFICIENCY

In this paper, we consider the ability of the RGP algorithm to generate low-resolution tertiary folds given sparse inter-residue distance information. A list of appropriate inter-residue distance restraints was selected for each native protein by selecting pairs of residues known to be between 4Å and 7Å away in the native coordinate file. RGP treated each restraint pair (j, k) with $bo_{j,k}$=2, until the addition of residue k. At this addition step, the RGP algorithm required that residue k be placed anywhere from 3.8Å to 7.4Å from residue j, rather than calculating z and ρ_{max} according to equations (1)-(7). Thus the difference in the distance between the restrained residues in the generated structure and the native structure could be as large as 3.4Å.

Figure 3.6 shows the probability of satisfying a first order restraint pair ($bo_{j,k}$=2) separated by N residues using the RGP method. For a sequence separation of 50 residues, the RGP method generates conformations that satisfy the restraint with >60% efficiency using just six-states per residue without a look-ahead step. Without a restraint coupling, the probability of generating a 50-residue segment with the terminal residues between 3.8Å and 7.4Å apart is less than 0.005. Thus, by identifying dead ends at each step in the buildup procedure, RGP leads to an efficiency increase by a factor of 120 over a random or exhaustive approach. Using a Silicon Graphics R10000 processor, the RGP-DMC method can generate 1000 50-residue polypeptides with restrained termini in less than 3 minutes (six-state per residue representation).

Figure 3.6 The probability of satisfying an inter-residue restraint of bond order two plotted versus the sequence separation of the restrained residues using the RGP-DMC method. A six state per residue representation coupled with a look-ahead step is denoted by (■) markers, a twelve state per residue representation without a look-ahead step is denoted by (◆) markers, and a six state per residue representation without a look-ahead step is denoted by (▲) markers.


1.00
0.90
0.80
0.70
0.60
Probability
10
20
30
40
50
Residues Between Restraint Pair
■ 6-State Lookahead
◆ 12-State
▲ 6-State

RESULTS

We will demonstrate how the RGP method can be used for making successful low-resolution tertiary structure predictions by applying the *generate-and-select* prediction hierarchy to two sequences with known tertiary structure. The first prediction target is the LexA repressor DNA binding domain from *E. coli* (1lea),[18] a 72-residue protein with helical and beta strand secondary structure. The second target is sea hare myoglobin (1mba),[19] a 146-residue protein with eight helices.

A. LexA repressor DNA binding domain (72 AA)

Table 3.1 shows the results obtained by applying the *generate-and-select* prediction hierarchy to the LexA repressor sequence using 2 (N/36), 3 (N/24), 6 (N/12), and 12 (N/6) inter-residue distance restraints. For each restraint set (see Table 3.3), the RGP algorithm generated an ensemble of S structures (column 2 of Table 3.1) using a six-state per residue representation. The closest match to the experimental structure in this ensemble has a CRMS as given in column 3 of table 3.1. From this original ensemble, a small subset of s structures (column 4) was selected according to the S-RB score. We then optimized the structure using D-RB minimization. Comparing these structures with the experimental structure, we found a near-native match with the rank given in column 5 and the CRMS given in column 6. We then took each structure in the selected set s and refined it to incorporate the results of a PHD[6] secondary structure prediction. We again applied D-RB and ranked the structures. We found a near native match to the experimental structure with the rank given in column 7 and the CRMS given in column 8.

Table 3.1 shows that 3 (N/24) inter-residue restraints combined with a secondary structure prediction is sufficient to identify a low-resolution native conformation (6.1Å CRMS) in the top 3 of all conformations. Less than 30 minutes on a Silicon Graphics R10000 processor was required to generate and score the 5,000 structure ensemble, and less than two hours was required to refine and minimize the selected set, resulting in an overall prediction time of less than three hours.

Given 6 (N/12) inter-residue distance restraints, an RGP ensemble of only 500 structures resulted in a low-resolution structure that was 6.3Å CRMS from the native. Adding the predicted secondary structure resulted in a 4.5Å CRMS structure ranked in the top 10 of all candidates. Figure 3.7 shows this structure compared to the native.

Generating the 500-structure set with N/12 restraints required less than 10 minutes on our single processor workstation. Increasing the restraint density results in lower efficiency, and thus rather than generate structures with N/6 restraints using RGP, we began with the set of 271 lowest S-RB energy N/12 structures, and added the remaining six restraints during D-RB minimization. This led to just 44 structures that satisfied all 12 restraints. The second lowest D-RB energy structure had the same overall fold as the native (6.13Å CRMS). Including the secondary structure prediction and carrying out D-RB minimization resulted in a near-native fold (CRMS=5.76A) tied for the best D-RB score.

Table 3.1 Results of the *generate-and-select* prediction hierarchy for LexA repressor (72 residues) using different sets of inter-residue restraints. As an example, consider the case of 3 distance restraints (row N/24). The RGP method generated S = 5,000 structures. One of the structures in this set has a CRMS of 6.57Å from the native structure. Applying S-RB criteria to this set, we selected the s = 209 lowest energy conformations and performed a D-RB minimization. We found a near native match in the top 7 (CRMS = 6.76A). We then incorporated a PHD[6] secondary structure prediction into the 209 conformations and ranked each structure according to its D-RB score. We found a near native match in the top 3 structures (CRMS = 6.11Å). The total time for this process was 180 minutes on an SGI workstation.

	RGP Ensemble		**Selected Set**			**Sec. Prediction**	
	S[a]	CRMS[b]	s[c]	Rank[d]	CRMS[e]	Rank[f]	CRMS[g]
N/36	30,000	6.85Å	395	24t	7.46Å	14t	6.67Å
N/24	5,000	6.57Å	209	6t	6.76Å	2t	6.11Å
N/12	500	6.28Å	271	1	6.43Å	7t	4.45Å
N/6	-	-	44	2	6.13Å	1t	5.76Å

[a] S denotes the total number of structures generated in the RGP ensemble (for N/6 constraints, the RGP method was not used to generate a conformation ensemble; the 271 structures in the selected set for N/12 were used as a starting set).

[b] The lowest α–carbon CRMS structure in the RGP ensemble.

[c] Set of s structures selected from the original RGP ensemble according to the SRBS.

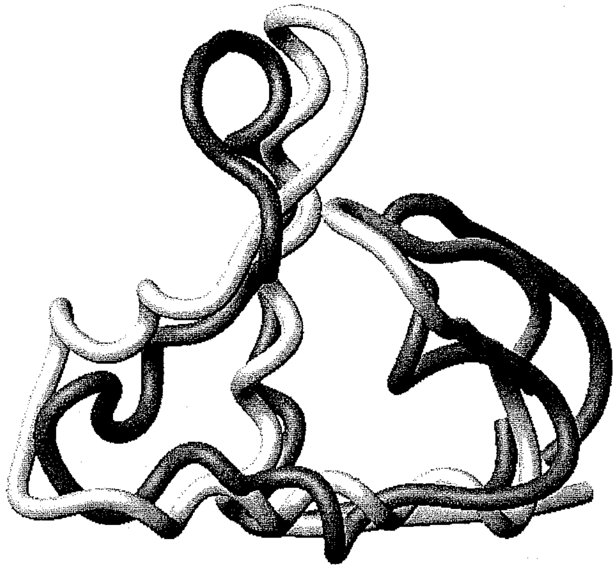
[d] Rank of the lowest energy structure possessing the native global fold using the DRB score (t denotes a tie).

[e] CRMS of ranked structure from column d.

[f] Rank of the lowest energy structure possessing the native global fold using the DRB score after incorporation of predicted sheet and helical regions from a PHD[6] secondary structure prediction.

[g] CRMS of ranked structure from column f.

Figure 3.7 The *generate-and-select* backbone prediction (4.5Å CRMS, dark strand) for 72-residue lexA repressor (light strand). Six inter-residue distance restraints were used in conjunction with predicted secondary structure to obtain this prediction. The complete *generate-and-select* hierarchy required less than 3 hours on a single processor R10000 Silicon Graphics workstation for this protein.



B. Myoglobin (146 AA)

The RGP method was also successful when applied to a much longer sequence using 12 (~N/12) restraints (Table 3.2). After an ensemble of S =50,000 structures was generated using the RGP algorithm, a smaller set of s = 117 conformations was selected based on S-RB score (the non-minimized residue burial score). Applying D-RB to this set led to a near-native match as number 11. Incorporating the secondary structure predicted by PHD[6] into each conformation and applying D-RB resulted in a near-native structure ranked fifth (7.01Å CRMS).

Starting with the 117 structures with 12 restraints, we added 12 more (N/6 total) during D-RB minimization. This led to 23 structures that satisfied all 24 restraints. The lowest energy fold in this set possessed the native overall fold topology. Incorporating the secondary structure prediction led to a highest-ranking structure with the correct overall fold (6.30Å CRMS). Figure 3.8 compares this structure to native myoglobin.

Table 3.2 Results of the *generate-and-select* prediction hierarchy for myoglobin (146 residues) using N/12 and N/6 inter-residue restraints. The definition of each column is similar to Table 3.1.

	RGP Ensemble		**Selected Set**			**Sec. Prediction**	
	S	CRMS	s	Rank	CRMS	Rank	CRMS
N/12	50,000	8.95Å	117	11	8.77Å	5	7.01Å
N/6	-	-	23	1	9.28Å	1	6.30Å

Figure 3.8 The *generate-and-select* backbone prediction (6.3Å CRMS, dark strand) for 146-residue myoglobin (light strand). Twenty-four inter-residue distance restraints were used in conjunction with predicted secondary structure to obtain this prediction. The complete *generate-and-select* hierarchy required less than 12 hours on a single processor R10000 Silicon Graphics workstation for this protein.

Table 3.3 List of inter-residue restraints used for the LexA repressor and myoglobin structure predictions. Each restraint set contained the restraints listed in the corresponding row along with the restraints listed in each prior row.

(a) LexA (N = 72 residues)			
N/36 (2)	1-72	25-64	
N/24 (3)	11-31		
N/12 (6)	8-50	28-44	55-68
N/6 (12)	2-53	11-47	18-26
	31-43	51-58	58-65
(b) Myoglobin (N = 146 residues)			
N/12 (12)	1-84	6-129	10-75
	16-119	22-65	30-51
	46-54	88-137	93-144
	102-141	109-131	113-127
N/6 (24)	3-79	9-125	10-130
	13-123	17-116	26-60
	41-48	78-84	101-146
	105-138	117-122	141-146

DISCUSSION

Both distance-geometry and dynamic MC methods produce correct low-resolution structure predictions given ~N/6 inter-residue restraints and accurate secondary structure assignments. We have demonstrated that a direct MC approach obtains predictions of similar precision with very little computational effort. Furthermore, since the RGP method employs a very simple packing force field, it can provide a coarse sampling of conformation space many orders of magnitude faster than more detailed dynamic simulation methods. Consequently, RGP is computationally feasible even when restraint information is very sparse.

Levitt and coworkers[20, 21] have considered the ability of many different types of functions to "recognize" correct low-resolution (near-native) folds from large sets of incorrect decoys. In order to measure the success of a particular function, they devised a quality factor,

$$Q = \log_{10}\left(M/_{nr}\right) \tag{18}$$

where M is the total number of structures in the set, r is the highest rank of a near-native structure, and n is the number of near-native structures in the set. While many selection functions recognize native crystal structures with Q-scores greater than 4, near-native structures (<4Å CRMS by their definition) are far more difficult to recognize, with Q rarely exceeding 2. Thus, selecting a near-native structure from a set of greater than 10^5 decoys is not feasible with current recognition potentials.

To successfully recognize near-native folds, a selection function must possess two important properties. First, it should rank the native structure as one of the lowest energy conformations. Second, the selection function should be insensitive to small structural changes, so that near-native structures appear similar in energy to the native. Unfortunately, selection functions that unambiguously identify the native typically do so at the cost of being highly sensitive to small changes in structure.

Though our S-RB selection function is very simple, it too is sensitive to small structural changes. By performing a short conformational minimization (D-RB) we greatly increase the structural invariance of the S-RB score, since each conformation is evaluated at a local minimum of the selection function.

Given just 3 restraints for the 72-residue 1lea, an ensemble of ~10^4 structures contains several native topology conformations. Thus, preserving the native topology in a smaller set of ~10^2 structures requires Q~1.5, a level of recognition attained by our S-RB function, and possibly many previously developed recognition approaches. Most importantly, since this reduced set of structures is very manageable in size, it is computationally feasible to use a dynamic selection procedure to select an even smaller set that will still contain the native topology. In this manner, the most promising topologies are analyzed by a selection procedure that possesses both properties required for successful near-native structure recognition.

CONCLUSION

We have developed a direct Monte Carlo method for efficiently generating the complete set of protein topologies consistent with a set of inter-residue distance restraints. We find that fewer than 10^4 distinct topologies are consistent with having 3 uniformly distributed restraints for a 72-residue protein. The RGP method can sample all of these topologies in less than one hour using a single Silicon Graphics R10000 processor workstation. Using the simple S-RB and D-RB criteria, it is possible to preserve a small set of structures that contains native topology. Since this remaining set is typically <10 structures, we suggest it is computationally feasible to perform much more detailed structure analysis to uniquely determine the native topology.

Future work will apply the *generate-and-select* hierarchy to the *de novo* prediction problem using predicted inter-residue contacts and biochemical structural restraints (such as disulfide bridges) as starting restraints for the RGP algorithm. A distinct benefit of the RGP method over previously developed methods is that only a very small number of restraints (<N/12) are needed to generate the native topology. For many protein sequences, knowledge of disulfide bond connectivity may be sufficient to lead to a correct low-resolution structure prediction. Furthermore, because the set of restraints is small compared to other methods, only a few of the most reliably predicted restraints[2] must be used, greatly minimizing the chance of including incorrect constraints in the predictions. Even so, it will be critical to understand how well the RGP method performs when some of the supplied tertiary restraints are inaccurate. Our results suggest that there is a sizable margin for error, since the present work allowed a 7.4Å distance between restrained residues.

REFERENCES

[1] Bystroff, C.; Baker, D.; *J. Mol. Biol.* **1998**, 281, 565.

[2] Goebel, U.; Sander, C.; Schneider, R.; Valencia, A.; *Proteins* **1994**, 18, 309.

[3] Ortiz, A. R.; Kolinski, A.; Skolnick, J.; *J. Mol. Biol.* **1998**, 277, 419.

[4] Benner, S. A.; Cannarozzi, G.; Gerloff, D.; Turcotte, M.; Chelvanayagam, G.; *Chem. Rev.* **1997**, 97, 2725.

[5] Kneller, D. G.; Cohen, F. E.; Langridge, R.; *J. Mol. Biol.* **1990**, 214, 171.

[6] Rost B.; Sander, C.; *J. Mol. Biol.* **1993**, 232, 584.

[7] Kolinski, A.; Skolnick, J.; Godzick, A.; Hu, W. P.; *Proteins* **1997**, 27, 290.

[8] DeWitte, R. S.; Michnick, S. W.; Shakhnovich, E. I.; *Prot. Sci.* **1995**, 4, 1780.

[9] Aszódi, A.; Gradwell, M. J.; Taylor, W. R.; *J. Mol. Biol.* **1995**, 251, 308.

[10] Skolnick, J.; Kolinski, A.; Ortiz, A. R.; *J. Mol. Biol.* **1997**, 265, 217.

[11] Ortiz, A. R.; Kolinski, A.; Skolnick, J.; *Proc. Natl. Acad. Sci. USA* **1998**, 95, 1020.

[12] Levitt, M.; *J. Mol. Biol.* **1976**, 104, 59.

[13] Sadanobu, J.; Goddard III, W. A.; *J. Chem. Phys.* **1997**, 106, 6722.

[14] Maiorov, V. N.; Crippen, G.M.; *Prot. Struct. Func Gen.* **1995**, 22, 273.

[15] Park, B.; Levitt, M.; *J. Mol. Biol.* **1995**, 249, 493.

[16] Rooman, M. J.; Wodak, S. J.; *Biochemistry* **1992**, 31, 10239.

[17] Mayo, S. L.; Olafson, B. D.; Goddard III, W. A.; *J. Phys. Chem.* **1990**, 94, 8897.

[18] Fogh, R. H.; Ottleben, G.; Ruterjans, H.; Schnarr, M.; Boelens, R.; Kaptein, R.; *EMBO J.* **1994**, 13, 3936.

[19] Bolognesi, M.; Onesti, S.; Gatti, G.; Coda, A.; Ascenzi, P.; Brunori, M.; *J. Mol. Biol.* **1989**, 205, 529.

[20] Park, B.; Levitt, M.; *J. Mol. Biol.* **1996**, 258, 367.

[21] Park, B. H.; Huang, E. S.; Levitt, M.; *J. Mol. Biol.* **1997**, 266, 831.

CHAPTER 4

FIRST PRINCIPLES PREDICTION OF PROTEIN FOLDING RATES

ABSTRACT

Experimental studies have demonstrated that many small, single-domain proteins fold via simple two-state kinetics. We present a first principles approach for predicting these experimentally determined folding rates. Our approach is based on a nucleation-condensation folding mechanism, where the rate-limiting step is a random, diffusive search for the native tertiary topology. To estimate the rates of folding for various proteins via this mechanism, we first determine the probability of randomly sampling a conformation with the native fold topology. Next, we convert these probabilities into folding rates by estimating the rate that a protein samples different topologies during diffusive folding. This topology-sampling rate is calculated using the Einstein diffusion equation in conjunction with an experimentally determined intra-protein diffusion constant. We have applied our prediction method to the 21 topologically distinct small proteins for which two-state rate data is available. For the 18 beta-sheet and mixed alpha-beta native proteins, we predict folding rates within an average factor of 4, even though the experimental rates vary by a factor of ~4×10^4. Interestingly, the experimental folding rates for the three four-helix bundle proteins are significantly underestimated by this approach, suggesting that proteins with significant helical content may fold by a faster, alternative mechanism. This method can be applied to any protein for which the structure is known and hence can be used to predict the folding rates of many proteins prior to experiment.

INTRODUCTION

One of the most important challenges in biology is to understand the relationship between the folded structure of a protein and its primary amino acid sequence. Consequently, there has been great interest in understanding how proteins fold. An important advance in 1991 was the experimental demonstration that stable intermediates were not present in the fast folding of Chymotrypsin Inhibitor 2 (Jackson & Fersht, 1991). Since then, two-state folding rates for 20 more small (<120 residues), topologically distinct proteins have been determined, providing sufficient rate data to begin testing quantitative aspects of proposed folding mechanisms (Jackson, 1998). Recently, Plaxco et al. reported a statistically significant correlation between the natural log of the two-state folding rate, $\ln(k_f)$, and a measure of the native state topological complexity (contact order) (Plaxco *et al.*, 1998b). This empirical observation suggests that the chemistry underlying the folding of simple, single-domain proteins may be universal, implying that a single mechanistic model might quantitatively account for the observed folding rates.

We recently proposed the Topomer-Sampling Model (TSM) of protein folding, wherein proteins fold by a two-state mechanism consisting of (Debe *et al.*, 1999a):

(i) topomer diffusion: random, diffusive sampling to find the native topomer (topomers are tubes of topologically equivalent conformations), followed by

(ii) intra-topomer ordering: non-random, local conformational changes within the native topology to find the unique native state.

Assuming step (i) represents the rate-limiting step in folding, the TSM folding rate is given by

$$k_f = \mathrm{P(Ntop)} \times k_{top}, \tag{1}$$

where P(Ntop) is the probability of randomly sampling a structure with the native tertiary topology (i.e., a structure in the native topomer), and k_{top} is the rate at which a protein samples different tertiary topologies as it folds [see (Debe et al., 1999a) for a precise definition of native tertiary topology in this context]. Previously, we developed a method to determine the total number of topomers, S_N, for a protein of length N, allowing us to estimate P(Ntop) = $(S_N)^{-1}$. Using this value of P(Ntop) in Eq. 1, we estimated that the rate for the topomer diffusion step (i) is ~10sec^{-1} for a 100-residue protein. This was an encouraging result, since this calculated rate is similar to experimentally observed folding rates. However, this calculation assumed that all topologies for a protein of length N have an equal probability of being sampled, and thus the predicted folding rate did not depend on the structure of the native fold. We now propose a quantitative first principles approach for predicting folding rates of specific proteins, where the probability of sampling the native topology is explicitly calculated from the native protein structure. This approach accurately predicts the folding rates of beta-sheet and mixed alpha-beta proteins.

METHODS AND RESULTS

We estimate P(Ntop) for a specific, native protein structure as follows. Consider choosing μ contacts (residue pairs whose alpha-carbons are within ~8Å) uniformly distributed throughout the protein structure. Given these μ contacts, we may write

$$P(\mathrm{Ntop}) = P(\mu) \times P(\mathrm{Ntop} \mid \mu), \tag{2}$$

where $P(\mu)$ is the probability of forming the μ contacts, and $P(\mathrm{Ntop} \mid \mu)$ is the conditional probability of sampling the native topology when the μ contacts are satisfied. We will focus on solving the terms in Eq. 2 to determine P(Ntop) for a native protein structure.

The determination of $P(\mu)$ is not trivial, since the probability of forming various inter-residue contacts in a protein depends on the location of these contacts along the protein sequence. For contact pairs that overlap in the sequence, the probability of forming one contact pair is influenced by the presence of the other contact, and thus the correlation between contact pairs must be considered while determining $P(\mu)$. Flory determined $P(\mu)$ for an average polymer of length N with μ arbitrary cross-links (the mean field approximation) (Flory, 1956). This mean field result has also been obtained using replica calculations (Gutin & Shakhnovich, 1994). Less progress has been made determining $P(\mu)$ for a particular set of μ contacts (no mean field approximation). Chan and Dill have determined correlation functions for up to three overlapping contacts with a non-arbitrary sequence separation using a cubic lattice polymer representation (Chan & Dill, 1990). However, $P(\mu)$ has not been determined for specific protein or polymer contact configurations for $\mu > 3$.

We determine P(μ) as follows. Let P(1) be the probability of sampling a conformation that satisfies one of the μ specified contacts. Then P(μ) can be written as

$$P(\mu) = P(\mu \mid 1) \times P(1), \tag{3}$$

where P(μ | 1) is the conditional probability of sampling a conformation that satisfies all μ contacts when one of the contacts is already satisfied. Similarly, the first term in Eq. 3 can be written as

$$P(\mu \mid 1) = P(\mu \mid 2) \times P(2 \mid 1). \tag{4}$$

By recursion, it follows that

$$P(\mu) = P(\mu \mid \mu{-}1) \times P(\mu{-}1 \mid \mu{-}2) \times \ldots \times P(2 \mid 1) \times P(1). \tag{5}$$

The individual probability terms in Eq. 5 can be solved using the Restrained Generic Protein (RGP) Direct Monte Carlo Method (Debe *et al.*, 1999b). The RGP method is a fast computational procedure for generating large ensembles of self-avoiding, off-lattice [ball-and-stick (Levitt, 1976)] protein conformations that comply with a set of user-defined inter-residue distance restraints. The term P(1) in Eq. 5 is given by the probability of satisfying one (or more) of the μ contacts in an unrestrained RGP ensemble of protein conformations. The next term, P(2 | 1), is given by the probability of satisfying two (or more) of the μ contacts in an ensemble of conformations that already comply with the inter-residue contacts that were satisfied during the determination of P(1). Hence, the i (or more) contacts satisfied in the conformations during the determination of P(i | i-1) are saved and used as inter-residue restraints to be satisfied by the conformations generated during the determination of P(i+1 | i). Each term in Eq. 5 can be determined by this approach, yielding P(μ).

Once P(μ) is determined, the remaining term required to solve for P(Ntop) by Eq. 2 is P(Ntop | μ), the probability of sampling the native topology when the μ contacts are satisfied. P(Ntop | μ) is determined by generating RGP conformations that satisfy all μ contacts and using the Native Topomer Test Procedure (Debe *et al.*, 1999a) to determine the fraction of the conformations possessing the native tertiary topology.

Thus, P(Ntop) can be calculated according to Eq. 2. The predicted rate of folding is computed using Eq. 1, where k_{top} is the rate a protein samples different topologies as it folds. We take this as the inverse of the time, τ_{top}, to diffuse from one topology to the next, approximated by the Einstein diffusion equation (Einstein, 1905):

$$k_{top} = \left(\tau_{top}\right)^{-1} = \left(\bar{x}^2 \middle/ 6D\right)^{-1}, \tag{6}$$

where $\bar{x}$ is the average CRMS distance between two neighboring topologies for a protein of length N as from our previous calculations [$\bar{x}$ = a + b(N)$^{1/3}$, where a = –4.12; b = 2.61; Debe et al., 1999a], and D ≈ 5×10^{-7}cm^2/s is an experimentally determined intra-protein diffusion constant (Hagen *et al.*, 1996). While the Einstein diffusion equation is certainly an approximation for a finite, constrained system such as a protein, the proportionality $\tau \propto \bar{x}^2$ has been shown to hold for proteins using molecular dynamics [see for example, trajectories 11 and 16 in Figure 2a of Lazaridis & Karplus, 1997]. Furthermore, since $\left(\bar{x}_{N=100} / \bar{x}_{N=50}\right)^2 \approx 2$, the variation in predicted folding rates is dominated by the

P(Ntop) term in Eq. 1, not by k_{top}. Hence, we expect that deviations from ideal diffusive behavior and chain length scaling errors do not significantly affect the rate predictions.

We refer to the overall procedure outlined above as the Native Topology Probability (NTP) method. Table 4.1 shows the predicted and experimental rate data for the 21 small proteins whose two-state folding rates have been determined. Figure 4.1 compares the experimentally determined $\ln(k_f)$ versus the predicted $\ln(k_f)$ for the set of 18 topologically distinct, beta-sheet and mixed alpha-beta proteins. The linear fit has a significant correlation (R=0.78), corresponding to an average prediction error of $e^{1.3} \approx 3.7$. This is similar to the error in rate, $e^{1.0} \approx 2.7$ arising from sequence changes in five structurally homologous protein families for which there is sufficient rate data (vertical error bars). Thus, the NTP method accurately predicts the general, sequence independent rate for alpha-beta and beta protein folding.

Figure 4.1 Experimental folding rate versus predicted folding rate for all 18 alpha-beta (●) and beta (▲) proteins for which there is rate data (Table 4.1). Over this data set the average error in the predicted rate is $e^{1.3} \approx 3.7$ ($R = 0.78$; $R=0.87$ for the fit excluding the outlier U1A/S6). The vertical error bars show the average error due to sequence changes across five structurally homologous families (average error is $e^{1.0} \approx 2.7$). The horizontal error bar represents the average error in the NTP rate predictions, $e^{0.5} \approx 1.6$. The predicted folding rates were calculated from Eq. 1. P(Ntop) in Eq. 1 was determined using Eq. 2. The term $P(\mu)$ in Eq. 2 was determined from the individual terms of the form $P(i+1 \mid i)$ in Eq. 5. These terms were determined from ensembles of protein conformations generated by the RGP method (see text). The radius of gyration (Rg) of the RGP conformations used in the determination of each $P(i+1 \mid i)$ term was limited to $Rg_{min} < Rg < 2Rg_{min}$, where $Rg_{min} = -1.26+2.79(N)^{1/3}$ (Maiorov & Crippen, 1995). This ensured that overly compact and non-compact conformations would not be considered. Inter-residue contacts were considered satisfied if their alpha-carbons were within 9.5Å. Two hundred conformations were generated for each $P(i+1 \mid i)$ determination. $P(i+1 \mid i)$ was typically ~0.2, yielding ~40 conformations out of the 200 that satisfied i+1 or more contacts. The ~40 sets of contacts satisfied during the determination of $P(i+1 \mid i)$ were saved and used during the determination of $P(i+2 \mid i+1)$. Two hundred new conformations were grown to determine $P(i+2 \mid i+1)$, so that on average, each of the ~40 different constraint sets was used to grow 5 of the new conformations (the algorithm cycles through the restraint sets in the order they were originally generated). Note that during the determination of $P(i+1 \mid i)$, more than i+1 contacts can be satisfied, for example i+3. In this case, all i+3 contacts are saved and used as restraints in the determination of the next probabilities, so that i+2

and i+3 contacts are necessarily satisfied when this contact set is used in the determination of P(i+2 | i+1) and P(i+3 | i+2), respectively. Note that the contact distance of 9.5Å is the only adjustable parameter in our model and is the same for all of the proteins considered. A distance of 9.5Å was chosen so that the calculated P(Ntop) values result in ln(k_f) predictions of the appropriate magnitude.

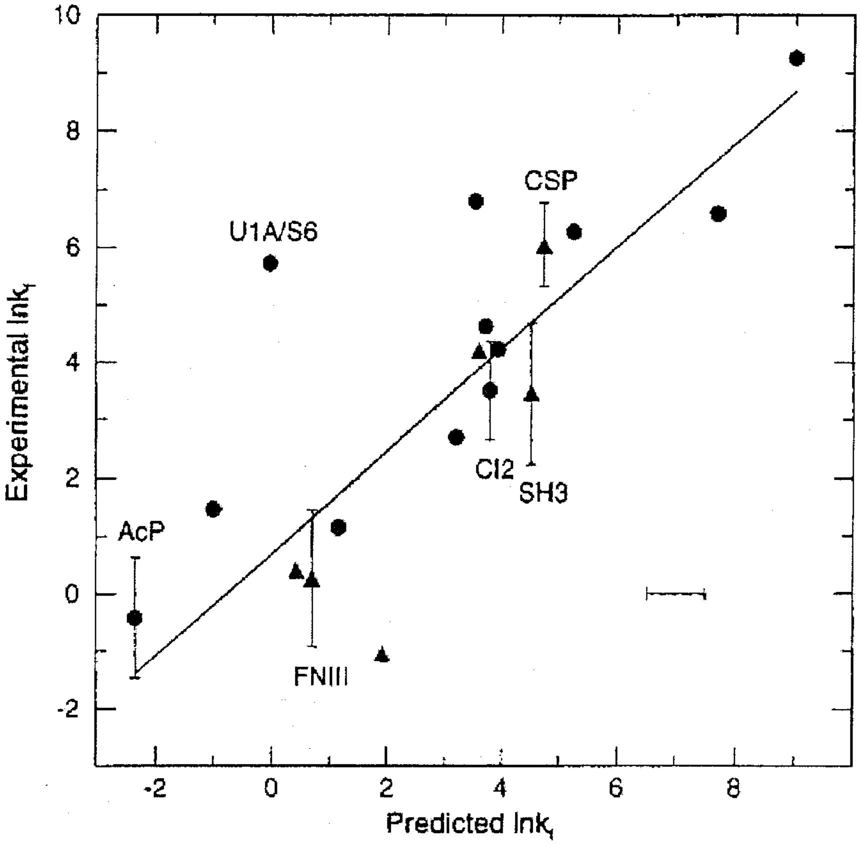

10
8
6
4
2
0
-2
-2
0
2
4
6
8
Experimental lnk$_f$
Predicted lnk$_f$
U1A/S6
CSP
CI2
SH3
AcP
FNIII

Table 4.1 Predicted and experimentally determined kinetic data for the 21 small, single domain, topologically distinct proteins (and protein families) that have been characterized. The predicted folding rates have an average uncertainty of $\sim e^{0.5} \approx 1.6$, based on calculations using at least two different sets of contacts to determine $P(\mu)$ and $P(\mathrm{Ntop} \mid \mu)$ for each protein (the contacts sets used for each protein are given in Tables 4.2-1 through 4.2-21). Column 7 lists the number and type of experiments that have been done on structurally homologous proteins of different sequence (M denotes point mutation experiments; H denotes homologous sequence experiments; T denotes experiments on proteins that are structural, but not sequence homologues). Such experiments estimate the extent that sequence specific effects influence the folding rate.

Protein	PDB		Pred	Pred	#	Exp.	Exp.
	Code	Length	$\ln k_f$	σ	Exp.	$\ln k_f$	σ
α proteins							
Monomeric λ–repressor[a]	1LMB	80	3.67	0.52	3M	10.23	1.26
ACBP[b]	2ABD	86	2.84	1.59	4H	6.62	1.04
Im9[c]	1IMQ	86	4.47	0.74	1	7.31	-
β proteins							
Tendamistat[d]	2AIT	74	3.60	0.13	1	4.20	-
Cold Shock Protein (A and B)[e]	1CSP 1MJC	67 69	4.73	0.83	4H 3M	6.04	0.72
SH3 domain (α-spectrin, src, fyn)[f]	1TUD 1NYF	64 67	4.51	0.24	3H	3.45	1.22
SH3 domain (PI3 kinase)[g]	1PKS	84	1.92	0.06	1	-1.05	-
Twitchin[h]	1WIT	93	0.43	0.98	1	0.41	-
^{9}FN-III & Tenascin (short and long)[i]	1FNF 1TEN	90 92	0.70	0.67	2H	0.26	1.18
α/β proteins							
Protein G B1 domain[j]	2GB1	56	5.23	0.26	2M	6.26	0.60
CI2[k]	1COA	64	3.80	0.58	90M	3.51	0.86
ADAh2[l]	1PBA	80	3.54	0.46	1	6.80	-
Arc repressor[m]	1ARR	53	9.01	0.51	1	9.27	-
Ubiquitin (Val26→Ala)[n]	1UBQ	76	3.71	0.27	1	4.63	-
IgG binding domain of protein L[o]	2PTL	62	3.94	0.47	10M	4.22	0.64
Hpr[p]	1HDN	85	3.20	1.06	1	2.70	-
FKBP12[q]	1FKB	107	-1.00	0.45	1	1.46	-
AcP (Muscle and Common Type)[r]	1APS 2ACY	98	-2.35	0.16	2H	-0.42	1.05
C-terminal Domain of protein L9[s]	1DIV	93	1.15	1.00	1	1.15	-
N-terminal Domain of protein L9[t]	1DIV	56	7.69	0.24	1	6.58	-
Spliceosomal U1A and Ribosomal S6[u]	1URN 1RIS	102	-0.04	0.56	2T	5.73	0.09

[a] Huang & Oas (1995), Burton *et al.* (1996), Ghaemmaghami *et al.* (1998).

[b] Kragelund *et al.* (1995), Kragelund *et al.* (1996).

[c] Ferguson *et al.* (1999).

[d] Schonbrunner *et al.* (1997).

[e] Schindler *et al.* (1995), Perl *et al.* (1998), Reid *et al.* (1998).

[f] Viguera *et al.* (1994), Viguera *et al.* (1996), Grantcharova & Baker (1997), Plaxco et al. (1998a).

[g] Guijarro *et al.* (1998).

[h] S. J. Hamill, unpublished observations. Data from Jackson (1998).

[i] Plaxco *et al.* (1997), Clarke *et al.* (1997), Hamill *et al.* (1998).

[j] Smith *et al.* (1996).

[k] Jackson & Fersht (1991), Itzhaki et al. (1995), Ladurner *et al.* (1998).

[l] Villegas *et al.* (1995).

[m] Robinson & Sauer (1996).

[n] Khorasanizadeh *et al.* (1996).

[o] Scalley *et al.* (1997).
[p] Van Nuland *et al.* (1998).
[q] S. E. Jackson, unpublished observations. Data from Jackson (1998).
[r] van Nuland *et al.* (1998), Taddei *et al.* (1999).
[s] Sato *et al.* (1999).
[t] Kuhlman *et al.* (1998).
[u] Silow & Oliveberg, (1997), Otzen et al. (1999).

The NTP procedure considerably underestimates the folding kinetics for the three α-helical bundle proteins (Table 4.1), where $\ln(k_{exp})$ exceeds $\ln(k_{pred})$ by $e^{6.6}\approx735$, $e^{3.8}\approx45$, and $e^{2.8}\approx16$. We expect that sequence-local helical conformational biases probably cause these helical proteins to fold faster via a diffusion collision mechanism (Karplus and Weaver, 1976). Future work will focus on understanding the amount of early helix formation required to produce the observed helical bundle folding rates.

The NTP rate predictions are based on the mechanistic assumption that the rate-limiting step in protein folding is a random, diffusive search for the native topology (Debe et al., 1999a). Thus we assume a nucleation-condensation mechanism (Fersht, 1995), where the transition-state required for condensation to the native state is the set of structures having the same tertiary topology as the native state. This is similar to the transition-state picture developed by the Fersht group from CI2 protein engineering studies (Itzhaki *et al.*, 1995). The accuracy of our first principles predicted rates provides evidence in favor of this nucleation-condensation mechanism. Furthermore, our calculations demonstrate how a nucleation-condensation mechanism accounts for the inverse relationship between folding rate and solvent viscosity recently observed for three small, two-state folding proteins (Jacob *et al.*, 1997; Plaxco & Baker, 1998; Bhattacharyya & Sosnick, 1999). This relationship is directly implied in our model by

Eq. 6 given the inverse proportionality between the diffusion constant and solvent viscosity in Stokes' Law.

We did not include any information about the stability of the native fold to accurately predict folding rates. Thus, we do not expect that stability is a primary factor in determining the folding rate (Plaxco et al., 1998b). However, a correlation could exist between native stability and folding rate for structurally homologous protein families (Plaxco *et al.*, 1998a). This correlation could arise from stabilizing sequence changes that increase the probability of trapping the protein once it has diffused into its native topomer.

The NTP predictions provide a useful framework for understanding factors that can change folding kinetics. The predicted folding rate is equivalent to the rate of randomly sampling a conformation in the transition-state ensemble. Thus, the folding rate is directly related to the difficulty of finding a conformation that can quickly condense to the native state. As the transition-state becomes more native-like (referred to as a tight transition state), the difficulty of finding a structure in the transition-state increases, and the folding rate decreases unless there is a mechanism to aid the search for a transition-state structure. Similarly, as the transition-state becomes less native-like (a loose transition state), the folding rate is expected to increase. Folding kinetics can also be affected by a change in the number of topologies that are available to the folding peptide. Faster folding rates might arise from sequence specific conformation biases, such as helical formation, which could preclude a protein from sampling a significant fraction of its possible topologies.

Careful examination of Figure 4.1 reveals that the NTP procedure accurately predicts the folding rates of all reported two-state beta-sheet and mixed alpha-beta proteins, with only one exception. The exception is spliceosomal protein U1A (and its recently characterized structural homologue ribosomal protein S6), which folds some $e^{5.8} \approx 330$ times more rapidly than predicted by the NTP method. U1A and S6 are quite long for two-state folding proteins (102 and 101 residues, respectively) and exhibit exceptional folding behavior, in that the $\log(k_{unfold})$ versus denaturant concentration and $\log(k_{fold})$ versus denaturant plots exhibit equal and opposite curvature in many of the mutants studied (Silow & Oliveberg, 1997; Otzen *et al.*, 1999). This curvature is interpreted by Silow and Oliveberg to imply that there is a change in the position of the folding transition-state with denaturant concentration (many of the mutants exhibit a very loose folding transition state in water, which qualitatively implies they should indeed fold faster than the NTP procedure predicts). Removing U1A from our data set on these grounds greatly improves the overall correlation (R=0.87), corresponding to an average prediction error of $e^{1.1} \approx 3.0$.

The NTP predictions apply to single domain proteins with single folding nuclei. Similar estimates can be made for multiple domain proteins if the nucleus formation events for different domains are independent (indeed, this assumption was used to predict the folding rates for the Arc repressor and the C-terminal domain of protein L9). Based on our previous estimates, we expect that single domain proteins longer than ~120 residues require more than a second to fold by a topomer sampling mechanism (Debe et al., 1999a), which would expose them to proteolysis *in vivo*. Possibly, the mechanisms that speed up the folding kinetics in alpha-helical proteins and the U1A family also allow protein domains beyond ~120 residues to fold on the shorter time scale appropriate for *in vivo* folding. Strong local structure propensities (Baldwin, 1993), early helix formation (Viguera *et al.*, 1997), and beta sheet inducing mechanisms such as glycosylation (O'Connor & Imperiali, 1998) are likely to play an important role during the *in vivo* folding of large protein domains.

To predict the folding rates for specific native proteins, we have developed a procedure for determining the probability of satisfying a specific set of contacts in a native protein structure. In addition to accurate rate prediction, our method explains the observed dependence of folding rate on solvent viscosity and provides a satisfying structural definition of the folding transition-state that correlates well with a nucleation-condensation picture of folding. Our approach is quite different from correlated energy landscape (Plotkin, *et al.*, 1996) and free energy functional (Shoemaker *et al.*, 1999) folding theories which use mean-field approximations to estimate the conformational entropy. Our approach avoids a mean-field treatment of contact probability, allowing it to be applied to native proteins very easily. However, unlike these theories, our method lacks any quantitative estimate of the enthalpy of various conformations [often given by the interaction energies of various contacts in other folding models (Miyazawa & Jernigan, 1985)]. Future work will seek to merge our approach into a theoretical framework that allows free energies to be estimated. We expect that our calculations can be used to tune the entropy estimates in mean-field approaches to specific native proteins [possibly using an interpolation between mean-field and specific contact probability formulations (Shoemaker *et al.*, 1999)], leading to unified theories that yield approximate free energy estimates simultaneously with accurate, experimentally verifiable rate predictions.

Supplementary Table 4.2-1. Monomeric λ Repressor (Helical)

PDB File	Contacts					$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
1LMB: $\mu = 10$	1-49	5-74	12-71	14-46	20-64	–13.83	–3.3
	28-42	31-63	35-56	54-79	57-68		
1LMB: $\mu = 8$	2-51	13-71	14-46	24-64	28-43	–12.63	–3.9
	35-59	54-79	57-68	-	-		
1LMB: $\mu = 7$	2-51	9-72	14-46	16-64	28-42	–12.51	–3.4
	35-59	54-79	-	-	-		

Supplementary Table 4.2-2. ACBP (Helical)

PDB File	Contacts					$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
2ABD: $\mu = 7$	3-70	6-35	15-85	21-50	27-73	–13.79	–3.9
	34-66	39-58	-	-	-		
2ABD: $\mu = 6$	3-70	6-35	15-85	21-50	34-66	–13.98	–5.0
	39-58	-	-	-	-		
2ABD: $\mu = 3$	6-35	15-85	21-50	-	-	–8.26	–6.9
	-	-	-	-	-		

Supplementary Table 4.2-3. Im9 (Helical)

PDB File	Contacts					$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
1IMQ: $\mu = 7$	7-44	9-84	15-40	16-68	23-36	–10.09	–4.8
	46-74	53-67	-	-	-		
1IMQ: $\mu = 6$	6-43	11-84	18-40	20-64	37-50	–12.38	–4.0
	46-74	54-79	57-68	-	-		

Supplementary Table 4.2-4. Tendamistat (Beta)

PDB File	Contacts					$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
2AIT: $\mu = 10$	7-71	11-27	16-67	19-59	23-54	–12.75	–3.9
	28-50	32-72	34-45	38-66	43-57		
2AIT: $\mu = 7$	7-71	11-27	19-59	28-50	32-72	–11.92	–4.6
	38-66	43-57	-	-	-		
2AIT: $\mu = 6$	7-71	19-59	28-50	32-72	38-66	–11.73	–5.1
	43-57	-	-	-	-		

Supplementary Table 4.2-5. Cold Shock Protein A & B (Beta)

PDB File	Contacts					ln*P(μ)*	ln*P(Ntop / μ)*
1CSP: $\mu = 10$	2-50	4-20	8-17	9-43	14-30	–9.63	–5.1
	18-26	26-59	34-64	36-67	51-60		
1CSP: $\mu = 8$	1-49	2-50	4-20	9-43	15-29	–10.23	–4.9
	26-59	34-64	36-67	-	-		
1MJC: $\mu = 10$	3-53	4-23	9-47	10-19	16-33	–12.49	–3.4
	20-29	30-62	37-67	49-69	54-64		
1MJC: $\mu = 7$	4-52	6-23	13-43	16-33	30-62	–12.68	–4.2
	36-65	50-69	-	-	-		

Supplementary Table 4.2-6. SH3 domains: α-spectrin, src & fyn (Beta)

PDB File	Contacts					ln*P(μ)*	ln*P(Ntop / μ)*
1NYF: $\mu = 5$	3-27	4-57	17-48	25-42	35-52	–11.30	–4.5
1TUD: $\mu = 7$	1-60	3-35	7-53	9-25	20-46	–12.10	–4.1
	26-40	44-59	-	-	-		

Supplementary Table 4.2-7. SH3 domains: PI3 kinase (Beta)

PDB File	Contacts					ln*P(μ)*	ln*P(Ntop / μ)*
1PKS: $\mu = 10$	1-30	3-76	8-23	9-70	18-65	–14.36	–3.9
	25-59	31-44	33-54	45-73	52-67		
1PKS: $\mu = 8$	1-76	2-31	8-23	9-70	18-64	–13.79	–4.6
	29-55	45-73	53-66	-	-		

Supplementary Table 4.2-8. Twitchin (Beta)

PDB File	Contacts					ln*P(μ)*	ln*P(Ntop / μ)*
1WIT: $\mu = 10$	2-82	5-26	9-87	16-93	19-62	–16.19	–4.4
	27-56	32-79	38-72	49-64	69-92		
1WIT: $\mu = 9$	1-82	5-25	7-86	15-92	18-64	–15.67	–4.3
	26-57	35-75	49-64	70-90	-		
1WIT: $\mu = 8$	3-84	5-25	15-92	19-62	26-57	–13.46	–4.8
	38-72	49-64	70-90	-	-		

Supplementary Table 4.2-9. ^{9}FN-III & Tenascin (Beta)

PDB File	Contacts					ln$P(\mu)$	ln$P(Ntop \mid \mu)$
1TEN: $\mu = 10$	2-26	8-20	12-89	14-63	22-54	–13.95	–4.5
	28-75	29-51	34-46	37-67	70-84		
1TEN: $\mu = 8$	1-80	4-24	14-89	15-61	28-75	–14.55	–5.5
	30-51	37-67	66-88	-	-		
1FNF: $\mu = 10$	3-27	4-82	15-90	16-62	22-55	–14.78	–4.8
	29-76	30-52	38-68	48-59	69-87		

Supplementary Table 4.2-10. B1 domain of protein G (Alpha-Beta)

PDB File	Contacts					ln$P(\mu)$	ln$P(Ntop \mid \mu)$
2GB1: $\mu = 8$	1-21	3-50	5-16	5-30	9-56	–11.36	–3.7
	23-45	31-40	42-55	-	-		
2GB1: $\mu = 6$	1-21	3-50	5-30	9-56	23-45	–10.51	–5.1
	42-55	-	-	-	-		

Supplementary Table 4.2-11. CI2 (Alpha-Beta)

PDB File	Contacts					ln$P(\mu)$	ln$P(Ntop \mid \mu)$
1COA: $\mu = 10$	1-24	3-63	5-20	12-56	17-29	–12.51	–4.6
	27-45	32-50	34-58	42-64	49-61		
1COA: $\mu = 6$	2-23	3-63	10-56	27-45	33-51	–12.35	–4.6
	44-64	-	-	-	-		
1COA: $\mu = 5$	2-23	10-56	27-45	33-51	44-64	–10.71	–5.1

Supplementary Table 4.2-12. ADAh2 (Alpha-Beta)

PDB File	Contacts					ln$P(\mu)$	ln$P(Ntop \mid \mu)$
1PBA: $\mu = 8$	3-62	11-79	12-55	19-48	26-73	–13.10	–4.1
	27-43	34-64	39-53	-	-		
1PBA: $\mu = 7$	3-62	11-79	11-56	19-48	27-44	–12.23	–3.9
	34-64	38-54	-	-	-		
1PBA: $\mu = 6$	3-62	11-55	13-77	19-49	27-43	–11.95	–4.6
	34-60	-	-	-	-		

Supplementary Table 4.2-13. Arc Repressor (Alpha-Beta)

PDB File	Contacts					$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
1ARR: μ = 2	9-34	22-37	-	-	-	–5.01	–6.1
1ARR: μ = 2	12-34	23-33	-	-	-	–4.32	–7.8

Supplementary Table 4.2-14. Ubiquitin (Alpha-Beta)

PDB File	Contacts					$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
1UBQ: μ = 10	1-18	2-64	4-14	8-69	15-30	–12.61	–3.9
	19-57	24-52	27-38	40-72	45-67		
1UBQ: μ = 7	1-17	2-64	8-69	15-30	23-53	–12.85	–4.1
	40-72	45-67	-	-	-		
1UBQ: μ = 6	1-17	8-69	15-30	23-53	40-72	–11.80	–4.4
	45-67	-	-	-	-		

Supplementary Table 4.2-15. IgG binding domain of protein L (Alpha-Beta)

PDB File	Contacts					$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
2PTL: μ = 10	1-22	5-55	7-17	8-35	11-61	–11.23	–4.8
	20-30	25-54	32-47	42-62	47-57		
2PTL: μ = 6	1-22	5-55	8-35	11-61	25-54	–11.48	–5.5
	42-62	-	-	-	-		

Supplementary Table 4.2-16. Hpr (Alpha-Beta)

PDB File	Contacts					$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
1HDN: μ = 10	1-66	6-78	7-60	14-85	15-55	–14.67	–3.5
	22-80	23-47	30-68	34-43	38-61		
1HDN: μ = 8	1-66	2-71	9-58	14-85	23-47	–13.00	–4.1
	30-68	34-43	38-61	-	-		
1HDN: μ = 7	1-66	9-58	14-85	23-47	30-68	–11.77	–3.8
	34-43	38-61	-	-	-		

Supplementary Table 4.2-17. FKBP12 (Alpha-Beta)

PDB File			Contacts			$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
1FKB: $\mu = 10$	2-76	10-70	16-106	19-50	21-107	–14.92	–5.5
	26-39	31-97	56-81	71-102	78-95		
1FKB: $\mu = 8$	3-75	10-70	19-50	21-106	26-38	–15.71	–5.7
	31-96	56-81	77-96	-	-		

Supplementary Table 4.2-18. Muscle and Common Type AcP (Alpha-Beta)

PDB File			Contacts			$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
1APS: $\mu = 8$	6-55	7-85	15-76	17-44	27-94	–16.00	–6.2
	35-52	40-98	53-89	-	-		
2ACY: $\mu = 10$	6-55	7-85	15-76	18-45	25-70	–16.02	–6.2
	27-94	35-52	40-98	53-89	62-80		
2ACY: $\mu = 9$	2-55	7-85	14-77	18-45	25-70	–15.65	–6.9
	27-94	33-57	40-97	52-89	-		

Supplementary Table 4.2-19. N-terminal domain of protein L9 (Alpha-Beta)

PDB File			Contacts			$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
1DIV: $\mu = 7$	1-21	3-38	6-14	7-35	15-51	–9.32	–3.7
	18-44	27-37	-	-	-		
1DIV: $\mu = 6$	1-23	3-39	5-17	8-35	15-51	–8.98	–3.8
	27-37	-	-	-	-		

Supplementary Table 4.2-20. Spliceosomal U1A & Ribosomal S6 (Alpha-Beta)

PDB File			Contacts			$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
1URN: $\mu = 10$	3-84	9-89	10-57	16-77	17-52	–14.36	–5.8
	22-46	30-42	36-63	43-55	68-83		
1URN: $\mu = 7$	10-62	11-87	15-54	18-78	23-45	–12.40	–6.9
	37-68	41-58	-	-	-		
1RIS: $\mu = 8$	3-66	5-91	11-84	12-58	26-79	–14.72	–5.9
	33-71	39-64	44-59	-	-		

Supplementary Table 4.2-21. C-terminal domain of protein L9 (Alpha-Beta)

PDB File			Contacts			$\ln P(\mu)$	$\ln P(Ntop \mid \mu)$
1DIV: $\mu = 9$	9-77	16-83	21-49	27-91	31-67	–12.59	–5.3
	39-56	43-56	58-76	69-89	-		
1DIV: $\mu = 8$	9-81	19-86	21-49	25-92	31-68	–13.98	–6.2
	39-56	59-75	69-89	-	-		

REFERENCES

Baldwin, R. L. (1993). Pulsed H/D-Exchange Studies of Folding Intermediates. *Curr. Opin. Struc. Biol.* **3**, 84.

Bhattacharyya, R. P. & Sosnick, T. R. (1999). Viscosity dependence of the folding kinetics of a dimeric and monomeric coiled coil. *Biochemistry* **38**, 2601.

Burton, R. E., Huang, G. S., Daugherty, M. A., Fullbright, P. W. & Oas, T. G. (1996). Microsecond protein folding through a compact transition state. *J. Mol. Biol.* **263**, 311.

Chan, H. S. & Dill, K. A. (1990). The Effects of Internal Constraints On the Configurations of Chain Molecules. *J. Chem. Phy.* **92**, 3118.

Clarke, J., Hamill, S. J. & Johnson, C. M. (1997). Folding and stability of a fibronectin type III domain of human tenascin. *J. Mol. Biol.* **270**, 771.

Debe, D. A., Carlson, M. J. & Goddard, W. A. (1999a). The topomer-sampling model of protein folding. *Proc. Natl Acad. Sci. USA* **96**, 2596.

Debe, D. A., Carlson, M. J., Sadanobu, J., Chan, S. I. & Goddard, W. A. (1999b). Protein fold determination from sparse distance restraints: The Restrained Generic Protein Direct Monte Carlo method. *J. Phys. Chem. B* **103**, 3001.

Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen, *Ann. Phys. (Leipzig)* **17**, 549.

Ferguson, N., Capaldi, A. P., James, R., Kleanthous, C. & Radford, S. E. (1999). Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9. *J. Mol. Biol.* **286**, 1597.

Fersht, A. R. (1995). Optimization of Rates of Protein-Folding - the Nucleation-Condensation Mechanism and Its Implications. *Proc. Natl Acad. Sci. USA* **92**, 10869.

Flory, P. J. (1956). Theory of elastic mechanisms in fibrous proteins. *J. Am. Chem. Soc.* **78**, 5222.

Ghaemmaghami, S., Word, J. M., Burton, R. E., Richardson, J. S. & Oas, T. G. (1998). Folding kinetics of a fluorescent variant of monomeric lambda repressor. *Biochemistry* **37**, 9179.

Grantcharova, V. P. & Baker, D. (1997). Folding dynamics of the src SH3 domain. *Biochemistry* **36**, 15685.

Guijarro, J. I., Morton, C. J., Plaxco, K. W., Campbell, I. D. & Dobson, C. M. (1998). Folding kinetics of the SH3 domain of PI3 kinase by real-time NMR combined with optical spectroscopy. *J. Mol. Biol.* **276**, 657.

Gutin, A. M. & Shakhnovich, E. I. (1994). Statistical-Mechanics of Polymers With Distance Constraints. *J. Chem. Phy.* **100**, 5290.

Hagen, S. J., Hofrichter, J., Szabo, A. & Eaton, W. A. (1996). Diffusion-limited contact formation in unfolded cytochrome c: Estimating the maximum rate of protein folding. *Proc. Natl Acad. Sci. USA* **93**, 11615.

Hamill, S. J., Meekhof, A. E. & Clarke, J. (1998). The effect of boundary selection on the stability and folding of the third fibronectin type III domain from human tenascin. *Biochemistry* **37**, 8071.

Huang, G. S. & Oas, T. G. (1995). Structure and Stability of Monomeric Lambda-Repressor - Nmr Evidence For 2-State Folding. *Biochemistry* **34**, 3884.

Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The Structure of the Transition-State For Folding of Chymotrypsin Inhibitor-2 Analyzed By Protein Engineering Methods - Evidence For a Nucleation-Condensation Mechanism For Protein-Folding. *J. Mol. Biol.* **254**, 260.

Jackson, S. E. (1998). How do small single-domain proteins fold? *Folding Design* **3**, R81.

Jackson, S. E. & Fersht, A. R. (1991). Folding of Chymotrypsin Inhibitor-2 .1. Evidence For a 2-State Transition. *Biochemistry* **30**, 10428.

Jacob, M., Schindler, T., Balbach, J. & Schmid, F. X. (1997). Diffusion control in an elementary protein folding reaction. *Proc. Natl Acad. Sci. USA* **94**, 5622.

Karplus, M. & Weaver, D. L. (1976). Protein folding dynamics. *Nature* **260**, 404.

Khorasanizadeh, S., Peters, I. D. & Roder, H. (1996). Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. *Nature Struc. Biol.* **3**, 193.

Kragelund, B. B., Hojrup, P., Jensen, M. S., Schjerling, C. K., Juul, E., Knudsen, J. & Poulsen, F. M. (1996). Fast and one-step folding of closely and distantly related homologous proteins of a four-helix bundle family. *J. Mol. Biol.* **256**, 187.

Kragelund, B. B., Robinson, C. V., Knudsen, J., Dobson, C. M. & Poulsen, F. M. (1995). Folding of a 4-Helix Bundle - Studies of Acyl-Coenzyme-a Binding-Protein. *Biochemistry* **34**, 7217.

Kuhlman, B., Luisi, D. L., Evans, P. A. & Raleigh, D. P. (1998). Global analysis of the effects of temperature and denaturant on the folding and unfolding kinetics of the N-terminal domain of the protein L9. *J. Mol. Biol.* **284**, 1661.

Ladurner, A. G., Itzhaki, L. S., Daggett, V. & Fersht, A. R. (1998). Synergy between simulation and experiment in describing the energy landscape of protein folding. *Proc. Natl Acad. Sci. USA* **95**, 8473.

Lazaridis, T. & Karplus, M. (1997). "New-View" of protein folding reconciled with the old through multiple unfolding simulations. *Science* **278**, 1928.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59.

Maiorov, V. N. & Crippen, G. M. (1995). Size-Independent Comparison of Protein 3-Dimensional Structures. *Proteins: Struct. Funct. Genet.* **22**, 273.

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **218**, 623.

O'Connor, S. E. & Imperiali, B. (1998). A molecular basis for glycosylation-induced conformational switching. *Chem. Biol.* **5**, 427.

Otzen, D. E., Kristensen, O., Proctor, M. & Oliveberg, M. (1999). Structural changes in the transition state of protein folding: Alternative interpretations of curved chevron plots. *Biochemistry* **38**, 6499.

Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M. A., Jaenicke, R. & Schmid, F. X. (1998). Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nature Struc. Biol.* **5**, 229.

Plaxco, K. W. & Baker, D. (1998). Limited internal friction in the rate-limiting step of a two- state protein folding reaction. *Proc. Natl Acad. Sci. USA* **95**, 13591.

Plaxco, K. W., Guijarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D. & Dobson, C. M. (1998a). The folding kinetics and thermodynamics of the Fyn-SH3 domain. *Biochemistry* **37**, 2529.

Plaxco, K. W., Simons, K. T. & Baker, D. (1998b). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985.

Plaxco, K. W., Spitzfaden, C., Campbell, I. D. & Dobson, C. M. (1997). A comparison of the folding kinetics and thermodynamics of two homologous fibronectin type III modules. *J. Mol. Biol.* **270**, 763.

Plotkin, S. S., Wang, J. & Wolynes, P. G. (1996). Correlated energy landscape model for finite, random heteropolymers. *Phys. Rev. E* **53**, 6271.

Reid, K. L., Rodriguez, H. M., Hillier, B. J. & Gregoret, L. M. (1998). Stability and folding properties of a model beta-sheet protein, Escherichia coli CspA. *Protein Sci.* **7**, 470.

Robinson, C. R. & Sauer, R. T. (1996). Equilibrium stability and sub-millisecond refolding of a designed single-chain arc repressor. *Biochemistry* **35**, 13878.

Sato, S., Kuhlman, B., Wu, W. J. & Raleigh, D. P. (1999). Folding of the multidomain ribosomal protein L9: The two domains fold independently with remarkably different rates. *Biochemistry* **38**, 5643.

Scalley, M. L., Yi, Q., Gu, H. D., McCormack, A., Yates, J. R. & Baker, D. (1997). Kinetics of folding of the IgG binding domain of peptostreptoccocal protein L. *Biochemistry* **36**, 3373.

Schindler, T., Herrler, M., Marahiel, M. A. & Schmid, F. X. (1995). Extremely Rapid Protein-Folding in the Absence of Intermediates. *Nature Struc. Biol.* **2**, 663.

Schonbrunner, N., Koller, K. P. & Kiefhaber, T. (1997). Folding of the disulfide-bonded beta-sheet protein tendamistat: Rapid two-state folding without hydrophobic collapse. *J. Mol. Biol.* **268**, 526.

Shoemaker, B. A., Wang, J. & Wolynes, P. G. (1999). Exploring structures in protein folding funnels with free energy functionals: the transition state ensemble. *J. Mol. Biol.* **287**, 675.

Silow, M. & Oliveberg, M. (1997). High-energy channeling in protein folding. *Biochemistry* **36**, 7633.

Smith, C. K., Bu, Z. M., Anderson, K. S., Sturtevant, J. M., Engelman, D. M. & Regan, L. (1996). Surface point mutations that significantly alter the structure and stability of a protein's denatured state. *Protein Sci.* **5**, 2009.

Taddei, N., Chiti, F., Paoli, P., Fiaschi, T., Bucciantini, M., Stefani, M., Dobson, C. M. & Ramponi, G. (1999). Thermodynamics and kinetics of folding of common-type acylphosphatase: Comparison to the highly homologous muscle isoenzyme. *Biochemistry* **38**, 2135.

van Nuland, N. A. J., Chiti, F., Taddei, N., Raugei, G., Ramponi, G. & Dobson, C. M. (1998). Slow folding of muscle acylphosphatase in the absence of intermediates. *J. Mol. Biol.* **283**, 883.

van Nuland, N. A. J., Meijberg, W., Warner, J., Forge, V., Scheek, R. M., Robillard, G. T. & Dobson, C. M. (1998). Slow cooperative folding of a small globular protein HPr. *Biochemistry* **37**, 622.

Viguera, A. R., Martinez, J. C., Filimonov, V. V., Mateo, P. L. & Serrano, L. (1994). Thermodynamic and Kinetic-Analysis of the Sh3 Domain of Spectrin Shows a 2-State Folding Transition. *Biochemistry* **33**, 2142.

Viguera, A. R., Serrano, L. & Wilmanns, M. (1996). Different folding transition states may result in the same native structure. *Nature Struc. Biol.* **3**, 874.

Viguera, A. R., Villegas, V., Aviles, F. X. & Serrano, L. (1997). Favourable native-like helical local interactions can accelerate protein folding. *Folding Design* **2**, 23.

Villegas, V., Azuaga, A., Catasus, L., Reverter, D., Mateo, P. L., Aviles, F. X. & Serrano, L. (1995). Evidence For a 2-State Transition in the Folding Process of the Activation Domain of Human Procarboxypeptidase-A2. *Biochemistry* **34**, 15105.

CHAPTER 5

UNDERSTANDING NON-RANDOM PROCESSES IN EARLY PROTEIN FOLDING

ABSTRACT

In chapter 4, we demonstrated that the rate of folding for single-domain proteins can be determined from first principles by assuming that the rate-limiting step in folding is a random, diffusive search for the native tertiary topology. In this chapter, we examine two different types of experimental data that further suggest a diffusive folding mechanism: (i) the ratio of the refolding rates for two single disulphide tendamistat variants, and (ii) the concentration of single disulphide bond trapped intermediates observed in the folding of Ribonuclease A and BPTI. Using the NTP method developed in chapter 4, we show that the experimentally determined folding rates for these variants behave according to a diffusive folding mechanism. In the refolding of RNase A and BPTI, disulphide bonds further apart in sequence are less frequently observed than those that are closer in sequence, also expected by the diffusive model. However, there are a few disulphide pairs that do not behave in strict accordance with a diffusive mechanism. These outliers provide clues about the nature of non-random interactions that can bias the predominantly diffusive search for the native state. In Ribonuclease A, the fast formation of a disulphide bond between residues 65 and 72 suggests that a β-turn centered at residues 68-69 biases the early folding process. In BPTI, the rate of disulphide formation is anomalous for a triad of disulphide bonds, where 14-30 forms very slowly, while 14-38 and 30-38 both form very quickly. Using the NTP procedure to investigate this non-random behavior, we find that these anomalous rates of disulphide bond formation are caused by the fast formation of a hydrophobic core and a propensity for amino acids with large, branched side-chains to lie in the extended, β-strand region of the Ramachandran plot.

INTRODUCTION

In chapter 4, proteins composed of predominantly helical residues all folded significantly faster than expected according to a diffusive mechanism. Folding rates for beta-sheet proteins and α/β proteins with low overall helical content were consistent with the diffusive mechanism predicted rates. This suggests that helix formation represents an early local interaction that increases the rate of folding. In this chapter, we investigate the nature of other non-random processes present in early protein folding.

METHODS AND DISCUSSION

TENDAMISTAT: A SYSTEM WITH PREFORMED DISULPHIDE BONDS

One protein that does not contain any helical residues is the 74-residue α-amylase inhibitor tendamistat. This small protein contains two native disulphide bonds, one between cysteines 11 and 27, the other between cysteines 45 and 73. In order to understand the effect of preformed tertiary structure on the rate of protein folding, Kiefhaber and coworkers have measured the folding rate of wild type tendamistat (both disulphides), as well as each of the two possible single disulphide variants (prepared using site directed mutagenesis) [1]. They find that the rate of the refolding reaction is decreased by 30-fold for the C45A/C73A mutant and 8-fold for the C11A/C27S mutant.

Given their results, it is interesting to ask whether they are adequately explained by a random, diffusive folding mechanism. The wild-type tendamistat contains two disulphide bonds. The mutants each possess one disulphide. In a diffusive mechanism, the rate of folding will increase as the number of accessible topologies decreases. This behavior is indeed observed in the Kiefhaber work, since the double-disulphide wild-type variant folds faster than either single-disulphide mutant. Thus, qualitatively, the results of these experiments are described by the diffusive folding mechanism.

We can also ask if the quantitative results of these experiments are accurately described by a diffusive folding mechanism. In the experiments, the variant lacking the long disulphide loop between residues 45 and 73 folds 3.75 times slower than the variant that lacks the short loop between residues 11 and 27.

The NTP method developed in chapter 4 can be used to determine if this folding rate reduction is expected given a diffusive folding process. Using NTP, it is possible to determine the probability of forming loop 45-73 when loop 11-27 is already formed, P(45-73|11-27), and conversely, the probability of forming loop 11-27 when loop 45-73 is already formed P(11-27|45-73). The ratio P(11-27|45-73)/P(45-73|11-27) represents the fraction of topologies that the variant with the 45-73 loop intact will not have to search compared to the protein with only the 11-27 loop intact. Thus, in our diffusive, topology sampling model, this ratio should be similar to the experimentally determined rate ratio of 3.75.

Using the NTP method, we find that P(11-27|45-73)/P(45-73|11-27) = (61.0/17.4) = 3.51. This correlates well with the experimental results, substantiating the topology-sampling model for tendamistat.

DISULPHIDE BOND INTERMEDIATES: RNaseA and BPTI

Experiments that monitor the concentration of single disulphide bond intermediates observed in the folding of proteins containing two or more disulphide bonds provide a unique fingerprint for determining other examples of local or non-local phenomena that bias the early search for the native topology. In a truly random folding process, one would expect the following:

(i) The various disulphide intermediates will collect with kinetic concentrations that decrease as the sequential distance between the participating cysteine residues increases, and

(ii) there will not be a bias towards those disulphide bonds that are found in the native state over those that are not present in the protein's native state.

For the two systems where there is a complete body of experimental data, Ribonuclease A (RNase A) and BPTI, points (i) and (ii) are satisfied by nearly all of the possible single disulphide pairings. However, there are some deviations from random behavior present in these systems. These deviations provide unique clues about the nature of non-random interactions that can bias a protein's search for the native state.

Bovine pancreatic RNase A is a 124-residue protein with four native disulphide bonds: [26-84], [40-95], [58-110], and [65-72]. Scheraga and coworkers have studied the formation of disulphide bond intermediates in this system for more than 20 years [2]. In 1999, they finished the complete characterization of the distribution of all one-disulphide bonds in the two-disulphide bond intermediates in the oxidative refolding of RNase A [3].

The first important finding from this work is that all of the 28 possible disulphide single disulphide pairs were found in the two-disulphide intermediates, demonstrating that all of the non-native disulphide bond pairs were accessed in the early folding events. Thus criteria (ii) above is satisfied in the RNase A system.

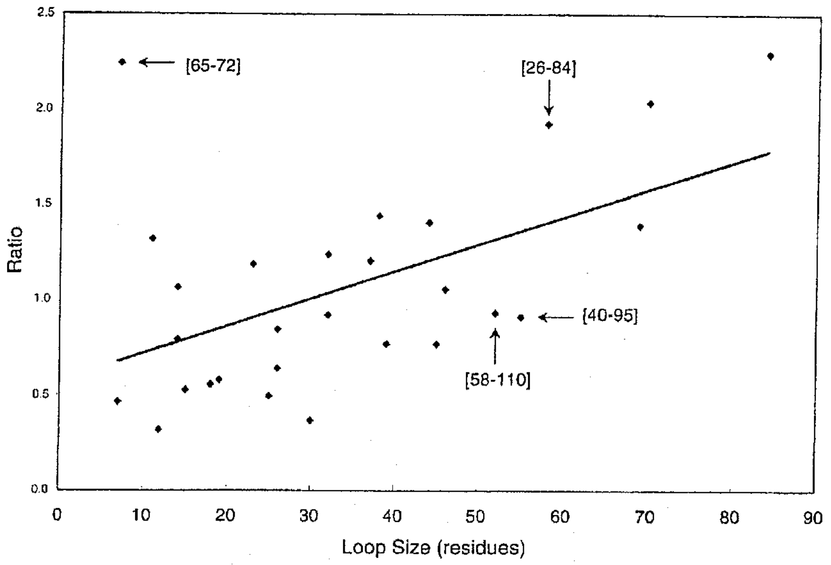
The second important finding in this work lies in an analysis of the observed single-disulfide bond concentrations. Two loop systems are simple enough that an expression for the expected entropy change for an arbitrary distribution of two disulphide loops in a protein may be calculated. In their work, Scheraga and coworkers use the following expression [4] for the entropy change for adding two disulphide loops, which is based on a random flight statistical mechanical model [5, 6] and on the Wang-Uhlenbeck expression for multivariate Gaussian distributions:

$$\Delta S^\circ = R(-6.94 + 6\ln(a) - 1.5\ln|C|), \qquad (1)$$

where R is the gas constant, a is the length of a chain element, and |C| is the determinant of C, where C is a 2x2 matrix whose elements are $C_{ij} = (a^2)$ x (# of shared residues by loops i and j). Thus equation (1) yields the entropy of formation for any set of independent or overlapping pair of disulphide bonds. After determining the entropic cost of forming each of the 210 possible two-disulphide bond intermediates for RNase A system, they calculated the expected distribution of the 28 possible single disulphide bonded species (includes the 4 pairs observed in the native state, as well as 24 non-native disulphide possibilities). Since equation (1) only considers the entropic contributions and is based on a random-walk model, these calculated distributions represent what would be expected if a random search process represented the rate-limiting step in the folding process.

Figure 5.1 on the following page shows the results of Scheraga and coworkers study. Since equation (1) suffers from a systematic error as the length of the disulphide loop size increase, the y-axis is the ratio of the experimentally determined percentage occupancy to the theoretically predicted percentage occupancy, while the x-axis is the loop length for each of the 28 single disulphide loops. Thus in this plot, deviations from ideal diffusive behavior exist for points that lie significantly off of the diagonal line. For this data set, there is only one such point, the native disulphide between residues 65 and 72. It lies well above the line, implying that its occupancy is much higher than predicted. Scheraga and coworkers suggest that this is due to an energetically favorable β-turn centered at residues 68 and 69. This appears to be one mechanism that reduces the number of accessible topologies during early folding.

Figure 5.1 The ratio of the experimental to theoretical abundance for each possible disulphide bond as a function of the sequence length of the disulphide loop. The four native disulphide bonds are labeled.


2.5
2.0
1.5
1.0
0.5
0.0
Ratio
0
10
20
30
40
50
60
70
80
90
Loop Size (residues)
[65-72]
[26-84]
[40-95]
[58-110]

It is very interesting that the occupancies for many of the non-native disulphide intermediates lie above the best line through the data, while two of the four native disulphide configurations lie below this line. Furthermore, *each and every* one of the possible 28 single disulphide configurations are observed with occupancies above or near their diffusion predicted value. This confirms that a large and *possibly* complete set of topologies is accessed during the rate-limiting search for the native topology.

Another system whose single disulphide intermediates have been thoroughly characterized is BPTI (bovine pancreatic trypsin inhibitor). BPTI is a 58-residue protein with three native disulphide bonds, residues 5-55, 14-38, and 30-51. Since there are six total cysteines, 15 one-disulphide bond intermediates are possible. As in RNase A, all of the single disulphide intermediates are observed during BPTI folding.

By monitoring the glutathione-mediated disulphide bond formation of the intramolecular disulphide bonds in BPTI, Dadlez and Kim [7] have determined the effective concentration, C_{eff}, for all 15 single disulphide bond intermediates for BPTI. The data they obtained is shown below in Table 5.1.

In addition to the Dadlez and Kim experimental data, Table 5.1 also contains the results obtained by applying the NTP method to this problem. In the column labeled "P(Diffusion)," we have the results of the NTP method where a purely diffusive mechanism has been assumed. An ensemble of 10,000 unconstrained BPTI chains was grown, and the probability of placing the residue pairs from column one within 9.5Å (a bond order of 3) of one another was determined (a 0.05 implies that 0.05 x 10,000 = 500 chains satisfied the 9.5Å restraint).

Cysteine Pair	C_{eff}	P(Phobic Core)	P(Core+Stiff)
5-14	16.9	0.11	0.08
5-30	4.3	0.03	0.02
5-38	6.9	0.04	0.04
5-51	6.2	0.01	0.01
5-55	6.3	0.00	0.01
14-30	4.1	0.05	0.01
14-38	27.0	0.22	0.13
14-51	7.5	0.04	0.02
14-55	6.3	0.03	0.02
30-38	17.1	0.06	0.07
30-51	12.3	0.03	0.04
30-55	6.5	0.01	0.01
38-51	9.5	0.03	0.06
38-55	8.1	0.04	0.03
51-55	14.0	-	-

Table 5.1 The experimentally observed effective concentrations of different single disulphide intermediates in BPTI. The experimental data lies in column "C_{eff}," while the NTP predictions for the probability of formation lie in the final two columns (see text for description). The final disulphide, 51-55, is only 4 residues long and is therefore too short to obtain a reliable value using the NTP method.

Since only the single-disulphide bonded intermediates were considered in this study, Dadlez and Kim were able to assess which disulphide bonds were deviating significantly from ideal diffusive behavior simply by plotting C_{eff} versus length. They found that pair 14-38 forms much faster than expected, while the pair 14-30 folds much slower.

In a follow-up to this original work [8], Dadlez studies a series of 24 BPTI mutants and found that eight non-polar or aromatic sidechain mutations (I18A, I19A, Y21A, F22A, Y23A, F33A, V34A, and Y35A) all decrease the rate of folding. Because of this, Dadlez proposed that the fast formation of hydrophobic core is responsible for the increased presence of the 14-38 intermediate.

The NTP method can be used to understand the effect that early hydrophobic core formation has on the rate of formation of the various disulphide intermediates. The data in Table 5.1 under the column heading "P(Phobic Core)," is the probability of forming each of the various disulphide pairs when residues 16 and 35 are constrained to be within 10Å of each other (only chains that meet this condition can contribute to the probability values). This approximates the existence of a crude hydrophobic core that contains or is adjacent to both residues 16 and 35. This data is plotted versus C_{eff} in Figure 5.2.

Figure 5.2 Predicted P(Phobic-Core) versus experimental C_{eff}.

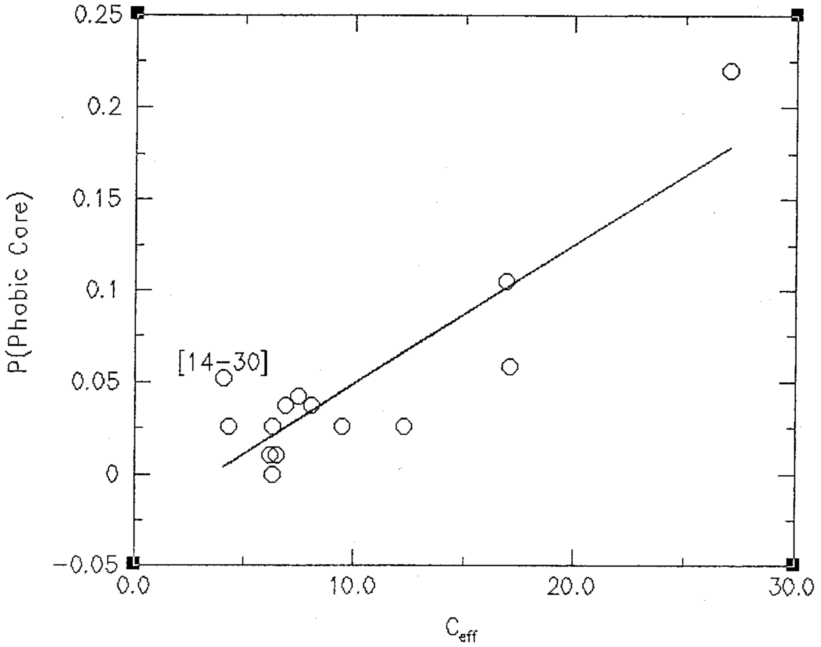
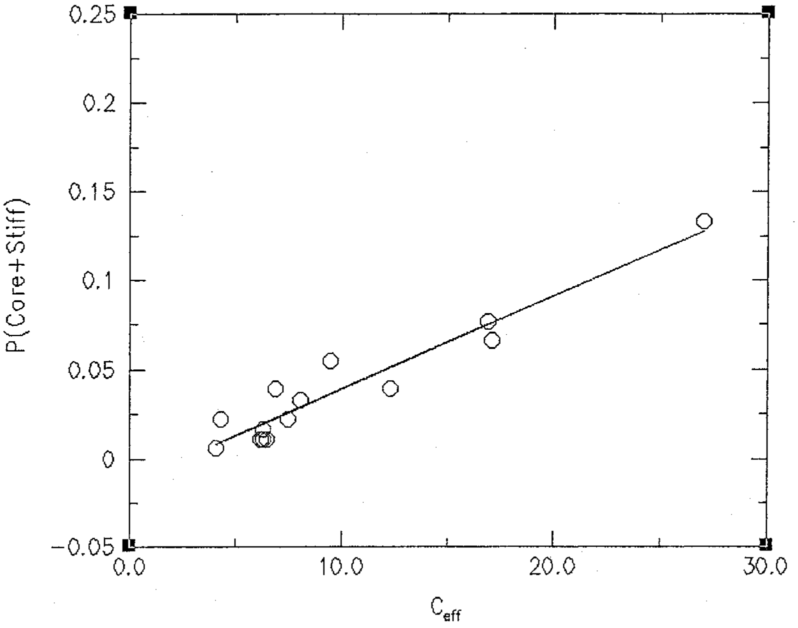

0.25
0.2
0.15
0.1
0.05
0
−0.05
P(Phobic Core)
[14–30]
0.0
10.0
20.0
30.0
C_{eff}

Figure 5.3 Predicted P(Core+Stiff) versus experimental C_{eff}.


0.25
0.2
0.15
0.1
0.05
0
−0.05
P(Core+Stiff)
0.0
10.0
20.0
30.0
C_{eff}

From Figure 5.2, we see that while the addition of the hydrophobic core restraint capably accounts for the observed increase in the 14-38 intermediate, it does not account for the observed decrease in the 14-30 intermediate. The fact that the 14-38 pair and the 30-38 pair both fold very quickly rules out the possibility that the observed decrease in the 14-30 pair is due to an experimental artifact caused by the local environment of either CYS14 or CYS30. However, Dadelez has not proposed a mechanism for the slower than expected formation of the 14-30 pair.

Examining the amino acid sequence between residues 18 and 25 in BPTI reveals that five out of the seven residues are large branched or aromatic sidechains. Street and Mayo [9] have demonstrated that the increased propensity for Threonine, Isoleucine, Valine, Tyrosine, and Phenylalanine residues to be involved in β-sheets is due to the van der Waals interactions between the large sidechains and the local backbone.

The data in Table 5.1, under the column heading "P(Core+Stiff)," is the probability of formation of the various disulphide pairs when the hydrophobic core restraint described above is imposed, while residues 19-23 are extended, like a stiff β-strand. This stiffness criterion is accomplished in the ensemble growth by limiting the ϕ torsion angle to 145, 165, 180, 195, and 215 degrees, instead of the usual 0, 60, 120, 180, 240, and 300 degrees. The data in this column is plotted versus C_{eff} in Figure 5.3. The correlation for the best line through the data is very high (R=0.96), and the nonrandom behavior of pairs 14-30 and 14-38 are adequately explained.

CONCLUSION

In this chapter we have demonstrated how the NTP method can be used to probe and understand non-random processes that occur during early protein folding. Analytical estimates for the entropy loss during disulphide formation become inaccurate when the number of bonds exceeds one. Furthermore, analytical estimates cannot address even simple important processes, such as β-turn catalysis and space-filling effects. The NTP computational approach provides a unique tool capable of taking into account many different physical aspects of the folding reaction.

REFERENCES

1. N. Schönbrunner, et al., *Biochemistry* **36**, 9057 (1997).
2. Y. Konishi and H. A. Scheraga, *Biochemistry* **19**, 208 (1980).
3. M. J. Volles, X. Xu, and H. A. Scheraga, *Biochemistry* **38**,7284 (1999).
4. S. H. Lin, Y. Konishi, M. E. Dentonand, and H. A. Scheraga, *Biochemistry* **23**, 5504 (1984).
5. P. J. Flory, Principles of Polymer Chemistry, Chapter X, Cornell University Press, Ithaca, NY (1953).
6. D. C. Poland and H. A. Scheraga, *Biopolymers* **3**, 379 (1965).
7. M. Dadlez and P. S. Kim, *Biochemistry* **35**, 16153 (1996).
8. M. Dadlez, *Biochemistry* **36**, 2788 (1997).
9. A. G. Street and S. L. Mayo, *Proc. Natl. Acad. Sci.* **96**, 9074 (1999).

CHAPTER 6

THE SIMPLEST EMPERICAL FOLDING MODEL:

THE HELIX-BIASED DIFFUSION MODEL

ABSTRACT

Chapters 2-5 have focused on the development and application of a first principles calculation whose foundation is a diffusive, topology-sampling model for protein folding. In this final chapter, we leave the first principles calculation to present a very simple Helix-Biased-Diffusion (HBD) empirical model that is also rooted in the topology diffusion model. This new empirical model accurately predicts the folding rate for all small two-state folding proteins, helical and non-helical alike. Furthermore, we show that this empirical model also applies to larger, multi-domain proteins as well. The rates for these larger, more complex systems can be determined once the rate-limiting folding unit (RLFU) is identified. Finally, we analyze the "contact order" empirical model for predicting folding rates, an alternative empirical model that has recently received significant literature attention.

INTRODUCTION

In 1991, Jackson and Fersht [1] measured the first observed two-state folding rate for a small protein, Chymotrypsin Inhibitor 2. Since then, two-state folding rates for more than 20 other proteins have been determined, providing sufficient rate data to begin testing quantitative aspects of proposed folding mechanisms [2].

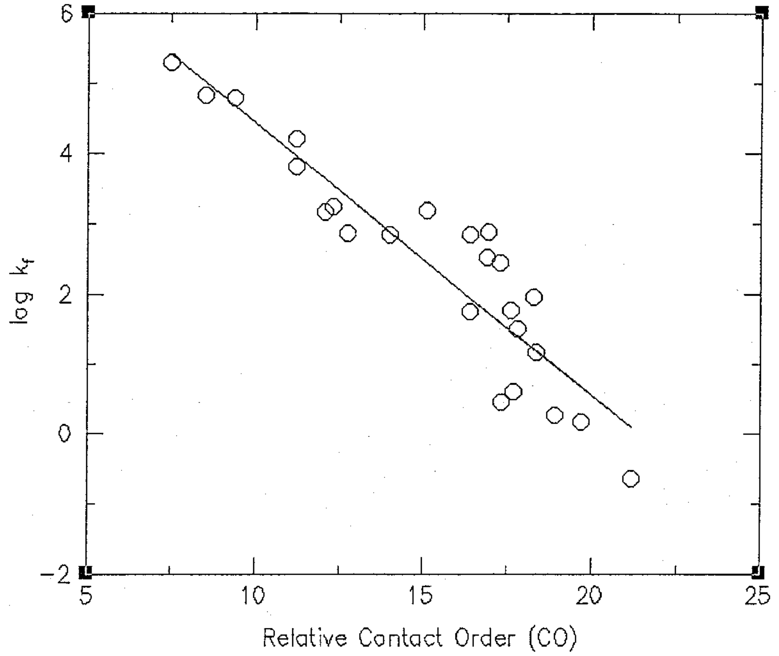
The first attempt to make some sense of all of the available two-state folding rate data was made in 1998 by Kevin Plaxco, then a post-doc in David Baker's laboratory at the University of Washington [3]. They reported a statistically significant correlation between the natural log of the two-state folding rate, ln(k_f), and a measure of the native state topological complexity, which they referred to as relative contact order (CO). While there were only 12 distinct proteins whose rates had been determined at the time of their publication, their ability to fit the rate date with a simple empirical equation suggested that the mechanism underlying the folding of single-domain proteins could be very simple.

Relative contact order is defined as the average sequence distance between all pairs of contacting residues normalized by the total sequence length:

$$CO = \frac{1}{LN} \sum^{N} \Delta S_{i,j} \, , \tag{1}$$

where N is the total number of contacts, $\Delta S_{i,j}$ is the sequence separation between contacting residues i and j, and L is the total number of residues in the protein (here residues are considered in contact if they have two non-hydrogen atoms within 6Å in the native structure). Figure 6.1 shows the correlation between the log of the folding rate, log(k_f), versus relative contact order for a set of 24 two-state folding proteins [4].

Figure 6.1 The log of the folding rate, log(k_f) versus the relative contact order for 24, 2-state folding proteins. The correlation for the best line through the data points is 0.92.


6
4
2
0
−2
log k_f
5
10
15
20
25
Relative Contact Order (CO)

While relative contact order can be described very simply, e.g., "the average sequence distance between all pairs of contacting residues normalized by the total sequence length," no one to date has been able to relate this empirical equation to a fundamental, physical folding mechanism. Fersht [5] has recently made some simple arguments that relate relative contact order to chain entropy; however, no direct correlation to a physical folding mechanism has been made.

In chapter 4, proteins composed of predominantly helical residues all folded significantly faster than what would be expected given a diffusive mechanism. This suggests that helix formation represents an early local interaction that biases the folding and dramatically increases the rate of folding. In chapter 5, we analyzed the concentration of disulphide intermediates in BPTI and RNase A in order to find other non-random processes that affect the predominantly early diffusion search for the native protein topology. Our analysis of the available experimental studies suggests that early β-turn formation plays a role in RNase A folding, while early hydrophobic collapse and β-strand stiffening play a role in BPTI folding.

Now that we have determined that the physical processes that play a role in the rate-limiting step in protein folding, it is interesting to see if we can develop a simple empirical model based on the physics that can accurately predict the rate that a protein will fold. This chapter is devoted to developing and testing such a model.

METHODS AND DISUCSSION

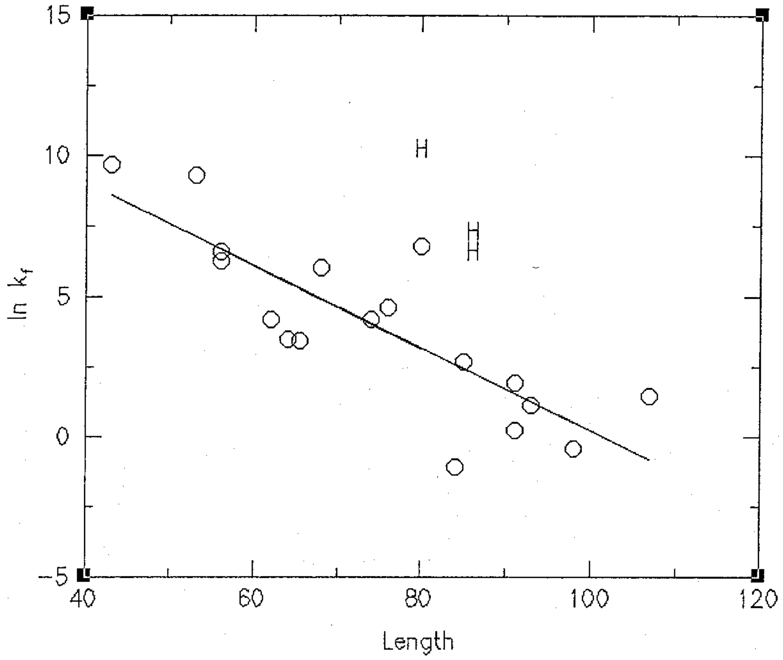
THE HELIX-BIASED DIFFUSION MODEL

Given the results of our studies in chapters 2-5, it is evident that the most dominant term in our empirical estimate should represent the process of random, diffusive sampling to find the native topology. In chapter 2, we used the Generic Protein model to demonstrate that this diffusive process is represented by a linear dependence between the log of the folding rate, $\log(k_f)$, and the length of the protein. Thus, our empirical rate estimate should have the form:

$$\ln(k_f) = A + B(N_{res}), \tag{2}$$

where N_{res} is the number of residues in the protein, A is a positive constant representing the maximum folding rate, and B is a negative constant so that the rate of folding increases as the number of residues increases. This simple equation contains the essence of a random diffusive search process for the native topology. Figure 6.2 shows a plot of the log of the experimentally observed rate, $\ln(k_f)$, versus protein length for 21, 2-state folding proteins (data from Table 6.1). The three predominantly helical proteins are labeled in the plot by the letter H. There is a strong correlation for the 18 β and α/β proteins (R=0.83).

Figure 6.2 Experimentally observed ln(k_f) versus length for 21, 2-state folding proteins. Each of the three predominantly helical proteins is labeled with an "H."


15
10
5
0
−5
40
60
80
100
120
ln k_f
Length
H
H
H

From Figure 6.2, it is evident that proteins with significant helical content fold faster than a random, diffusive rate-limiting mechanism would suggest. Thus, it will be necessary to add a term to our empirical estimate if we expect our model to be applicable to all two-state folding proteins. We will add a single term to equation (2) above to accomplish this:

$$\ln(k_f) = A + B(N_{res}) + C(N_{helix}), \tag{3}$$

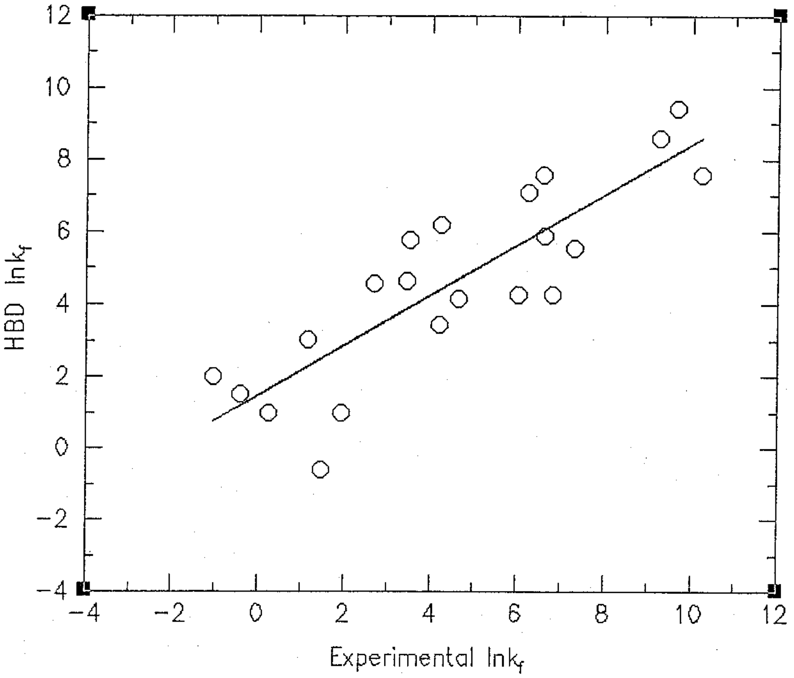
where N_{helix} is the number of helical residues in the protein, C is a positive constant so that the folding rate increases as the proportion of helical residues increases, and N_{res}, A, and B are the same as in equation (2). Based on our studies in chapters 2-5, equation (3) now contains the two most important physical processes that are present in the rate-limiting step of protein folding. Table 6.1 below shows the experimental $\ln(k_f)$ rate data for the 21 two-state folding proteins for which rate data is available. Note that we have omitted four proteins that are present in the relative contact order plot in Figure 6.1. Cytochrome b_{562} was omitted because it contains a heme [6], and apo-myoglobin, villin and spliceosomal protein U1A are omitted here but considered below because they exhibit kinetic "rollover" (slower than expected folding rates at low denaturant concentrations).

PDB	N_{res}	$N_{helical}$	N_{sheet}	Exp. ln(k_f)
2AIT	74	0	30	4.20
1CSP/1MJC	68	0	35	6.04
1TUD/1NYF	65.5	0	23.5	3.45
1PKS	84	0	26	-1.05
1WIT/1TIT	91	0	40	1.94
1FNF/1TEN	91	0	49	0.26
2GB1	56	13	22	6.26
1COA	64	11	14	3.51
1PBA	80	20	9	6.80
1ARR	53	26	0	9.27
1UBQ	76	12	24	4.63
2PTL	62	12	24	4.22
1HDN	85	32	26	2.70
1FKB	107	8	41	1.46
1APS	98	18	36	-0.42
1DIV(Cterm)	93	27	35	1.15
1DIV(Nterm)	56	19	11	6.58
1LMB	80	59	0	10.23
2ABD	86	49	0	6.62
1IMQ	86	45	0	7.31
2PDD	43	19	0	9.68

Table 6.1 Experimental Data for 21, two-state folding proteins.

In Figure 6.3, we plot the HBD predicted ln(k_f) versus experimental ln(k_f) for the 21 two-state folding proteins. The predicted ln(k_f) values predicted by the Helix-Biased-Diffusion model are determined by equation (3), with A= 13.93, B= $-$ 0.14, and C = 0.084. The correlation obtained for the best linear fit is R= 0.87. On average, the experimental folding rates are predicted within a factor of $e^{1.2} \approx 3.3$.

Figure 6.3 HBD predicted ln(k_f) versus experimental ln(k_f) for the 21 two-state folding proteins in Table 6.1.


HBD $\ln k_f$
Experimental $\ln k_f$
12
10
8
6
4
2
0
−2
−4
−4
−2
0
2
4
6
8
10
12

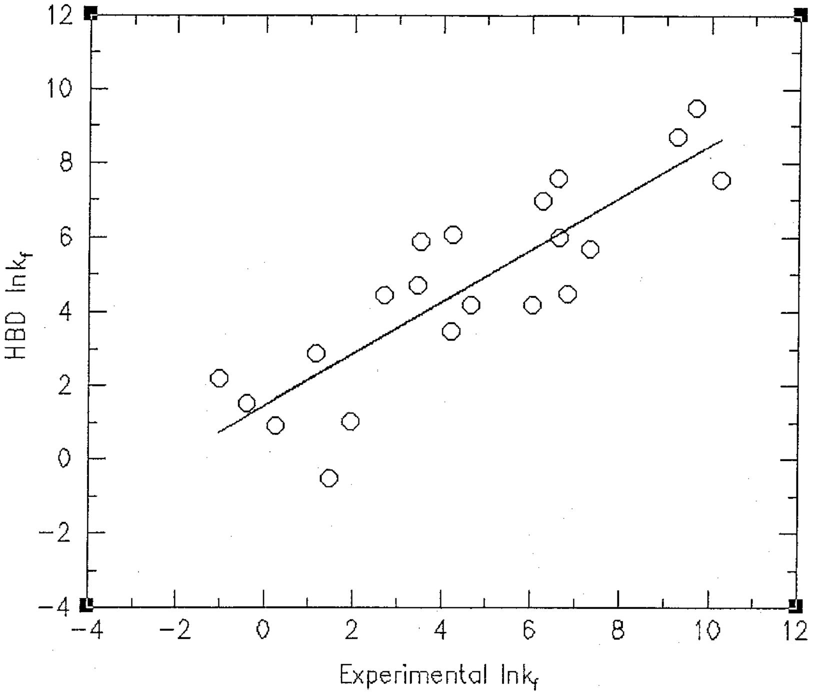
From Figure 6.3, it is evident that the HBD model fits the known experimental data very well. The coefficients A, B, and C in equation 3 behave just as expected. The coefficient B is a negative constant, and C is a positive constant approximately one half the magnitude of B. Thus. In this simple model, the average "freedom" of a helical residue is half of that of a non-helical residue.

Given the success of the HBD model, it is also important to ask whether or not a correction for sheet residues should be included as well. Modifying equation 3 to include the effects of possible β-sheet effects yields:

$$\ln(k_f) = A + B(N_{res})+C(N_{helix})+D(N_{sheet}). \qquad (4)$$

Again, the coefiicients A, B, C, and D can be determined by linear regression analysis. The coeeficients A, B, and C behave as before, with A= 13.84, B= $-$ 0.13, and C = 0.075. The coefficient D is very small in magnitude compared to B and C (D= $-$ 0.015), suggesting that the presence of sheets in the final folded state is not as important for determining of folding kinetics. In Figure 6.4, we plot the $\ln(k_f)$ predicted by equation (4) versus experimental $\ln(k_f)$ for the 21 two-state folding proteins. The correlation for the best line through the points is R= 0.88, similar to what was obtained without the additional degree of freedom, D.

Figure 6.4 HBD+Sheet predicted ln(k_f) versus experimental ln(k_f) for the 21 two-state folding proteins in Table 6.1.


12
10
8
6
4
2
0
−2
−4
−4
−2
0
2
4
6
8
10
12
HBD $\ln k_f$
Experimental $\ln k_f$

Now that we have demonstrated that the physically satisfying HBD model accurately predicts two-state folding rates, it is important to address why the relative contact order model can also seemingly be successfully applied to rate prediction. There are two primary aspects to the HBD model:

(i) folding rate decreases as length increases, and

(ii) folding rate increases with increasing helical content.

The empirical equation for relative contact order accidentally captures both of these elements. First, since relative contact order is equivalent to the average sequential distance between contacts in the final folded protein structure, longer proteins naturally have the possibility of forming longer-range contacts, which steadily drives the relative contact order lower with increasing protein length, despite the presence of protein length in the denominator. Second, the i to i+4 hydrogen bonds in helices naturally reduce the contact order by contributing many terms of $\Delta S_{i,j} = 4$. Because of this, *all* predominantly helical proteins will be predicted to fold very quickly by the relative contact order model.

Despite the accidental similarity between relative contact order and the HBD model, there is one class of protein where the models will produce decidedly different results: long helical proteins. As noted above, the relative contact order model predicts that such proteins will fold quickly. However, if the proteins are long enough, the HBD model predicts that they will eventually begin to fold slower, despite the high helical content. Since most two-state folding proteins are reasonably short, there are no helical proteins in the set that are long enough to fold relatively slowly. Plaxco et al. [4] include deoxymyoglobin is included in their set, but the presence of the heme prevents the experiment from starting from an unfolded state [7]. Recently, Cavagnero and Wright have shown that apomyoglobin folds much more slowly [8], at a rate that is closely determined by the HBD model.

Given the success of the HBD model for predicting two-state folding rates, it is important to ask whether or not the model can predict the folding rates in three-state folding systems. Such systems are often characterized by "roll-over" in their plots of folding rate versus denaturant concentration as the concentration approaches zero. Thus, the two-state extrapolation to zero denaturant concentration in these systems will overestimate the actual folding rate of these systems. However, for many three-state systems, the error due to roll-over is not significant and is less than the average error of the HBD and relative contact order predictions. Thus, we expect that if the large three-state protein are folding by the same overall mechanism as the small, two-state folders, the HBD model should be able to predict their folding rates as well.

In order to accurately predict the folding rate for proteins that exhibit roll-over in their chevron plots, it is necessary to define the concept of a rate-limiting folding unit (RLFU):

> The *rate limiting folding unit* in a protein consists of the regions of sequence expected to compete during the folding process for hydrophobic burial in the native structure's largest hydrophobic core.

Consider the example of a 150-residue protein comprised of two independent domains, where the first domain is 100 residues, and the second domain is only 50 residues. The 100-domain protein will naturally have the larger hydrophobic core. Since the 50 residues in the separate domain are not expected to compete for the larger hydrophobic core, the number of residues competing for this core is expected to be 100 and not 150, the RLFU for this protein is 100 residues.

In order to apply the HBD model to multi-domain proteins and proteins with modest levels of roll-over in their chevron plots, we will replace the total number of residues in the protein N_{res} with the number of residues in the RLFU, N_{RLFU}. Table 2 shows experimental rate data along with the HBD predictions for 16 multi-domain or three-state folding proteins. Figures 6.5 and 6.6 show the native structures of the 12 proteins with an RLFU that is smaller than the entire protein ($N_{RFLU} < N_{res}$).

PDB	N_{RLFU}	$N_{helical}$	N_{sheet}	Exp. $\ln(k_f)$
1URN/1RIS	65	14.5	27.5	5.73
2VIK	82	26	25.0	6.80
1BTA	89	37	16	4.61
1HNG	98	0	54	2.22
1BNI	85	12	29	2.56
1HEL	102	33	8	1.32
1JON	100	42	26	0.83
2RN2	123	41	44	-0.51
1PHPN	109	34	19	-0.65
1PHPC	156	66	25	-3.51
3CHY	129	58	22	0.99
1BKS	113	56	16	1.83
1IET	80	6	6	3.00
1FNF9/10	76	0	42	2.06
1MBC	156	113	0	1.67
1ADW	93	0	44	0.69

Table 6.2 Experimental data for 16 multi-domain and three-state folding proteins who exhibit modest rollover in their chevron plots.

Figure 6.5 The rate-limiting folding units from 6 of the 12 proteins in Table 6.2 with $N_{RLFU} < N_{res}$. (A) 1ADW. (B) 1BKS. (C) 1BNI. (D) 1FNF10. (E) 1HEL. (F) 1IET.

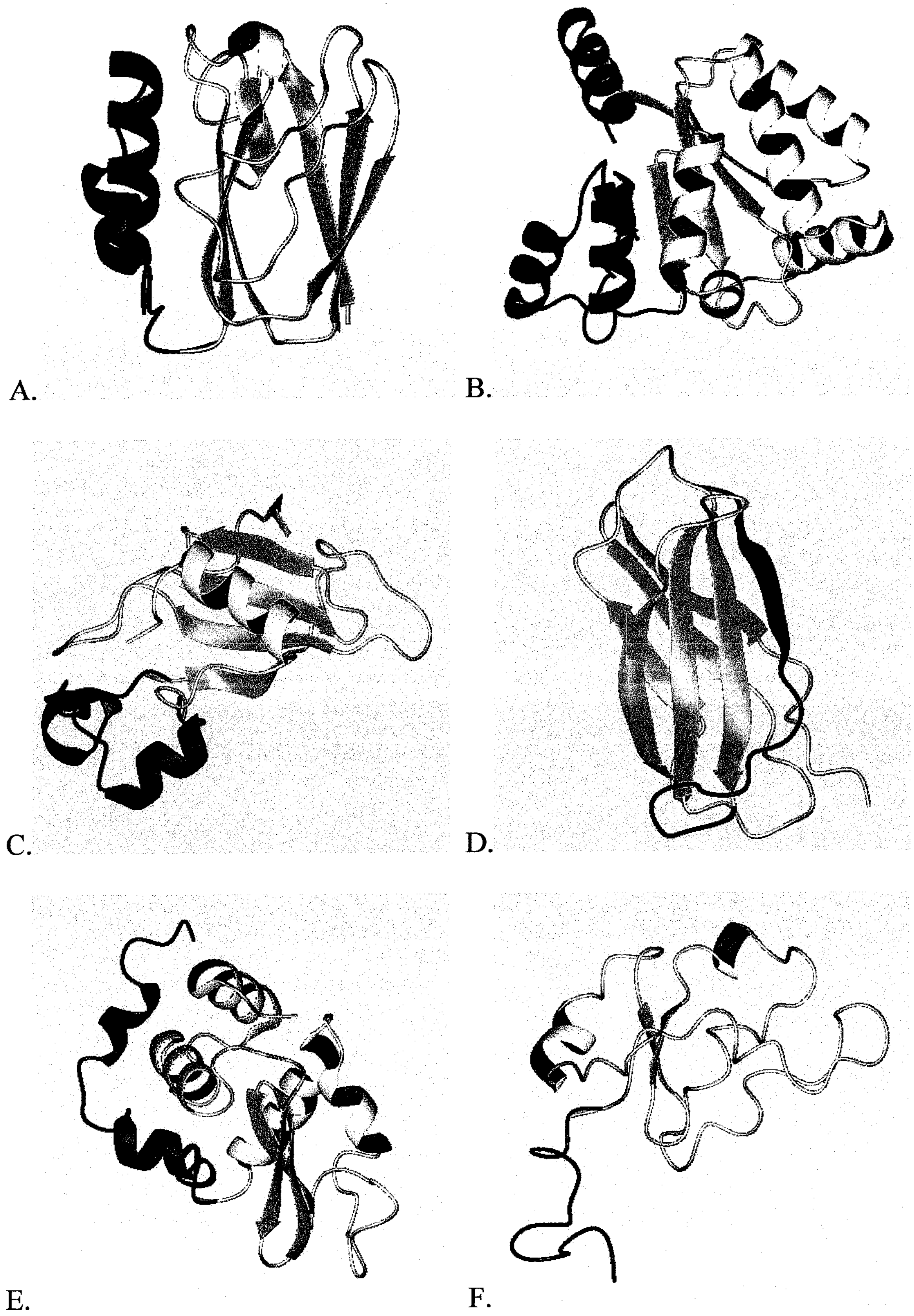
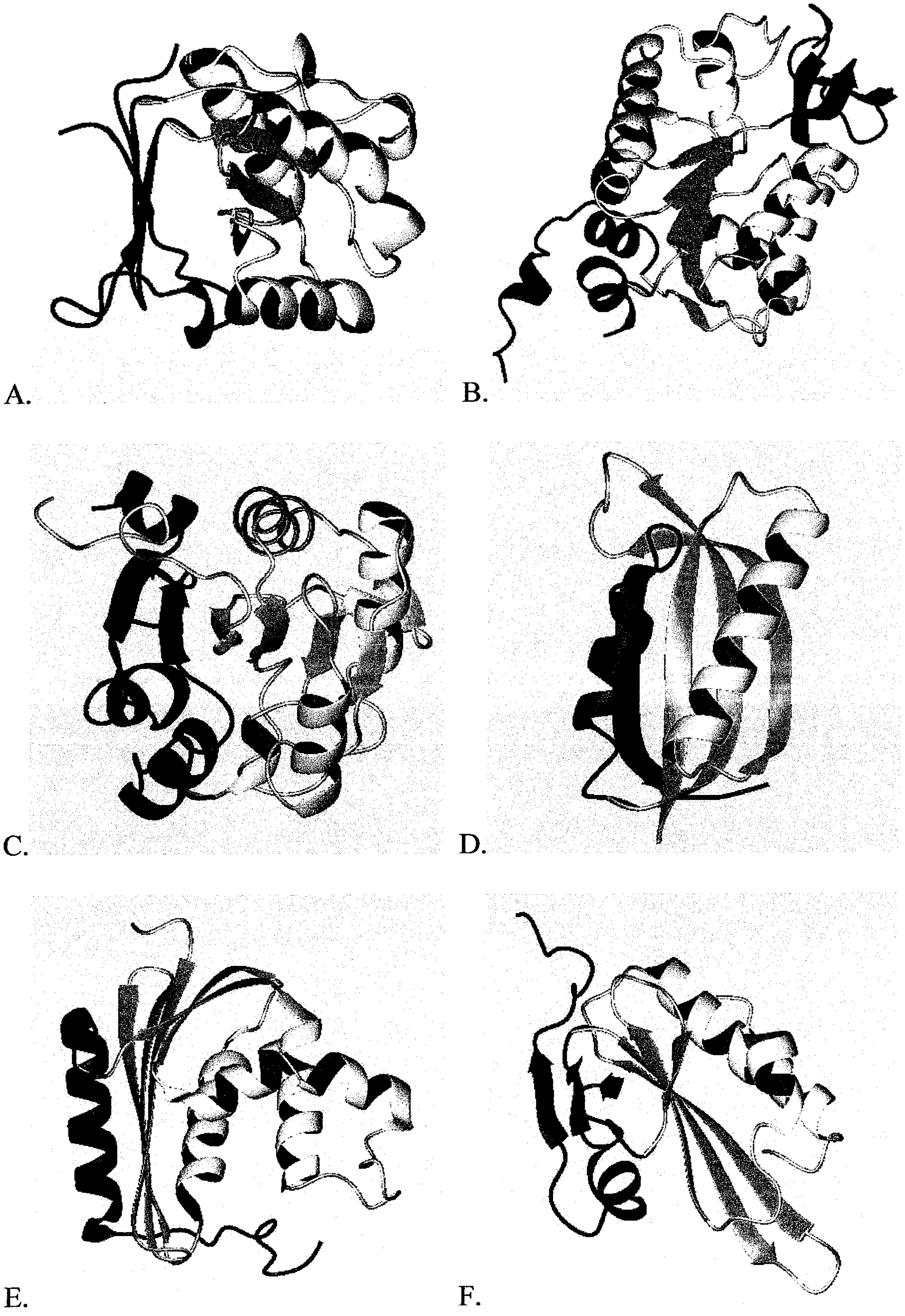
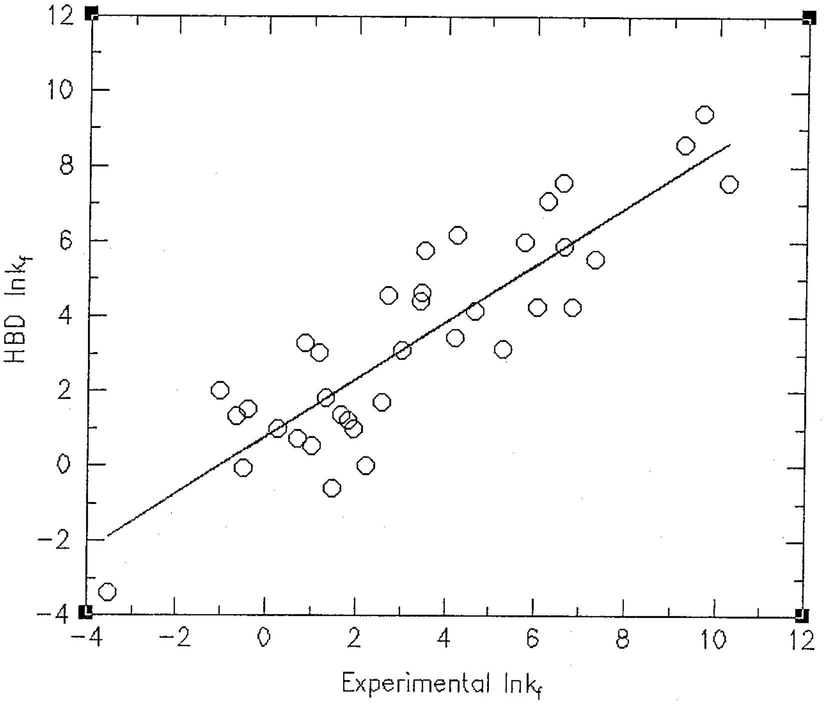

A.
B.
C.
D.
E.
F.

Figure 6.6 The rate-limiting folding units from 6 of the 12 proteins in Table 6.2 with $N_{RLFU} < N_{res}$. (A) 1JON. (B) 1PHPC. (C) 1PHPN. (D) 1RIS. (E) 2RN2. (F) 2VIK.


A.
B.
C.
D.
E.
F.

Now that we have assigned the RLFU's, we can add the 16 new proteins to our previous set of 21 smaller, two-state folding proteins. In Figure 6.7, we plot the HBD predicted ln(k_f) versus experimental ln(k_f) for the entire 37 protein set. The correlation obtained for the best line fit is R= 0.88. These results suggest that the HBD model coupled with the concept of rate-limiting folding units capably predict the folding rates of three-state and multi-domain proteins.

Figure 6.7 HBD predicted ln(k_f) versus experimental ln(k_f) for the 37 two- and three-state folding proteins considered in Tables 6.1 and 6.2.


HBD $\ln k_f$
Experimental $\ln k_f$
12
10
8
6
4
2
0
-2
-4
-4
-2
0
2
4
6
8
10
12

CONCLUSION

We have demonstrated that the Helix-Biased Diffusion model can predict the folding rates for all known two-state folding proteins, as well as three-state folding proteins with only modest rollover in their experimental chevron plots. By introducing the concept of rate-limiting folding units, we have successfully extended the Helix-Biased Diffusion folding model to large, multi-domain proteins. This represents the first successful treatment of folding rates in these larger systems and the first model generalized to include all classes of proteins.

REFERENCES

1. S. E. Jackson and A. R. Fersht, *Biochemistry* **30**, 10428 (1991).

2. S. E. Jackson, *Folding Design* **3**, R81 (1998).

3. K. W. Plaxco, K. T. Simons, & D. Baker, *J. Mol. Biol.* **277**, 985 (1998).

4. K. W. Plaxco, K. T. Simons, I. Ruczinski, and D. Baker, *Biochemistry* **39**, 11177 (2000).

5. A. R. Fersht, *Proc. Natl. Acad. Sci.* **97**, 1525 (2000).

6. P. Wittung-Stafshede, J. C. Lee, J. R. Winkler, and H. B. Gray, *Proc. Natl. Acad. Sci.* **96**, 6587 (1999).

7. Harry B. Gray, personal communication.

8. S. Cavagnero, H. J. Dyson, and P. E. Wright, *J. Mol. Biol.* **285**, 269 (1999).