Heterogeneity of the Folding Mechanism

Testing the Predictions of Free Energy Functional Theory

by

BARIŞ ÖZTOP

B.Sc., Bilkent University, 2002

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Department of Physics and Astronomy)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

July 16, 2004

Library Authorization

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

BARIS OZTOP
Name of Author *(please print)*

19/07/2004
Date (dd/mm/yyyy)

Title of Thesis: Heterogeneity of the Folding Mechanism Testing the Predictions of Free Energy Functional Theory

Degree: Master of Science Year: 2004

Department of Physics and Astronomy
The University of British Columbia
Vancouver, BC Canada

Abstract

The free energy functional theory of protein folding presents a framework to explain the effects of heterogeneity in the folding mechanism. These heterogeneity effects introduce changes in the folding free energy barriers that govern the rates for 2-state folding proteins. Here in this thesis, we focused on checking the validity of the predictions of free energy functional theory by using the data from simulations of C_α, Gō proteins and from experiments. Our results show that folding rates correlate with the degree of heterogeneity in the formation of native contacts for both simulated structures and real proteins.

Contents

List of Tables

List of Figures

Acknowledgements

I want thank Steven S. Plotkin for sharing his incredible knowledge of physics with me and guiding me through the dark dungeons of the protein realm. I am grateful to M. Reza Ejtehadi for always having time to answer my endless questions about computer work and proteins with a huge patience. I also want to thank all those friends in the Physics and Astronomy Department of UBC, who made the tough graduate student life more livable, including Bruno C. Mundim and Ignacio I. Olabarrieta and the nameless friends back in Turkey who didn't let me stay alone even though being on the other side of the planet. Last but obviously not the least thanks go to my family for their endless support.

A Poem

Attila crossed the Danube
Hannibal crossed the Alps
Caesar crossed the Rubicon
and I crossed
my self
burning all the flowers behind me

—Can Yücel

Part I

Thesis

Chapter 1

Introduction

Proteins are polypeptide structures of covalently bonded amino acids, folded into nearly-unique 3-dimensional shape for specific functioning. The unbranched polymer that consists of amino acids before folding is the primary sequence. There are 20 types of amino acids in the nature which are distinct in their side chain groups. Other than this side chain, the remaining structure of amino acids are same for all of them; a central Carbon atom (C_α) attached to a Hydrogen (H), an amino group (NH_2) and a carboxyl group (COOH).

In the cell, hereditary information is stored in 1-dimensional sequence of DNA base pairs [57] and it is transmitted through RNA for protein synthesis in ribosomes. Ribosomes read the instructions from messenger RNA and link the amino acids in the prescribed order, which forms the backbone (primary sequence) of proteins. The information stored in primary sequence has shown to be sufficient for protein to fold into specific 3-dimensional structure without the aid of any cellular machinery [1].

1.1 Driving Interactions

Interactions that drive the folding mechanism are various:

- Hydrogen bond interactions: Attractive intermolecular force between a Hydrogen atom and a strongly electronegative atom (Oxygen, Nitrogen). In proteins it can be between two amino acid atoms or an amino acid atom and a water molecule's atom.

- Hydrophobic interactions: Some amino acids are hydrophobic (non-polar amino acids that are immiscible with water) and some are hydrophilic (polar amino acids that are attracted to water molecules). It is known that these interactions play an important role in the folding process in the formation of the folding nucleus.

- Electrostatic interactions: Some amino acids are electrically charged, so there is Coulomb interactions present depending on the distances, in addition to the hydrogen bond and hydrophobic interactions.

Even by knowing all these interactions that drive the folding process, it is very hard (time wise) even computationally to understand protein folding at the atomic level since proteins are large and complex systems, and it is very hard to keep track of all atoms separately. It is because of this reason, some effective interactions and statistical mechanical models have been introduced which use these interactions to understand some specific aspects protein folding (e.g. rates, barriers, importance of topology, etc.) both theoretically and computationally [40, 41, 42, 44, 45].

Folding can be thought of as a thermodynamic process where the system (1-dimensional polymer chain in solution) searches to find the unique[1] low energy ground state for given amino acid sequence. During folding, protein tends to twist into shapes that achieve a low energy state in which amino acids fit comfortably together (for example hydrophobic amino acids usually cluster in the middle of a protein structure while hydrophilic ones move to the surface). But how does the sequence find the unique stable native state, is it totally a random search in all configuration space?

1.2 Levinthal's Paradox and Funneled Energy Landscape

We can make an estimation for the time a protein needs to find its native state if it searches all possible conformations available. Let's think about a 100 amino acid long protein, and guess roughly that each amino acid in the chain has 10 conformational states to search. So this makes a complete configuration space for the protein spanning 10^{100} total states. Even if the sampling time (amount of time which a residue makes an attempt to find its native state) is assumed to be as small as 10^{-15} seconds, the mean first passage time becomes $(10^{-15}$ sec$)\times 10^{100} = 10^{35}$ sec $\sim 3 \times 10^{77}$ years, which is about 10^{67} times the age of the universe. So, because the protein folding process occurs in physiological time scales on the order of seconds, all conformations available to the protein are not searched [3, 41]. For random heteropolymers (RHP, polymers which have random primary sequences), Levinthal paradox is actually real where the collapse of an RHP can take very long amount of times in comparison to protein folding times. So, one thinks that there should be an evolutionary mechanism which prevents the need to search the whole configurational space for a protein before finding its native state, a biased search.

A theory that resolves the Levinthal paradox for proteins and gives answers to the search problem is energy landscape theory and its prediction: folding funnels. Folding kinetics can be understood in the energy landscape perspective as the organization of an ensemble of partially folded structures (with their associated free energies and entropies), which the protein passes through in the folding process (many routes to folding) [31, 32]. The folding landscape of proteins are thought to be rugged because of the fact that polymers have many available conformations during the process and there is always possibility for residues to form inappropriate contacts (non-native, the ones that are not present in the folded structure) with other residues on the way to the native state. In a simple

[1] It has been shown [13, 14, 51] that physiologically active state is not just this lowest energy one but a number of states which differ at least in side chain orientations.

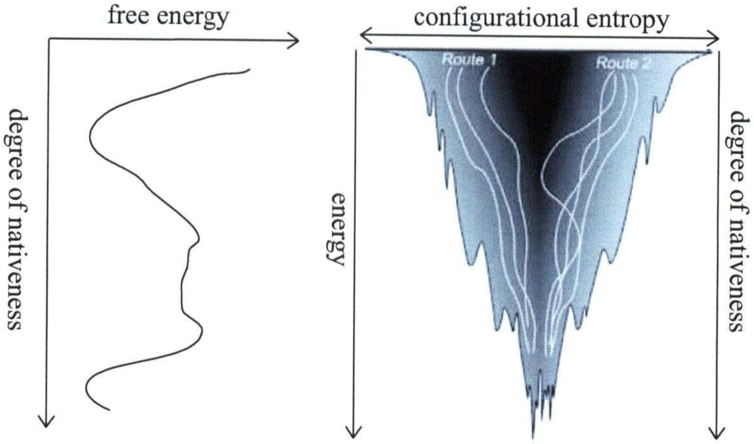

Figure 1.1: Protein folding process is explained as the configurational diffusion on a funnel shaped energy landscape [41]. The depth of the funnel typically represents the energy and the width typically represents the configurational entropy of a conformational state. Usually, there is not a perfect cancellation between energy and the entropy, so when we project the folding process onto a free energy surface, one observes free energy barriers for folding and unfolding, which determine the rate of diffusion. Folding barriers (several T_f) are much smaller than the total binding energy ($\sim 100T_f$) [34].

model of folding, these non-native contacts are usually assumed to have random energetic contributions [31, 40].

Because the native contacts are highly stabilizing interactions, there is an overall slope of the energy landscape, that gives its funneled shape, toward the native structure. In realistic models of folding, proteins are considered to be minimally frustrated polymers. It means that the rugged energy landscape of folding for real proteins is not flat with random fluctuations imposed on it, but has an overall inclination and a preferred direction of flow [31, 32]. The local roughness in the landscape shows the temporary trapping of the configurations in local free energy minima.

Appropriate order parameters are needed to describe the ensemble of partially folded structures, which is another big topic of research [43]. One useful order parameter that is being used in folding literature to describe the position of an ensemble of states in the funnel picture is the fraction of native contacts, Q (some other order parameters that were introduced; fraction of correct dihedral angles in the backbone, fraction of the correct secondary structure, etc.). In our analysis, Q is the appropriate order parameter.

By looking at the funneled energy landscape, one can see that the thermodynamics of folding can be understood as a process where energy and entropy have competing contributions. So when we project this process to a free energy surface as a function of the order parameter, for short, single domain, 2-state folding proteins (which fold without a metastable intermediate), it reduces down to a simple barrier crossing problem where the rate of folding is determined by the free energy barrier and given by the Arrhenius rate law:

$$k_f = k_0 e^{-\Delta F^{\ddagger}/T} \tag{1.1}$$

The rates and free energy barriers for different proteins are different. The question is what factors determine the free energy barrier of proteins? A free energy functional theory has been developed to understand the determinants of folding rates and barriers [39, 40, 49, 50]. In this thesis, our aim is to understand the results of the theory and check the predictions of it by using the data available from both experiments and simulations.

Chapter 2

Free Energy Functional Theory and Its Predictions

In the process of protein folding, both the energetic and the topological factors play important roles [21, 29, 37]. When we say energetics, it means the contact energies of two or more residues that are in proximity at any stage during folding. Topology here means the distribution of contact loop lengths in the protein. In functional approach, energetics is characterized by the distribution of contact energies $\{\epsilon_{ij}\}$, where i and j label the residues in the protein and run from 1 to N (total number of residues). The overall native topology is characterized by the distribution of contact loop lengths $\{\ell_{ij}\} = \{|i-j|\}$. So if there is a native contact (see the definition of contact in Methods section) between the residues i and j, ℓ_{ij} becomes the the length of the loop in terms of number of amino acids and ϵ_{ij} becomes the strength of the interaction. As discussed before, the fraction of native contacts, Q, was chosen to be the order parameter in the theory [40].

$$Q = \frac{1}{M}\sum_{i>j} Q_{ij} \tag{2.1}$$

where M is the total number of native contacts and Q_{ij} is the probability of residue i having a contact with residue j at an overall nativeness during folding[2]. So, given the contact energies and loop length distributions, the free energy can be written as a functional of contact probabilities, $F(\{Q_{ij}(Q)\}|\{\ell_{ij}\},\{\epsilon_{ij}\})$.

2.1 Hamiltonian of the Theory

The Hamiltonian written for the theory to start with is

$$\mathcal{H}(\{\Delta_{ij}\}|\{\Delta_{ij}^N\}) = \sum_{i<j}[\epsilon_{ij}^N \Delta_{ij}\Delta_{ij}^N + \epsilon_{ij}^{nn}\Delta_{ij}(1-\Delta_{ij}^N)] \tag{2.2}$$

where Δ_{ij} (Δ_{ij}^N) is 1 if residues i and j are in contact in a configuration (in native state) and 0 otherwise. ϵ_{ij}^N and ϵ_{ij}^{nn} are the energies of the native and non-native contacts respectively. The goal of the theory is to find the energy functional, and for this purpose one needs to find analytic expressions for thermal energy

[2] In other words, it is the fraction of time the contact between residues i and j is formed at equilibrium in the ensemble with MQ native contacts or fraction of proteins in a macroscopic sample having some degree of nativeness (Q) with the contact between i and j formed [40].

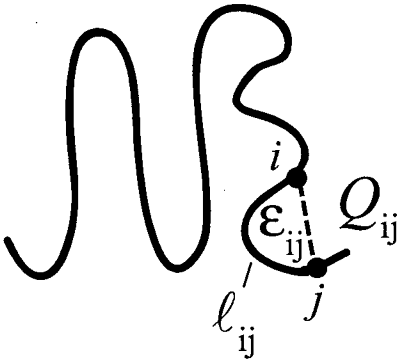

Figure 2.1: Schematic description of the protein's native structure [40]. The native state can be described by the distribution of contact loop lengths $\{\ell_{ij}\}$ and the distribution of contact energies $\{\epsilon_{ij}\}$. Q_{ij} is the probability of contact formation between residues i and j with energy ϵ_{ij} and loop length ℓ_{ij}.

and thermal entropy. The usual way to this is first to calculate the density of states at particular energy E for a given distribution of contact probabilities $\{Q_{ij}\}$, this is $n(E|\{Q_{ij}\})$. And it is equal to the number of states for the specific distribution $\{Q_{ij}\}$, times the probability of having energy E with that distribution, given the native state having a fixed energy E_N:

$$n(E|\{Q_{ij}\}) = \Omega(\{Q_{ij}\})P(E|E_N, \{Q_{ij}\}). \tag{2.3}$$

Conditional probability $P(E|E_N)$ can be written as:

$$P(E|E_N) = \frac{P(E, E_N)}{P(E_N)} \tag{2.4}$$

where the probability of native configuration and configuration Q_{ij} to have energies E_N and E respectively is $P(E, E_N)$ and the probability that native state has energy E_N is $P(E_N)$.

In writing the theory [40], non-native contact energies were considered as an average background field, and taken to be random which in turn gives a Gaussian distribution (with variance b^2) and it is thought to be a good approximation for uncorrelated minimally frustrated energy landscapes [40, 42]. By using this information, the probability $P(E|E_N, \{Q_{ij}\})$ can be calculated by taking the average of the delta functions over the non-native contact energy distribution:

$$P(E|E_N, \{Q_{ij}\}) = \frac{\langle \delta[E - \mathcal{H}(\{\Delta_{ij}\})] \;\; \delta[E_N - \mathcal{H}(\{\Delta_{ij}^N\})] \rangle_{nn}}{\langle \delta[E_N - \mathcal{H}(\{\Delta_{ij}^N\})] \rangle_{nn}}. \tag{2.5}$$

Now one can calculate the thermal energy using $\partial \log n(E)/\partial E = 1/T$. It is

$$E(T|\{Q_{ij}\}) = \overline{E}_{nn} + \sum_{i<j} \epsilon_{ij} Q_{ij} - \frac{Mb^2}{T}(1-Q) \tag{2.6}$$

where $\overline{E}_{nn}$ is the average total non-native energy. The last term in the right hand side of Eq. 2.6 corresponds to decrease in thermal energy due to non-native traps [31, 40] (ruggedness of energy landscape). By using this relation for energy one can calculate the thermal entropy as

$$S(T|\{Q_{ij}\}) = S(\{Q_{ij}\}) - \frac{Mb^2}{2T^2}(1-Q). \tag{2.7}$$

First term in the right hand side of Eq. 2.7 is the entropy of the polymer with the geometric constraints $\{Q_{ij}\}$ and the second term is the decrease in entropy due to non-native traps.

The next step in writing the free energy functional is to find an expression for the geometric entropy term $S(\{Q_{ij}\})$. This term can be written in 3 parts [40]:

$$S(\{Q_{ij}\}) = Ns_0 + S_{ROUTE}(\{Q_{ij}\}) + S_{BOND}(\{Q_{ij}\}|\{\ell_{ij}\}). \tag{2.8}$$

Here s_0 is the entropy per monomer, so Ns_0 becomes the entropy of the unconstrained polymer chain. $S_{ROUTE}(\{Q_{ij}\})$ is the entropy due to the ensemble of states having the same contact formation probability distribution $\{Q_{ij}\}$, so clearly $S_{ROUTE}(\{Q_{ij}\}) > 0$. And finally $S_{BOND}(\{Q_{ij}\}|\{\ell_{ij}\})$ is the configurational entropy loss due to forming contacts, so $S_{BOND}(\{Q_{ij}\}|\{\ell_{ij}\}) < 0$. A detailed analysis and rigorous calculations have been done in [40] to find analytic expressions for these entropy terms. What we are interested in when writing this thesis is not to discuss ways of performing these calculations but to use the results and predictions and check their agreement with experiments and simulations. Because it is not going to give any insight to what we are doing, we will use the expressions taken from [40] and not repeat the calculations here.

By using the expressions for $S_{ROUTE}(\{Q_{ij}\})$, $S_{BOND}(\{Q_{ij}\}|\{\ell_{ij}\})$, Eq. 2.6 and Eq. 2.7, one can write the functional for the free energy barrier height ($F = E - TS$). The result can be written in terms of a perturbative expansion[3] around a mean field term $\Delta F^{\ddagger}_{MF}$ which only depends on mean of the contact energy and loop length distributions ($\overline{\epsilon}$, $\overline{\ell}$). The first order terms are zero since $\sum_{i<j} \delta\epsilon_{ij} = \sum_{i<j}(\epsilon_{ij} - \overline{\epsilon}) = 0$ and $\sum_{i<j} \delta\ell_{ij} = \sum_{i<j}(\ell_{ij} - \overline{\ell}) = 0$. By plugging the appropriate coefficients of the expansion, the free energy barrier becomes:

$$\frac{\Delta F^{\ddagger}}{MT}(\{\epsilon_{ij}\}, \{\ell_{ij}\}) = \frac{\Delta F^{\ddagger}_{MF}}{MT}(\overline{\epsilon}, \overline{\ell}) - \frac{Q^{\ddagger}}{2T^2}\overline{\delta\epsilon^2} - \frac{9}{8}Q^{\ddagger}\frac{\overline{\delta\ell^2}}{\overline{\ell}^2} - \frac{3}{4}\frac{Q^{\ddagger}}{T}\frac{\overline{\delta\ell\delta\epsilon}}{\overline{\ell}}. \tag{2.9}$$

This is the free energy barrier in terms of a mean field term and some fluctuation terms due to the varying contact energies and loop lengths, so they can be written in terms of the variances of corresponding distributions.

[3] This can be done by perturbing the free energy of a homogenous system with $\ell_{ij} = \overline{\ell}$, $\epsilon_{ij} = \overline{\epsilon}$ and $Q_{ij} = Q^{\ddagger}$ ($Q^{\ddagger}$ is the value of Q at the barrier), by taking ℓ_{ij} to $\overline{\ell} + \delta\ell_{ij}$ and ϵ_{ij} to $\overline{\epsilon} + \delta\epsilon_{ij}$ [38].

The second term in the right hand side of the Eq. 2.9 is the correction to the mean field barrier due to variance in the contact energy distribution. As it is clear, this term always decreases the barrier. Teh third term in the right hand side is the fluctuation due to variance in contact loop length distribution. Like the energetic variance, structural variance also decreases the barrier. The barrier can also be lowered by making more likely contacts stronger and shorter contacts more likely. In that case the last term also decreases the barrier (since $\overline{\delta\ell\delta\epsilon}$ becomes positive).

Chapter 3

Results

In this part of the thesis, we present the predictions of the free energy functional theory and see if the results from molecular dynamics simulations and experiments agree with those predictions. [4]

The main question that we are interested in answering is what factors determine the folding free energy barrier for short proteins that fold via 2-state kinetics. It has been shown that one factor is the stability of the folded structure -the barrier decreases as the energetic stability of the folded structure increases [8]. It has also been shown that the native topology is a very important predictor of rates. A topological measure, named relative contact order;

$$RCO = \frac{\overline{\ell}}{N} = \frac{1}{MN}\sum_{i<j}|i-j|\Delta_{ij}^{N} \tag{3.1}$$

was found to be a good predictor of experimental rates that were measured at room temperature in water for 2-state folders [37]. After some time, it was discovered that mean loop length ($\overline{\ell}$) itself is better predictor for both 2 and 3-state (those that fold via a metastable intermediate state) proteins [21].

Here, we first reexamined the trend of experimental rates at the transition midpoint (see Methods) and simulated free energy barriers with $\overline{\ell}$. For this purpose we plotted the log folding rate k_f *vs* $\overline{\ell}$ for a representative set of 20 2-state proteins (See Fig. 3.1A). This graph shows a significant anti-correlation between $\ln(k_f)$ and $\overline{\ell}$. This was clearly an expected result, because one can think that for a protein, during the folding process, it would usually take more time for a long contact to be formed than a shorter contact since the corresponding residues would have to search longer. So if a protein has longer contacts on average, one would expect it to fold slower (or have a larger barrier) than a short contact protein. We can observe the same effect for simulations when we plot the barrier height *vs* $\overline{\ell}$ for 18 structures of known 2-state folders (See Fig. 3.1B). Again we observe a statistically significant correlation between those quantities. However, the fluctuations around the best fit lines of both Fig. 3.1A and Fig. 3.1B tells us that there should be some other factors that affect the barriers and rates.

In the theory section we mentioned that the effects of native topology and energetics can be described analytically by free energy functional theory. It has been shown in Eq. 2.9 that the free energy barrier may be written in terms of

[4] For more information on molecular dynamics simulations and related data see Appendix A and Appendix B, for information about the experimental data see Methods section.

an expansion involving moments of distributions of native contact interaction energies $\{\epsilon_{ij}\}$ and native contact lengths $\{\ell_{ij}\}$. Our earlier discussion on the second order fluctuation terms leads us to the notion that proteins with more heterogeneous folding mechanisms are expected to fold faster, since heterogeneity decreases the free energy barrier. Here heterogeneity refers to variance in contact formation probabilities, loop lengths and contacts energies.

The next step is to check the theory prediction that fluctuations in the contact loop lengths $(\overline{\delta\ell^2}/\overline{\ell}^2)$ really decrease the barrier height. From Eq. 2.9, the change in the barrier due to presence of structural variance is:

$$\frac{(\Delta F^{\ddagger} - \Delta F^{\ddagger}_{MF})}{MT} \equiv \frac{\delta\Delta F^{\ddagger}}{MT} \approx -Q^{\ddagger}\frac{\overline{\delta\ell^2}}{\overline{\ell}^2} \tag{3.2}$$

Even when we ignore variations due to different mean loop lengths $(\overline{\ell})$ of different proteins, the energetic variance term $(\overline{\delta\epsilon^2})$ and the cross term $(\overline{\delta\ell\delta\epsilon})$, there is still an observable effect on barriers and rates. This can be seen from the plots of log experimental folding rate (over M) and simulated free energy barrier height (over MT) *vs* $\overline{\delta\ell^2}/\overline{\ell}^2$ (See Fig. 3.2A and Fig. 3.2B). They both show statistically significant correlations with the measure of structural heterogeneity $(\overline{\delta\ell^2}/\overline{\ell}^2)$, telling us that the free energy barrier of folding decreases with increasing structural heterogeneity. But as one can see, there are large deviations present in the graphs: neglecting the effects of $\overline{\ell}$ and energetic variance, may have introduced some errors.

Using functional theory, one can relate the fluctuations in contact energies and loop lengths to fluctuations in contact formation probabilities. As we discussed before, shorter and more stabilizing contacts are more probable to be formed, so the contact probability distribution Q_{ij} can be written as a function of $\{\epsilon_{ij}\}$ and $\{\ell_{ij}\}$ and the variance in contact formation probabilities can be written in terms of $\overline{\delta\epsilon^2}$ and $\overline{\delta\ell^2}$. If we rewrite the change in the barrier in terms of $\overline{\delta Q^2} \equiv (1/M)\sum_{i<j}(Q_{ij} - \overline{Q})^2$ by using the appropriate relations [40], it becomes:

$$\frac{\delta\Delta F^{\ddagger}}{MT} \approx -\frac{1}{2Q^{\ddagger}}\overline{\delta Q^2}. \tag{3.3}$$

Here neither Q_{ij} nor the variance $\overline{\delta Q^2}$ is a practical quantity to extract from folding experiments. Rather, a more practical quantity named ϕ-value (see Methods) is easier to determine and very closely related to the Q-values. Since ϕ-value, like Q-value, is a measure of both energetics and entropics (topology) for a residue, it should better capture the effects of heterogeneity in folding mechanism. We can estimate the variance in Q-values in terms of variance in ϕ-values $(\overline{\delta\phi^2} \equiv (1/N)\sum_i(\phi_i - \overline{\phi})^2)$ in the approximation that all contacts are fully formed in the native structure $(Q^F = 1)$ and unformed in the unfolded structures $(Q^U = 0)$. The approximate relation is

$$\overline{\delta\phi^2} \approx \frac{1}{z}\overline{\delta Q^2} \tag{3.4}$$

where z is the average number of contacts per residue. Eq. 3.4 tells that the variance in ϕ-values is proportional to the variance in Q-values up to a proportionality constant of order unity. We checked the validity of this approximation from the simulations, since both Q_{ij} and ϕ_i values are available, by plotting $\overline{\delta\phi^2}$ against $\overline{\delta Q^2}$. As we can see from Fig. 3.3, they correlate extremely well and this result shows the validity of Eq. 3.4 except for the proportionality factor $1/z$. z is typically ~ 5 for the proteins used in our analysis. According to Eq. 3.4, one expects the slope of Fig. 3.3 to be ~ 0.2, which is not the case. The reason for that is when we use the exact relation between the ϕ-values and the Q-values (see Eq. A.7 in Appendix A), there are some fluctuating quantities from one protein to another (like Q_U and Q_F) and z is also different for different proteins, which may change the value of the proportionality constant.

Now we can write the total change in the barrier height due to both energetic and contact loop length fluctuations in terms of ϕ variance:

$$\frac{\delta\Delta F^{\ddagger}}{MT} \approx -\frac{z}{2Q^{\ddagger}}\overline{\delta\phi^2}. \tag{3.5}$$

This equation tells us that the free energy barrier of folding should be smaller for proteins with more polarized folding nucleus (larger variance in ϕ-values). To check this, we used ϕ-value data extracted from experiments for 12 proteins and plotted it against the log of experimental folding rates (over M) at transition midpoints (Fig. 3.4A). The graph shows a statistically significant correlation which is what theory predicted about the change in the barrier (and so the rates) with $\overline{\delta\phi^2}$. Furthermore we plotted the simulated folding barriers (over MT) against the variance in ϕ-values extracted from simulations (Fig. 3.4B). What we observe is again strong, statistically significant correlation, telling that the barrier go down with increasing heterogeneity.

We plotted the whole barrier over MT against the variances but not the change, because the mean field barrier is not a measurable quantity from the experiments or simulations. The quantity $\delta\Delta F^{\ddagger}/MT$ is actually the residual barrier after subtracting out the mean field barrier (which only depends on the mean loop length $\bar{\ell}$ and the mean of the contact energies $\bar{\epsilon}$). One way to approximate this residual barrier could be to subtract the effects of $\bar{\ell}$ (since we don't know $\bar{\epsilon}$ for experimental proteins and it is 0 for simulated structures). Looking at the correlations of the residuals of $-\Delta F^{\ddagger}/MT$ *vs* $\bar{\ell}$ with $\overline{\delta\phi^2}$ and $\overline{\delta\ell^2}$, the results are comparable (statistical significance within 10%).

For simulations, there is a strong and statistically significant correlation between $\overline{\delta\phi^2}$ and $\overline{\delta\ell^2}/\bar{\ell}^2$ (Table 3.1) as one expects. It is because in our simulation models all the contact energies are same, so the energetic variance is 0. This means that the second and fourth terms on the right hand side of Eq. 2.9 vanish and there remains only the term due to fluctuations in the contact loop length which is a determinant of the barrier by itself like the variance in the ϕ-values. However for experiments, we didn't observe any significant correlation between these two quantities (Table 3.1). This tells that there may be variance present in the native contact energies of real proteins. This is also the reason why we do not see any significant correlation between the variances of experimental and simulated ϕ-values (Table 3.1).

In testing the theory we divided the barrier by the total number of contacts M and plotted it against variances. We want to note that the total number of contacts increases linearly with the chain length (N) for the proteins used for our analysis, which can be seen by looking at the extremely good correlation between them (Table 3.1). One might divide the barrier by the chain length instead of number of contacts and still observe statistically significant correlations with the structural and energectic variances (Table 3.1).

Data for wild type and P^{13-14} circular permutant of protein S6 was not used in Fig. 3.2A, because both the wild type and the permutant have significantly correlated contact energies and loop lengths. For the wild type, longer contacts are stronger whereas the circular permutant was engineered to have stronger short contacts [26]. So, the effect of structural heterogeneity $(\overline{\delta\ell^2}/\overline{\ell}^2)$ is not enough to explain the change in the barrier and this is why using the variance in ϕ-values is more accurate and significant (If we use those 2 data points in Fig. 3.2A, the correlation becomes $r = 0.57$ and $P(r) = 9.6 \times 10^{-3}$, which is still significant).

Table 3.1: Correlation coefficient and statistical significance for various quantities.

y	vs	x	r	$P(r)$[5]	τ	$P(\tau)$[5]
$\ln(k_f)$		$\bar{\ell}$	-0.69	9×10^{-4}	-0.46	5.3×10^{-3}
$-\Delta F^{\ddagger}_{sim}/T_f$		$\bar{\ell}$	-0.71	10^{-3}	-0.61	4.0×10^{-4}
$\ln(k_f)/M$		$\overline{\delta\phi^2_{exp}}$	0.78	2.8×10^{-3}	0.52	2.0×10^{-2}
$-\Delta F^{\ddagger}_{sim}/MT_f$		$\overline{\delta\phi^2_{sim}}$	0.67	2.3×10^{-3}	0.47	7.2×10^{-3}
$\ln(k_f)/M$		$\overline{\delta\ell^2}/\bar{\ell}^2$	0.62	6.6×10^{-3}	0.48	5.7×10^{-3}
$-\Delta F^{\ddagger}_{sim}/MT_f$		$\overline{\delta\ell^2}/\bar{\ell}^2$	0.53	2.7×10^{-2}	0.36	3.7×10^{-2}
M		N[6]	0.94	$< 10^{-6}$	0.84	$< 10^{-6}$
$\ln(k_f)/N$		$\overline{\delta\phi^2_{exp}}$	0.78	2.6×10^{-3}	0.49	2.8×10^{-2}
$-\Delta F^{\ddagger}_{sim}/NT_f$		$\overline{\delta\phi^2_{sim}}$	0.59	1.0×10^{-2}	0.40	2.1×10^{-2}
$\ln(k_f)/N$		$\overline{\delta\ell^2}/\bar{\ell}^2$	0.63	5.7×10^{-3}	0.54	1.7×10^{-3}
$-\Delta F^{\ddagger}_{sim}/NT_f$		$\overline{\delta\ell^2}/\bar{\ell}^2$	0.56	1.8×10^{-2}	0.40	2.1×10^{-2}
$\bar{\ell}$		$\overline{\delta\ell^2}/\bar{\ell}^2$[6]	-0.14	0.52	-0.07	0.70
$\bar{\ell}$		$\overline{\delta\phi^2_{exp}}$	-0.64	2.5×10^{-2}	-0.43	5.5×10^{-2}
$\bar{\ell}$		$\overline{\delta\phi^2_{sim}}$	0.16	0.52	0.15	0.38
$\overline{\delta\phi^2_{sim}}$		$\overline{\delta\ell^2}/\bar{\ell}^2$	0.71	1.0×10^{-3}	0.32	6.4×10^{-2}
$\overline{\delta\phi^2_{exp}}$		$\overline{\delta\ell^2}/\bar{\ell}^2$	0.29	0.37	0.18	0.41
$\overline{\delta\phi^2_{exp}}$		$\overline{\delta\phi^2_{sim}}$	-0.16	0.80	0.20	0.63
$\overline{\delta\phi^2_{sim}}$		$\overline{\delta Q^2_{sim}}$	0.94	$< 10^{-6}$	0.77	9.0×10^{-6}

[5] 2-sided statistical significance has been used.

[6] Data from both simulated and experimental proteins used.

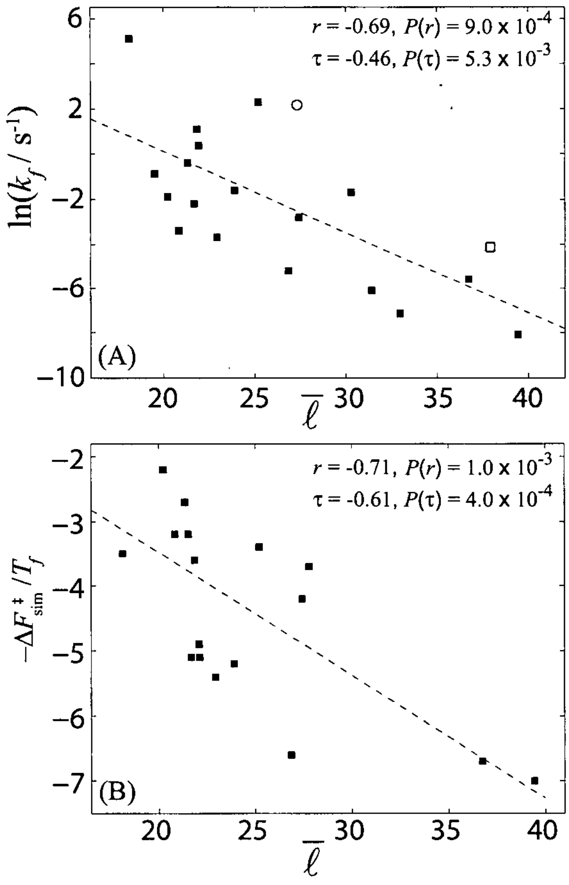

Figure 3.1: (A) Log of experimental folding rates (in sec^{-1}) at the proteins' transition midpoints *vs* the mean loop length ($\overline{\ell}$). Wild type protein S6 is shown by an open square and P^{13-14} circular permutant of S6 (formed by linking the ends of the protein and cutting the covalent bond between residues 13 and 14, so there is a new distribution of contact lengths) is shown by an open circle [26]. (B) Equivalent measure for logarithm of rates in simulations is $-\Delta F^{\ddagger}_{sim}/T_f$ plotted *vs* $\overline{\ell}$. Both graphs show statistically significant anti-correlations. As a measure of statistical correlation, linear correlation coefficient r and Kendall's τ have been used. Statistical significance is given by the corresponding probabilities to observe a given correlation coefficient or greater by chance. If the probability values associated with a correlation coefficient is smaller than 5%, the correlation is thought to be significant [56]

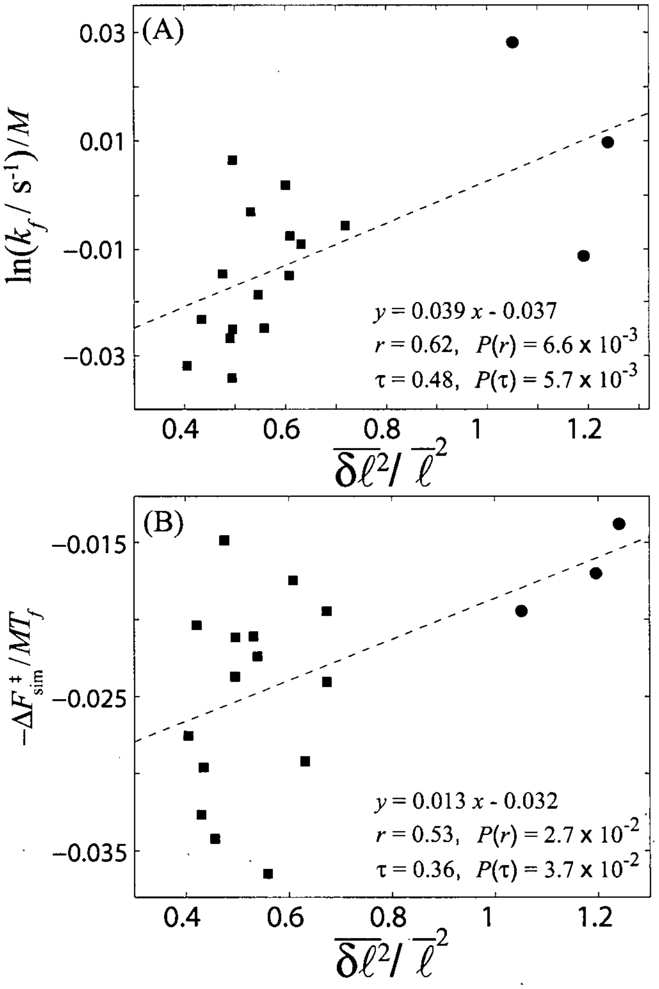

Figure 3.2: (A) Logarithm of the experimental rate data (over M, at the transition midpoints) is plotted against the structural dispersion $(\overline{\delta\ell^2}/\overline{\ell}^2)$. (B) Simulated barrier heights (over MT) at the proteins' folding temperatures T_f is plotted against $\overline{\delta\ell^2}/\overline{\ell}^2$. Both plots show statistically significant correlations. In the graphs 3 outliers with large $\overline{\delta\ell^2}/\overline{\ell}^2$ (shown by filled circles) are α/β proteins (λ-repressor chain 3, cytochrome c, yeast iso-1-cytochrome c) which both have large variance in contact length distributions and relatively fast folding rates.

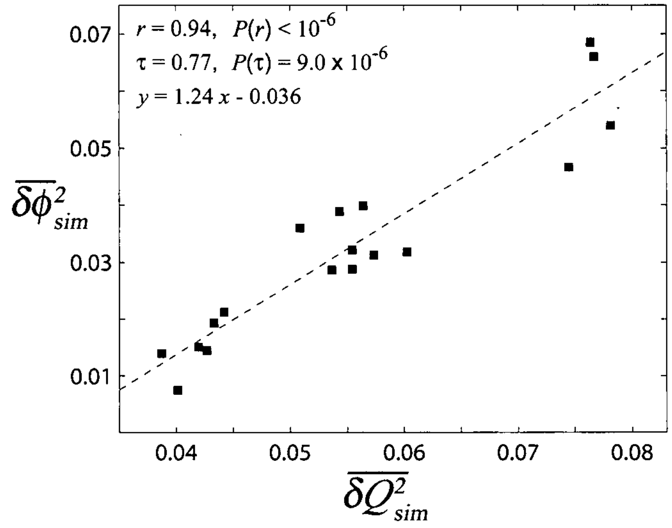

Figure 3.3: Variance in ϕ-values for 18 simulated proteins shows a very strong and statistically significant correlation as expected with the variance in Q-values which were also extracted from the simulations. So that we can safely recast the change in the barrier in terms of $\overline{\delta\phi^2}$.

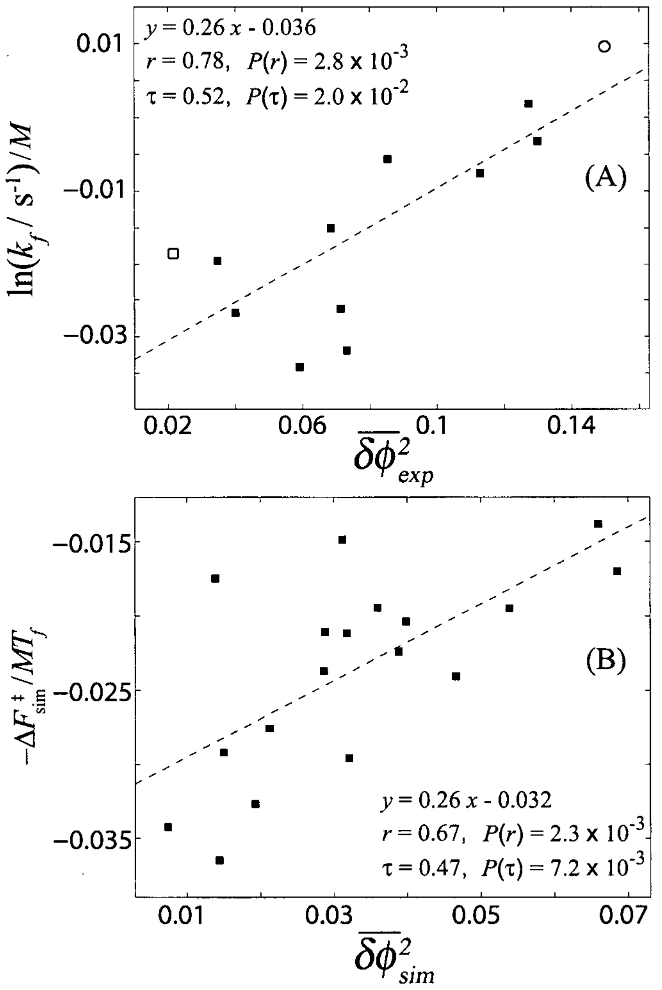

Figure 3.4: (A) Logarithm of experimental folding rate (over M) is plotted against the variance in experimentally measured ϕ-values for a subset of the proteins in Fig. 3.1. Wild type protein S6 is again marked by an open square and P^{13-14} circular permutant of S6 is shown by an open circle [26]. (B) Minus simulated free energy barrier height (over MT) for 18 proteins is plotted against the variance in simulated ϕ-values. Both graphs show strong, statistically significant correlations. Especially, despite the fact that the number of data points in (A) is small, it is important to note the strong correlation.

Chapter 4

Conclusions and Future Prospects

In this thesis, we aimed to understand the results and the predictions of the free energy functional theory [40] on determinants of folding rates and corresponding free energy barriers for proteins that fold via a 2-state mechanism and check the validity of these predictions. To this end, we used the data available from folding experiments and from our own molecular dynamics simulations.

We started by checking earlier results [21] of the dependence of rates on the mean loop length ($\bar{\ell}$) by using our data. Results showed us that there are significant correlations between the mean loop length and the experimental and simulated free energy barriers of proteins (Fig. 3.1A and B). Proteins with longer loop lengths have larger barriers and as a result, smaller rates. Free energy functional theory tells that apart from the dependence on mean loop length, heterogeneity present in the folding mechanism can effectively reduce the free energy barrier and speed up the folding process. This heterogeneity can be thought as the non-uniform ordering of the contacts, where shorter and more stabilizing contacts are more probable to be formed. So, one can talk about non-zero variances in the contact loop length ($\{\ell_{ij}\}$), contact energy ($\{\epsilon_{ij}\}$) and contact formation probability ($\{Q_{ij}\}$) distributions. By using mean field approach, an expansion of the free energy barrier can be written around the uniform folding scenario in terms of contact energy and loop length variances which was predicted to decrease the barrier. In order to see this, first we plotted the simulated and experimental barriers against the structural variance term ($\overline{\delta\ell^2}/\bar{\ell}^2$) in the expansion by using the available data (Fig. 3.2A and B). Statistically significant correlations tells that the structural heterogeneity indeed reduces the barrier. But it is not the end of the story since the result did not capture the whole heterogeneity present in the proteins but only the one due to loop length distribution. Total heterogeneity (structural and energetic) can be written as the variance in contact formation probabilities ($\overline{\delta Q^2}$). This quantity is not practical to extract from experiments, but it is very closely related to another quantity, variance in ϕ-values, which can be obtained both from experiments and simulations. So, we showed that the barrier can be written in terms of $\overline{\delta\phi^2}$. We plotted the experimental and simulated barriers against $\overline{\delta\phi^2}$ and observed that proteins with more polarized nucleus (larger ϕ variance) have smaller free energy barriers as theory predicted.

The free energy functional theory is able to capture the overall effects on the folding barriers. However, one can go further in the analysis by including the many body effects, which were shown to be present in some proteins [9]. This may introduce some corrections and increase the accuracy of the predictions of the theory. It may also be extended to include some predictions for 3-state folding proteins which have different determinants for their rates (like the chain length N).

One of the problems that we have encountered during this work was the fact that experimental ϕ-values available for 2-state proteins are very limited for this kind of statistical analysis. These effects could be observed better with more data points. Correlations and the significance values might be more accurate if an analysis will be done in the future by using a larger set of proteins.

Chapter 5

Methods

5.1 Experimental Rates

Experimental rates for 20 proteins were taken from different articles [4, 6, 12, 16, 18, 19, 22, 23, 24, 26, 27, 28, 30, 36, 46, 47, 48, 54, 55]. Instead of rates at room temperature in water, the rates at proteins' transition midpoints (where the stability of the folded and unfolded states are equal, at various denaturant concentrations) were used in plots and calculations. This was done to eliminate the effects due to the presence of different stabilities for different proteins and to make a consistent analysis together with the results from simulations where all proteins are at their folding temperature (T_f is the temperature at which the folded and unfolded structures are at the same stability).

5.2 Experimental ϕ-values

ϕ-value is a measure that determines the structure of residues and their close proximity in the transition state. Since the knowledge of the transition state structures is very important for understanding the protein folding process, ϕ-value is a very useful quantity to examine. For 2-state proteins, experimental ϕ-values are measured by mutations. A point mutation (changing a particular amino acid type) is done for a residue, than the change in the folding barrier ($\Delta\Delta G_{\ddagger-U}$) and the change in the stability of the folded structure ($\Delta\Delta G_{F-U}$) are measured. ϕ-value for that residue is defined as:

$$\phi \equiv \frac{\Delta\Delta G_{\ddagger-U}}{\Delta\Delta G_{F-U}} \tag{5.1}$$

When the mutation can be considered as a small perturbation, ϕ-value can be accepted as a good measure of the fraction of native structure formed in the transition state ensemble for the mutated part. A ϕ-value close to 1 means that the free energy change in the transition and the native state are very close to each other for the mutant and the wild type protein. And this indicates that the native contacts for the mutated residue are already formed in the transition state. On the other hand, a ϕ-value close to 0 means that the free energy change in the transition state and the unfolded state are very close to each other, so that it looks more like unfolded in the transition state around the mutated residue.

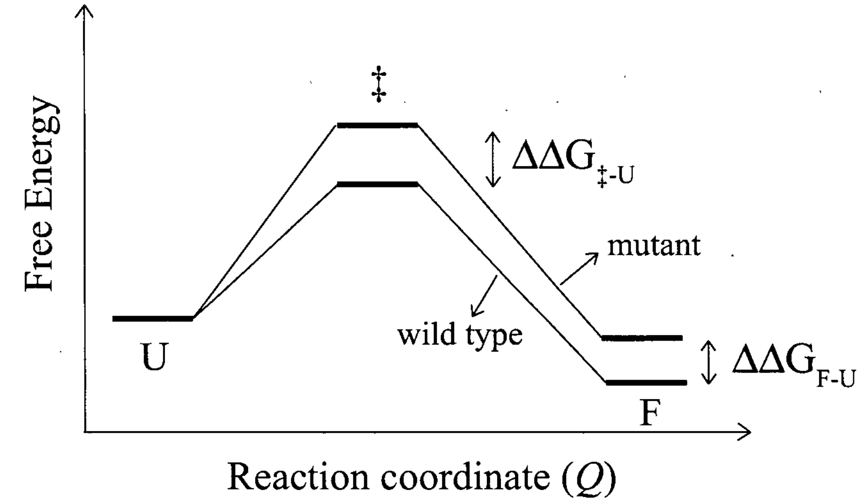
Free Energy
‡
$\Delta\Delta G_{‡\text{-}U}$
mutant
wild type
U
F
$\Delta\Delta G_{F\text{-}U}$
Reaction coordinate (Q)

Figure 5.1: Figure shows the case of a protein which folds via 2-state mechanism. A mutation causes a change in the stability of the folded state (F) with respect to the unfolded state (U), $\Delta\Delta G_{F-U}$, and a change in the free energy of the transition state (‡) with respect to unfolded state, $\Delta\Delta G_{‡-U}$. ϕ-value, the ratio of these two free energy changes, depends on the amount of structure that has been formed in the transition state around the position of mutation.

Experimental ϕ-value Data

Data for experimental ϕ values were taken from [5, 6, 12, 15, 16, 18, 20, 24, 26, 28, 47].

5.3 Topological Quantities

Topological quantities for all proteins (simulated and experimental) were calculated by using corresponding Protein Databank (PDB) entries. PDB files have structural information about the folded proteins, including the amino acid types and order in the primary sequence and the coordinates for all the atoms in the folded structure. In calculating the topological measures ($\bar{\ell}$ and $\overline{\delta\ell^2}$), a contact between two residues has been taken to be formed if in the native structure either heavy side chain atoms or C_α atoms of two amino acids are within a cut-off distance of 4.8 Å. So, mean loop length was calculated by using:

$$\bar{\ell} \equiv \frac{1}{M} \sum_{i<j} |i-j| \Delta_{ij}^{N} \tag{5.2}$$

where

$$\Delta_{ij}^{N} = \begin{cases} 1 & \text{if residues } i \text{ and } j \text{ are in contact in native state} \\ 0 & \text{otherwise} \end{cases}$$

And the structural dispersion was calculated by using:

$$\overline{\delta\ell^2} \equiv \frac{1}{M} \sum_{i<j} (\ell_{ij} - \overline{\ell})^2 \Delta_{ij}^{N}. \tag{5.3}$$

PDB Codes of Experimental Proteins

PDB entries for 19 experimental proteins are: 1AEY, 1APS, 1BF4, 1FKB, 1HRC, 1LMB, 1MJC, 1NYF, 1PGB, 1RIS, 1SRL, 1TEN, 1TIT, 1UBQ, 1YCC, 2AIT, 2CI2, 2PTL, 2VIK. In calculation of topological quantities for P^{13-14} circular permutant of protein S6, the PDB entry 1RIS has been modified.

Bibliography

[1] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223, 1973.

[2] H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[3] H. Bohr and S. Brunak. *Protein Folds: A Distance Based Approach.* CRC Press Inc., Boca Raton; Florida, 1996.

[4] C. K. Chan, Y. Hu, S. Takahashi, D. L. Rousseau, W. A. Eaton, and J. Hofrichter. Submillisecond protein folding kinetics studied by ultrarapid mixing. *Proc. Natl. Acad. Sci. U. S. A.*, 94:1779–1784, 1997.

[5] F. Chiti, N. Taddei, P. M. White, M. Bucciantini, F. Magherini, M. Stefani, and C. M. Dobson. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.*, 6:1005–1009, 1999.

[6] S. E. Choe, L. W. Li, P. T. Matsudaira, G. Wagner, and E. I. Shakhnovich. Differential stabilization of two hydrophobic cores in the transition state of the villin 14T folding reaction. *J. Mol. Biol.*, 304:99–115, 2000.

[7] C. Clementi, H. Nymeyer, and J. N. Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and en-route intermediates for protein folding? an investigation for small globular proteins. *J. Mol. Biol.*, 298:937–953, 2000.

[8] A. R. Dinner and M. Karplus. The roles of stability and contact order in determining protein folding rates. *Nat. Struct. Biol.*, 8:21–22, 2001.

[9] M. R. Ejtehadi, S. P. Avall, and S. S. Plotkin. Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proc. Natl. Acad. Sci. U. S. A. submitted.*

[10] A. M. Ferrenberg and R. H. Swendsen. New monte-carlo technique for studying phase-transitions. *Phys. Rev. Lett.*, 61:2635–2638, 1988.

[11] A. M. Ferrenberg and R. H. Swendsen. Optimized monte-carlo data-analysis. *Phys. Rev. Lett.*, 63:1195–1198, 1989.

[12] S. B. Fowler and J. Clarke. Mapping the folding pathway of an immunoglobulin domain: Structural detail from phi value analysis and movement of the transition state. *Structure*, 9:355–366, 2001.

[13] H. Frauenfelder, F. Parak, and R. D. Young. Conformational substates in proteins. *Ann. Rev. Biophys. Biophys. Chem.*, 17:451–479, 1988.

[14] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254:1598–1603, 1991.

[15] K. F. Fulton, E. R. G. Main, V. Daggett, and S. E. Jackson. Mapping the interactions present in the transition state for unfolding/folding of fkbp12. *J. Mol. Biol.*, 291:445–461, 1999.

[16] R. Guerois and L. Serrano. The SH3-fold family: Experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.*, 304:967–982, 2000.

[17] Zhuyan Guo, D. Thirumalai, and J. D. Honeycutt. Folding kinetics of proteins: a model study. *J. Chem. Phys.*, 97(1):525–535, 1992.

[18] S. J. Hamill, A. Steward, and J. Clarke. The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.*, 297:165–178, 2000.

[19] G. S. Huang and T. G. Oas. Submillisecond folding of monomeric lambda-repressor. *Proc. Natl. Acad. Sci. U. S. A.*, 92:6878–6882, 1995.

[20] L. S. Itzhaki, D. E. Otzen, and A. R. Fersht. 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.*, 254:260–288, 1995.

[21] D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker, and A. V. Finkelstein. Contact order revisited: Influence of protein size on the folding rate. *Protein Sci.*, 12:2057–2062, 2003.

[22] S. E. Jackson and A. R. Fersht. Folding of chymotrypsin inhibitor-2 .1. evidence for a 2-state transition. *Biochemistry*, 30:10428–10435, 1991.

[23] S. Khorasanizadeh, I. D. Peters, T. R. Butt, and H. Roder. Folding and stability of a tryptophan-containing mutant of ubiquitin. *Biochemistry*, 32:7054–7063, 1993.

[24] D. E. Kim, C. Fisher, and D. Baker. A breakdown of symmetry in the folding transition state of protein l. *J. Mol. Biol.*, 298:971–984, 2000.

[25] N. Koga and S. Takada. Roles of native topology and chain-length scaling in protein folding: A simulation study with a Go-like model. *J. Mol. Biol.*, 313:171–180, 2001.

[26] M. Lindberg, J. Tangrot, and M. Oliveberg. Complete change of the protein folding transition state upon circular permutation. *Nat. Struct. Biol.*, 9:818–822, 2002.

[27] E. R. G. Main, K. F. Fulton, and S. E. Jackson. Folding pathway of FKBP12 and characterisation of the transition state. *J. Mol. Biol.*, 291:429–444, 1999.

[28] E. L. McCallister, E. Alm, and D. Baker. Critical role of beta-hairpin formation in protein G folding. *Nat. Struct. Biol.*, 7:669–673, 2000.

[29] C. Micheletti. Prediction of folding rates and transition-state placement from native-state geometry. *Proteins*, 51:74–84, 2003.

[30] G. A. Mines, T. Pascher, S. C. Lee, J. R. Winkler, and H. B. Gray. Cytochrome c folding triggered by electron transfer. *Chem. Biol.*, 3:491–497, 1996.

[31] J. N. Onuchic, Z. LutheySchulten, and P. G. Wolynes. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.*, 48:545–600, 1997.

[32] J. N. Onuchic, H. Nymeyer, A. E. Garcia, J. Chahine, and N. D. Socci. The energy landscape theory of protein folding: Insights into folding mechanisms and scenarios. *Adv. Protein Chem.*, 53:87–152, 2000.

[33] J. N. Onuchic, N. D. Socci, Z. LutheySchulten, and P. G. Wolynes. Protein folding funnels: The nature of the transition state ensemble. *Folding & Design*, 1:441–450, 1996.

[34] J. N. Onuchic and P. G. Wolynes. Theory of protein folding. *Curr. Opin. Struct. Biol.*, 14:70–75, 2004.

[35] B. Öztop, M. R. Ejtehadi, and S. S. Plotkin. Protein folding rates correlate with heterogeneity of folding mechanism. *Preprint.*

[36] K. W. Plaxco, J. I. Guijarro, C. J. Morton, M. Pitkeathly, I. D. Campbell, and C. M. Dobson. The folding kinetics and thermodynamics of the Fyn-SH3 domain. *Biochemistry*, 37:2529–2537, 1998.

[37] K. W. Plaxco, K. T. Simons, I. Ruczinski, and B. David. Topology, stability, sequence, and length: Defining the determinants of two-state protein folding kinetics. *Biochemistry*, 39:11177–11183, 2000.

[38] S. S. Plotkin and J. N. Onuchic. Structural and energetic heterogeneity in protein folding. *Preprint.*

[39] S. S. Plotkin and J. N. Onuchic. Investigation of routes and funnels in protein folding by free energy functional methods. *Proc. Natl. Acad. Sci. U. S. A.*, 97:6509–6514, 2000.

[40] S. S. Plotkin and J. N. Onuchic. Structural and energetic heterogeneity in protein folding. I. theory. *J. Chem. Phys.*, 116:5263–5283, 2002.

[41] S. S. Plotkin and J. N. Onuchic. Understanding protein folding with energy landscape theory - Part I: Basic concepts. *Q. Rev. Biophys.*, 35:111–167, 2002.

[42] S. S. Plotkin and J. N. Onuchic. Understanding protein folding with energy landscape theory - Part II: Quantitative aspects. *Q. Rev. Biophys.*, 35:205–286, 2002.

[43] S. S. Plotkin and P. G. Wolynes. Non-markovian configurational diffusion and reaction coordinates for protein folding. *Phys. Rev. Lett.*, 80:5015–5018, 1998.

[44] J. J. Portman, S. Takada, and P. G. Wolynes. Microscopic theory of protein folding rates. I. Fine structure of the free energy profile and folding routes from a variational approach. *J. Chem. Phys.*, 114:5069–5081, 2001.

[45] J. J. Portman, S. Takada, and P. G. Wolynes. Microscopic theory of protein folding rates. II. Local reaction coordinates and chain dynamics. *J. Chem. Phys.*, 114:5082–5096, 2001.

[46] K. L. Reid, H. M. Rodriguez, B. J. Hillier, and L. M. Gregoret. Stability and folding properties of a model beta-sheet protein, Escherichia coli cspa. *Protein Sci.*, 7:470–479, 1998.

[47] D. S. Riddle, V. P. Grantcharova, J. V. Santiago, E. Alm, I. Ruczinski, and D. Baker. Experiment and theory highlight role of native state topology in SH3 folding. *Nat. Struct. Biol.*, 6:1016–1024, 1999.

[48] N. Schonbrunner, K. P. Koller, and T. Kiefhaber. Folding of the disulfide-bonded beta-sheet protein tendamistat: Rapid two-state folding without hydrophobic collapse. *J. Mol. Biol.*, 268:526–538, 1997.

[49] B. A. Shoemaker, J. Wang, and P. G. Wolynes. Structural correlations in protein folding funnels. *Proc. Natl. Acad. Sci. U. S. A.*, 94:777–782, 1997.

[50] B. A. Shoemaker, J. Wang, and P. G. Wolynes. Exploring structures in protein folding funnels with free energy functionals: The transition state ensemble. *J. Mol. Biol.*, 287:675–694, 1999.

[51] D. L. Stein. A model of protein conformational substates. *Proc. Natl. Acad. Sci. U. S. A.*, 82:3670–3672, 1985.

[52] R. H. Swendsen. Modern methods of analyzing monte-carlo computer-simulations. *Physica A*, 194:53–62, 1993.

[53] H. Taketomi, Y. Ueda, and N. Go. Studies on protein folding, unfolding and fluctuations by computer-simulation .1. effect of specific amino-acid sequence represented by specific inter-unit interactions. *Int. J. Peptide Protein Research*, 7:445–459, 1975.

[54] N. A. J. van Nuland, F. Chiti, N. Taddei, G. Raugei, G. Ramponi, and C. M. Dobson. Slow folding of muscle acylphosphatase in the absence of intermediates. *J. Mol. Biol.*, 283:883–891, 1998.

[55] A. R. Viguera, J. C. Martinez, V. V. Filimonov, P. L. Mateo, and L. Serrano. Thermodynamic and kinetic-analysis of the sh3 domain of spectrin shows a 2-state folding transition. *Biochemistry*, 33:2142–2150, 1994.

[56] R. von Mises. *Mathematical Theory of Probability and Statistics.* Academic Press, New York, 1964.

[57] J. D. Watson and F. H. C. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 177:964, 1953.

Appendix A

Molecular Dynamics Simulations

A.1 Hamiltonian for the Model

In order to check the predictions of the free energy functional theory, we followed the dynamics of the protein by using a Gō-like Hamiltonian [53] to calculate the energy of the protein for a given configuration. Gō-like means that the Hamiltonian takes into account only native interactions. Herein our model, each of these interactions has the same amount of energy, if any two residues are within a certain cut-off distance, they are given a fixed value of contact energy [7, 17].

Residues of the protein can be thought as droplets centered in their C_α positions. Residues form a chain by bond and angle interactions. The geometry of the native state is given in the dihedral angle potential and a non-local potential. Energy of a configuration Γ of a protein having a native state configuration Γ_0 is given by [7, 25]

$$\begin{aligned} E(\Gamma, \Gamma_0) &= \sum_{\text{bonds}} K_r(r_i - r_i)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 \\ &+ \sum_{\text{dihedral}} K_\phi^{(n)}[1 + \cos(n \times (\phi - \phi_0))] \\ &+ \sum_{i<j-3} \left\{ \epsilon_1(i,j)\left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10}\right] \right. \\ &\left. +\epsilon_2(i,j)\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \right\} \end{aligned} \tag{A.1}$$

In Eq. A.1, r is the distance between two adjacent residues at configuration Γ and r_0 is the distance between them at native configuration Γ_0. Similarly $\theta(\theta_0)$ is the angle formed by three subsequent residues and $\phi(\phi_0)$ is the dihedral angle formed by four subsequent residues at configuration $\Gamma(\Gamma_0)$. The dihedral potential (third term on the right hand side of Eq. A.1) is a sum of two terms for every four subsequent C_α atoms, one with period $n = 1$ and the other with period $n = 3$. The last term on the right hand side consists of two terms; first one is the non-local native interactions and the second one is the short-range repulsive potential for non-native interactions. If the residues i and j are native contact pair then $\epsilon_1(i,j) = 1$ and $\epsilon_2(i,j) = 0$. If they are non-native then $\epsilon_1(i,j) = 0$ and $\epsilon_2(i,j) = 1$. σ_{ij} is the distance between residues i and j in the native state. For native pairs it is equal to the native distance between the residues and for non-native contacts it was taken to be 4 Å in our simulations. K_r, K_θ and K_ϕ are the strengths of different interactions, in our simulations $K_r = 100$, $K_\theta = 20$, $K_\phi^{(1)} = 1$ and $K_\phi^{(3)} = 0.5$.

For the calculation of the native contact map for a protein, native contacts between pairs of residues (i,j) are taken to be zero if $j \leq i+3$, since three or four adjacent residues are already assumed to be interacting in the angle and dihedral terms [7]. We defined that residues i and j are in native contact if either the heavy side chain atoms or C_α atoms are within a cut-off distance 4.8 Å. The measure of the nativeness for a configuration $Q(\Gamma)$, is the fraction of the formed native contacts at that configuration. Since this is not an all atom simulation, we do not keep track of all the atoms, but only C_α atoms. So, during the simulation, a contact is taken to be formed if the C_α atoms of the residues i and j are within a distance $1.2\sigma_{ij}$.

We used a simulation package named AMBER which uses Berendsen algorithm [2] to run constant temperature molecular dynamics simulations, which solves the Newtonian equations of motion numerically by rescaling the velocity to keep the temperature constant (by using Berendsen algorithm to couple the system to an external bath). In the simulations, both temperature and energy are measured in the units of the folding temperature T_f.

A.2 Free Energy Profile

For every protein structure we ran the molecular dynamics simulations numerous times to have enough sampling. After that we used the results from the WHAM algorithm [10, 11, 52] to get the free energy profile $F(Q)$ as a function of the reaction coordinate Q. This algorithm estimates the free energy profile $F(Q)$ at a specific temperature by using the approximation that logarithm of the probability distribution of the order parameter Q at fixed temperature can be considered as an estimate for the free energy profile. In a canonical ensemble, probability of variable Q to have value Q_1 can be calculated by

$$P_T(Q_1) = \frac{W(Q_1)e^{-E(Q_1)/T}}{Z_T} \tag{A.2}$$

where $E(Q_1)$ is the energy of the system at Q_1, $W(Q_1)$ is the density of states available for the value Q_1 and Z_T is the canonical partition function at temperature T. The entropy of the system can be described in terms of the density of states:

$$S(Q,T) \sim \ln[W(Q)]. \tag{A.3}$$

So, free energy can be written by the well known formula by using related quantities:

$$F(Q) = E(Q) - TS(Q). \tag{A.4}$$

Since the free energy barrier is equal to the difference of the free energies of the transition ($Q^{\ddagger}$) and unfolded (Q^{U}) states, by using this formulation:

$$\frac{P_T(Q^U)}{P_T(Q^{\ddagger})} = \frac{W(Q^U)e^{-E(Q^U)/T}}{W(Q^{\ddagger})e^{-E(Q^{\ddagger})/T}} = \frac{e^{-F(Q^U)/T}}{e^{-F(Q^{\ddagger})/T}} \tag{A.5}$$

and the barrier becomes:

$$\Delta F^{\ddagger} \equiv F(Q^{\ddagger}) - F(Q^U) = T \ln \frac{P_T(Q^U)}{P_T(Q^{\ddagger})}. \tag{A.6}$$

We calculated the corresponding probability distributions for different Q values by sampling the configuration space during all the molecular dynamics simulations.

A.3 ϕ-values

For the simulated proteins, kinetic ϕ-values are calculated by using [9, 33, 35]:

$$\phi_i = \frac{\langle n_i \rangle_{\ddagger} - \langle n_i \rangle_U}{\langle n_i \rangle_F - \langle n_i \rangle_U} = \frac{\sum\limits_{j \neq i} (Q_{ij}^{\ddagger} - Q_{ij}^{U}) \Delta_{ij}^{N}}{\sum\limits_{j \neq i} (Q_{ij}^{F} - Q_{ij}^{U}) \Delta_{ij}^{N}} \tag{A.7}$$

where $\langle n_i \rangle_U$, $\langle n_i \rangle_{\ddagger}$ and $\langle n_i \rangle_F$ are the thermally averaged number of contacts for residue i in the unfolded state, transition state and folded state, respectively.

Simulated PDB Structures

18 simulated PDB structures are: 1AB7, 1AEY, 1APS, 1CSP, 1FKB, 1HRC, 1LMB, 1MJC, 1NMG, 1NYF, 1SHG, 1SRL, 1UBQ, 1YCC, 2AIT, 2CI2, 2PTL, 2U1A.

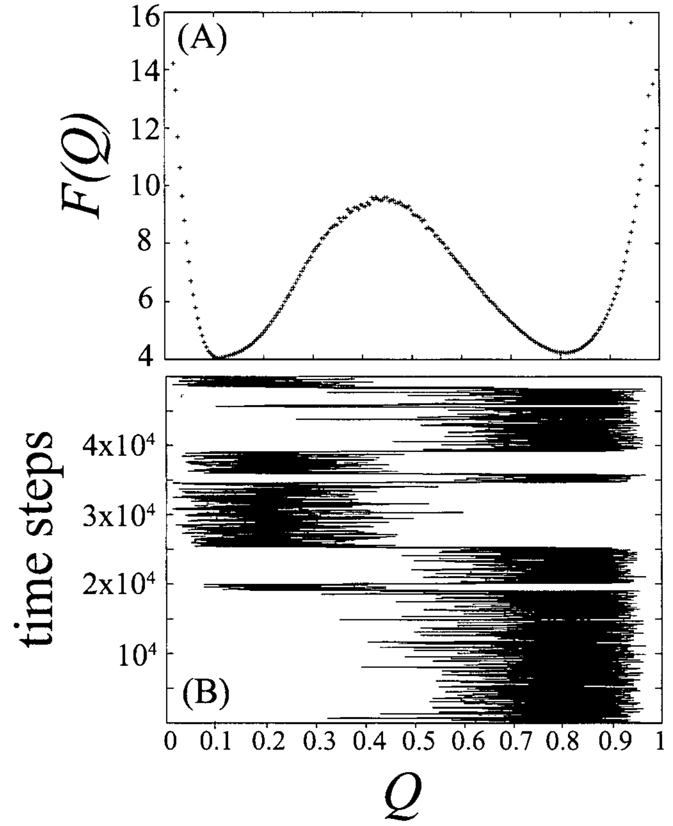

Figure A.1: (A) Free energy $F(Q)$ as a function of the reaction coordinate Q at the folding temperature T_f from our molecular dynamics simulations for the *major cold-shock protein* (PDB code 1CSP). Unfolded and native states are separated by a free energy barrier at around $Q \sim 0.45$ (B) A typical simulation (again for *major cold-shock protein*) at the folding temperature. The graph is the reaction coordinate Q as a function time that was measured in arbitrary units of molecular dynamics steps. Both graphs show 2-state behavior for this protein in our simulations.

Appendix B

How to Use Scripts and Codes to Run Simulations

This chapter explains the steps how we ran AMBER molecular dynamics simulations starting from the PDB code of a specific structure. In addition to that, it also explains the codes that were used to prepare the required data format to run the simulations and the codes that were used to extract data from the output of the simulations (ϕ-values, barriers, etc.).

First step to start a simulation is to download the necessary PDB structure from the Protein Databank. The file "PDBCODE.pdb" has some information that is not necessary for running minimalist Gō-like simulations, such as the type of amino acids and explanatory text. Some PDB structures also have more than one model (different experimental results). In this case one needs to extract only the necessary information, which is just the coordinates of all the atoms in the protein for one specific model (in our case we chose to use the first model in the beginning of the PDB file). For this purpose, we used a Perl code "model.pl". When we run this,

```
> model.pl PDBCODE.pdb
```

it creates an output file "PDBCODE.coord" with several columns including the residue number, atom type and the coordinates of the atoms. By using this information, now we can calculate the values of the necessary variables for the Hamiltonian (e.g. θ angles, dihedral angles ϕ and covalent bond distances r) in Eq. A.1. For this purpose we used the executable part of the C code named "gen.c";

```
> gen.exe PDBCODE.coord prm.crd PDBCODE.contacts > prmtop
```

which uses the coordinate file "PDBCODE.coord" as the input and creates several output files with various information. "prm.crd" is the file with all the coordinates of the C_α atoms in the structure. "PDBCODE.contacts" has 3 columns; first two columns show the residue numbers in the protein which have a native contact (having either heavy side chain or C_α atoms closer than the cut-off distance 4.8 Å) and the last column shows the distance between those residues. The place we extract the actual information that is necessary for running the simulations is "prmtop" file. It includes the energy information that is going to be used in the recipe given by the Hamiltonian in Eq. A.1 (all the relative strengths of different interactions; K_r, K_θ, K_ϕ, ϵ_1, ϵ_2, data for r, θ and ϕ angles, etc.). This file has the correct input format to be used directly by AMBER package.

After getting all this necessary input, before starting the main run for the molecular dynamics simulations, we need to thermalize the protein structure which is in the folded state in the beginning and wait for it to come to equilibrium depending on the folding temperature T_f. To do this, we need to use the script "`looprun_pre`"

```
> looprun_pre
```

which reads the "`prmtop`" file as the input and runs the molecular dynamics simulation to find protein's equilibrium state and it prepares a new file with the new coordinates of C_α atoms, named "`coord_pre.TEMPERATURE`", to use in the main simulation run.

Next step is to start the molecular dynamics simulation. For this purpose, we used the script "`looprun`"

```
> looprun
```

which tells the AMBER software to run the simulation with the chosen options; such as the temperature, number of steps between the consecutive samplings, maximum number of time steps, etc. In the end of each run, we get two output files; "`mdcrd.TEMPERATURE.NUMBEROFRUN.gz`" and "`mden.TEMPERATURE.NUMBEROFRUN.gz`" which have all the coordinate information and all the energy information for different samplings in that run, respectively. We keep running the simulation and sampling until we get enough ($\sim$ 15) barrier crossings (folding-unfolding event).

When we are done with simulations and got enough sampling, we could extract the necessary information from the output data. For this purpose, we first run the executable part of the C++ code "`GetAll.cpp`";

```
> GetAll.exe PDBCODE TEMPERATURE FIRSTRUN LASTRUN CUTOFF
```

where "`FIRSTRUN`" and "`LASTRUN`" are the corresponding numbers for the first and the last runs respectively (e.g. 1 and 50) and "`CUTOFF`" is the parameter that we need to choose which multiplies native distance (σ_{ij}) (see Eq. A.1) to calculate a cut-off value for C_α atoms (it is 1.2 in our case). It creates two output files; "`PDBCODE.TEMPERATURE.QFE`" has the total energy and free energy profile as a function of the reaction coordinate Q and "`PDBCODE.TEMPERATURE.All`" has the kinetic, potential and total energy information for any snapshot from the simulation (snapshots can be taken periodically with a period of desired number of time steps). By looking at the free energy profile, one can find the Q-value for the unfolded, transition and the folded states. Thermal transition state (TTS) can be approximated by first calculating the Q-value that corresponds to the maximum of the barrier ($Q^\ddagger$) and by finding the interval of Q where free energy drops until 20% of its maximum value on both sides of $Q^\ddagger$, the interval is the thermal transition state.

To calculate ϕ-values, we first need to run the executable part of another C++ code `qAverage.cpp`;

```
> qAverage.exe PDBCODE TEMPERATURE FIRSTRUN LASTRUN
```

which reads the data from "`mdcrd.TEMPERATURE.NUMBEROFRUN.gz`" files and calculates the average number of contacts for each residue in the protein as a function of the reaction coordinate Q (e.g. $\langle n_i \rangle_U$, $\langle n_i \rangle_\ddagger$, $\langle n_i \rangle_F$ are the average number of contacts residue i has in the unfolded $Q = Q^U$, transition

$Q = Q^{\ddagger}$ and folded $Q = Q^{F}$ states respectively) and creates the output file "`PDBCODE.qAverage`". Once we get this information, it is straightforward to calculate the ϕ value for each residue by using Eq. A.7.