ABSTRACT

Title of Document: ONE-DIMENSIONAL FREE ENERGY SURFACE MODELS OF PROTEIN FOLDING: CONNECTING THEORY AND EXPERIMENTS

Urmi R. Doshi, Doctor of Philosophy, 2007

Directed By: Dr. Victor Muñoz, Department of Chemistry and Biochemistry

Experimental techniques have now reached the sub-microsecond timescale necessary to study fast events in protein folding. However, analysis of fast folding experiments still commonly rely on conventional procedures that provide an oversimplified picture i.e. an all-or-none transition between the unfolded and native states, which is not valid for all cases. Moreover, due to the presence of discrepancies between theoretical predictions and experimental observations, discerning the correct mechanisms of protein folding becomes difficult. This is true even for the most elementary processes such as α-helix formation. Recent laser-induced temperature jump experiments on α-helical peptides have revealed unprecedented complexity in relaxation kinetics. These observations are suggested to be incompatible with the nucleation-elongation theory for α-helix formation. However, the detailed kinetic model based on nucleation-elongation theory developed in this work quantitatively reproduces all the observed complex kinetics. The results are rationalized using a simple one-dimensional projection of free energy surface. It is concluded that the observed probe-dependent and thermal perturbation size-dependent multiphasic relaxation kinetics are

consequences of helix fraying and heterogeneity of peptide sequence. Remarkably, all the kinetic behaviors predicted by the detailed model are successfully reproduced by diffusion on one-dimensional free energy surface. The one-dimensional free energy approach thus validated empirically is then extended for the analysis of protein folding experiments. For this purpose a simple mean field model is formulated that is consistent with the size-scaling properties of thermodynamic parameters as well as with the observation of entropy convergence at high temperatures. The model describes the effects of chemical and thermal denaturation, making it amenable for direct comparison with experimental observables i.e. folding rates and heat capacity changes on a quantitative level. The main advantage of the model is the treatment in which free energy barrier on one-dimensional profile is allowed to modulate by just one parameter, that can be directly related to protein size, structure- and sequence-dependent energetics. Recently the one-dimensional free energy surface model has been applied for analyzing the dependence of rates on temperature and chemical denaturant in fast folding proteins. This analysis has allowed simultaneous investigation of energetic and dynamic factors governing folding kinetics. Unlike traditional methods the model serves as an analytical tool without making any *a priori* assumptions about the presence of a barrier. With its simplicity and versatility the model provides the foundation for exploring general trends in protein folding as well as prediction of folding properties at the level of individual proteins.

ONE-DIMENSIONAL FREE ENERGY SURFACE MODELS OF PROTEIN FOLDING: CONNECTING THEORY AND EXPERIMENTS

By

Urmi R. Doshi

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Dr Victor Muñoz, Chair
Dr. George Lorimer
Dr. David Fushman
Dr. Dorothy Beckett
Dr. Norma Allewell
Dr. Wolfgang Losert, Dean's representative

Acknowledgements

This PhD thesis is the culmination of an exciting journey that began in Fall 2001 in the Biochemistry department at the University of Maryland.

Foremost I would like to express my sincere gratitude to my advisor Dr Victor Muñoz for guiding me throughout this journey. His interdisciplinary approach to solving problems and attention to detail has been very motivating. I thank him for his patience and showing the confidence in me that was necessary for the completion of this project.

Thanks to all the past and present members of Muñoz lab for making the work environment very friendly and conducive for sharing views. In particular, I thank Athi Naganathan and Jianwei Liu with whom I have had several interesting discussions.

I thank Eric Henry at NIH for providing the CVODE program that helped a great deal in the helix-coil calculations.

My deepest thanks go to my husband Jitesh Jasapara for his patience and support over the years. Finally I would like to thank my parents and brother for their love and unwavering support when I was with them and being a phone call away when at a distance of few thousand miles and time zones.

Table of Contents

Acknowledgements ii
Table of Contents iii
List of Tables v
List of Figures vi
Chapter 1: Introduction and Specific Aims 1
1.1 Microscopic picture: statistical description of protein folding 2
1.2 Low-dimensional projection of free energy landscape 3
1.3 Macroscopic picture: experimental observation and theoretical analysis of various protein folding scenarios 7
1.4 Scaling laws in protein folding 13
1.5 Simple native-centric models 16
1.6 Folding dynamics and nature of free energy barriers 20
1.7 Elementary events in protein folding 22
1.7.1 Characteristics of helix-coil transition 23
1.8 Scope of the present work 26
Chapter 2: Analysis of T-jump experiments on ^{13}C-labeled peptides with a detailed kinetic model of α-helix formation 28
2.1 Introduction 28
2.2 Model and Methods 30
2.2.1 Description of the equilibrium model 30
2.2.2 Modeling FTIR amide I band spectra 34
2.2.3 Description of the kinetic model 34
2.3 Results and Discussion 38
2.3.1 Comparison between the equilibrium behavior of α-helical peptides observed in experiments and that predicted by the model 38
2.3.2 Comparison between T-jump relaxation kinetics of α-helical peptides observed in experiments and those predicted by the model 44
2.3.3 Analysis of the observed complex kinetics in helical peptides using one-dimensional (1-D) free energy surface 51
2.3.4 Investigation of helix nucleation with the detailed kinetic model 58
Chapter 3: Calculation of helix-coil kinetics as diffusion on 1-D free energy surface 60
3.1 Introduction 60
3.2 Model and Methods 60
3.2.1 Calculation of 1-D free energy functional 60
3.2.2 The diffusion model 61
3.3 Results and Discussion 62
3.3.1 Comparison between predictions of 1-D diffusive model and detailed kinetic model: Site-specific relaxation kinetics of α-helical peptides 62
3.3.2 Comparison between predictions of 1-D diffusive model and detailed kinetic model: T-jump size-dependent relaxation kinetics of α-helical peptides 65

3.3.3 Comparison between predictions of 1-D diffusive model and detailed kinetic model: Length dependence of relaxation kinetics 67
Chapter 4: Formulation of a mean field 1-D free energy surface model of protein folding 73
4.1 Introduction 73
4.2 Model and Methods 74
4.2.1 Description of thermodynamics 74
4.2.2 Modeling temperature effects 75
4.2.3 Modeling chemical denaturation effects 78
4.2.4 Calculation of free energy barrier heights and folding rates 79
4.2.5 Modeling DSC profiles and Chevron Plots 80
4.2.6 Inclusion of size, structure and sequence effects 82
4.3 Protein Database Used in the Analysis 83
4.3.1 Selection criteria 83
4.4 Results and Discussion 91
4.4.1 Particularities of the 1-D free energy surface model 91
4.4.2 Simulation of DSC and chemical denaturation experiments 94
4.4.3 Prediction of folding rates at mid-transition 98
4.4.4 Prediction of folding rates at native conditions 111
Chapter 5: Estimation of conformational entropy from statistical analysis of protein structure database 114
5.1 Introduction 114
5.2 Methods 116
5.3 Estimation of conformational entropies of test proteins 121
5.4 Results and Discussion 125
5.4.1 Comparison of conformational entropies obtained from experiments and theory 125
5.4.2 Comparison of conformational entropies of natural proteins and random heteropolymers 129
Chapter 6: Analysis of protein folding experiments with 1-D free energy surface model 135
6.1 Introduction 135
6.2 Methods 136
6.2.2 Fitting DSC profiles 136
6.2.3 Fitting Chevron plots 138
6.3 Results and Discussion 139
6.3.1 Analysis of DSC thermograms and Chevron plots 139
6.3.2 Analysis of temperature dependence of relaxation rates of fast folding proteins 144
Chapter 7: Summary and concluding remarks 146
Appendices 150

List of Tables

2.1 Spectral parameters used in modeling amide I band spectra…….41

4.1 Proteins/Protein domains used in the analysis……...86

4.2 Summary of results: Prediction of mid-point folding rates with 1-D free energy surface model ……103

5.1 Backbone conformational entropies (ΔS_{bb}^{conf}) of different amino acids.……..122

5.2 Side chain conformational entropies (ΔS_{sc}^{conf}) of different amino acids……..123

5.3 Comparison of conformational entropies from theory and experiment…….…..126

5.4 Comparison of estimates of conformational entropy per residue obtained from thermodynamic and kinetic data and from protein structure statistics……..129

5.5 Mean conformational entropy per residue for proteins used in the prediction of folding rates……..130

5.6 Sequence composition of natural proteins and average conformational entropy for individual amino acids……131

6.1 Barrier heights obtained from the analysis of DSC thermogram and Chevron plots……144

A1 Experimental protein folding rates at thermal or chemical midpoints and in absence of denaturant…..…….……...…….150

A2 Summary of structural information of proteins used in the analysis…..…….152

List of Figures

1.1 Funnel-shaped energy landscape picture……..……...3

1.2 Projection of free energy landscape onto few dimensions……..5

1.3 One-dimensional free energy profiles with different folding scenarios……..6

2.1 The nucleation-elongation model of α-helix formation……31

2.2 Equilibrium amide I band spectra of non-labeled and labeled peptides as a function of temperature……...…39

2.3 Difference amide I band spectra of non-labeled and labeled peptides as a function of temperature……42

2.4 Theoretical equilibrium thermal transition……43

2.5 Relaxation kinetics observed at selected regions of the peptide……46

2.6 Relaxation kinetics observed at different probing frequencies……..48

2.7 Relaxation kinetics of middle labeled peptides after T-jumps of different sizes…….……...……..50

2.8 Characteristics of α-helix formation explained with 1-D projections of free energy surface……..56

3.1 Comparison between predictions of 1-D diffusive model and detailed kinetic model: Site-specific relaxation kinetics of α-helical peptides.……63

3.2 Comparison between predictions of 1-D diffusive model and detailed kinetic model: T-jump size-dependent relaxation kinetics of α-helical peptides……66

3.3 Length dependence of relaxation kinetics after T-jump of same size to the same final temperature……...……..68

3.4 Length dependence of relaxation kinetics after T-jump of same size to different apparent T_m's ……...71

4.1 Correlation of thermodynamic parameters with protein size (number of residues *N*)……..…77

4.2 Functionals used in generating 1-D free energy surface and temperature dependence of free energy barrier heights……...……93

4.3 Simulations of thermal and chemical denaturation experiments……...…95

4.4 Dependence of destabilization energy on nativeness……..…97

4.5 Prediction of mid-point folding rates with 1-D free energy surface model…….100

4.6 Comparison of average mid-point folding rates of structural scaffolds from prediction and experiments……..…..104

4.7 Comparison of calculated and observed rates at mid-point conditions for various atomic models……110

4.8 Comparison of calculated and observed rates in absence of dentaurant……..112

5.1 Distribution and clustering of ϕ-ψ dihedral angles……..119

5.2 Distribution and clustering of χ and ω dihedral angles……...……120

6.1 Fits of DSC thermograms and Chevron plots to 1-D free energy surface model……. ……140

6.2 Fits of DSC thermograms and Chevron plots to 1-D free energy surface model…… ……142

A1 Three-dimensional structures and contact maps of proteins used in the analysis……157

A2 Distribution of ϕ-ψ dihedral angles for each individual amino acid……164

A3 Distribution of side-chain and peptide bond dihedral angles for each amino acid……166

Chapter 1: Introduction and Specific Aims

The prerequisite to connecting the genetic blueprint of a protein to its biological function is the folding of its amino acid chain to a specific three-dimensional structure. Most simple proteins, upon *in vitro* denaturation, have the ability to self-assemble in a reversible and reliable manner without the aid of any cellular machinery[1]. Protein folding is essentially a concerted process of isomerization reactions around several single bonds not involving any breaking of covalent linkages. Due to the astronomical number of degrees of freedom that a protein possesses, this macromolecule has the possibility of adopting a large number of conformations. Remarkably, however, proteins find the set of relatively unique conformations corresponding to a free energy minimum in biologically relevant timescales. The questions of how and why a protein achieves its native conformation have been central in biochemistry for more than five decades and referred to as the 'protein folding' problem. The problem is addressed by two rather distinct but overlapping approaches – one involves the determination of the physico-chemical principles that underlie the folding process whereas the other deals with the prediction of native structures from amino acid sequence alone. The objectives of the present work are associated with the former. Besides being a fundamental problem, protein folding is of great practical and clinical importance. The study of the basic physics guiding protein folding can provide vital clues to the cure of the many unrelated diseases such as cystic fibrosis, Alzherimer's disease, Parkinson's disease, Huntington's disease, Type II diabetes, spongiform encephalopathies and certain

types of cancers, the molecular cause for all of which is defective protein folding. An interdisciplinary approach towards solving the folding problem has allowed remarkable progress both on theoretical and experimental fronts.

1.1 Microscopic picture: statistical description of protein folding

As per Levinthal, proteins carry out a directed search following a well-defined sequence of events ('the pathway') to avoid taking cosmological time to traverse through conformational hyperspace[2]. According to this old view folding was described as a chemical reaction with many distinct and obligatory intermediates between the unfolded and the fully folded state. However, not all protein molecules may follow one unique pathway, i.e. they may take alternate folding routes from a set of possible ones. The new view describes protein folding in terms of statistical ensembles and a biased conformational search along a multi-dimensional potential energy surface sloped towards the native state[3]. The so-called 'funnel' landscape represents the idea of decreasing energy along with concomitant decrease in conformational entropy as folding progresses. The unfolded state at the broad end of the funnel consists of a rather degenerate ensemble of structures with large root-mean-square fluctuations. On the other hand the native state ensemble comprises of far fewer structures with low energy and small fluctuations in the relative positions of all the residues. The landscape picture cannot be said to be completely incompatible with the old view because its inherent multi-dimensionality arises due to the complex network of conformational ensembles that are kinetically coupled to each other in a sequential manner[4].

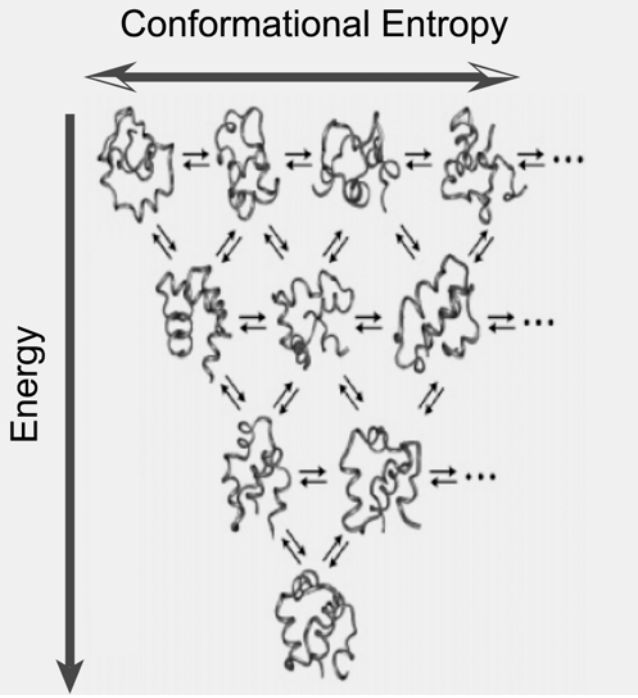


Figure 1.1 Funnel-shaped energy landscape picture

The energy landscape is shown as a chemical reaction network. The width of the funnel represents conformational entropy. As folding progresses the number of possible conformations decrease, reducing the width of the funnel. The height signifies energy that decreases as more and more native interactions are made. The inter-conversions between different conformations are shown.

(Reproduced from ref. 5)

1.2 Low-dimensional projection of free energy landscape

By intuition a complete description of a folding landscape would require the specification of all atomic coordinates of a protein (3*N* Cartesian coordinates for a protein with *N* atoms) and its surrounding solvent molecules. Due to steric constraints and concerted motions of different parts of the protein the effective number of

degrees of freedom is greatly reduced. Despite this fact the resulting energy landscape is still hyper-dimensional and impossible to comprehend. The detailed picture masks the relevant physical features of the folding process. A practical solution to this problem is to project the high dimensional landscape onto few collective degrees of freedom or coordinates. However, it is not known *a priori* which degrees of freedom are germane in describing the folding properties of proteins. The chosen collective coordinate should be able to distinguish between the folded ensemble and the manifold of unfolded conformations (i.e. be easily interpreted in terms of protein structure); be directly related to experimental probe; and act as a progress variable/kinetic ruler (i.e. change slowly relative to the total change in the reaction coordinate) to reflect the distance from the native structure and capture the kinetic features[6].

By accounting for the average energy and entropy of ensembles at each value of the collective coordinates a low-dimensional free energy surfaces can be built. Such free energy surfaces determine the thermodynamic properties (i.e. relevant conformational ensembles and free energy barriers) and predict the folding mechanisms. When a mismatch in the rate at which interaction energy (for protein-protein, protein-solvent and solvent-solvent interactions) and conformational entropy decay occurs it gives rise to a barrier separating the unfolded and native ensembles. This is the Type I scenario described by the energy landscape theory[3]. However, if the decrease in conformational entropy perfectly balances the decrease in the interaction energy the barrier may significantly decrease and even completely disappear resulting in a free energy profile with only one minimum i.e. Type 0 scenario.

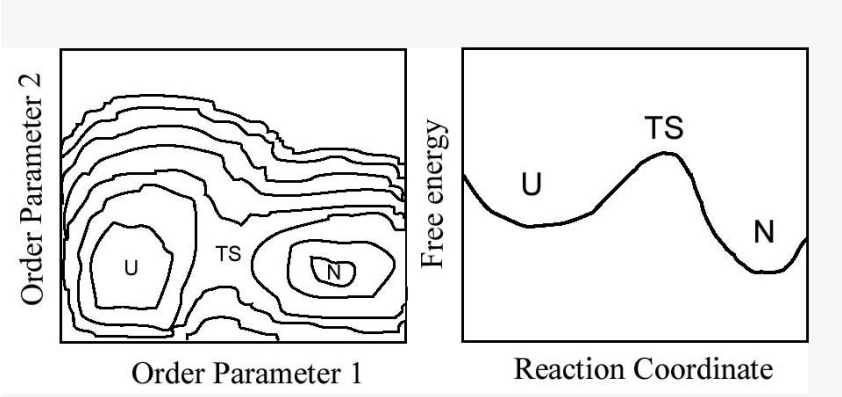


Figure 1.2 Projection of free energy landscape onto few dimensions

(This is a sketch to illustrate low-dimensional free energy surface generally produced by theoretical models and molecular dynamic (MD) simulations)

(Left Panel) Free energy surface as a function of two order parameters. (Right Panel) One-dimensional (1-D) free energy profile. Order parameters are generally number or fraction of native contacts, radius of gyration, R_g, end-to-end or intra-protein distance. If free energy contours are assumed to be marked every 3RT the free energy barrier separating the Unfolded (U) and Folded (N) ensembles is ~12RT. Also shown is the Transition state region (TS).

If the free energy barriers are large enough (>>3 RT) such that there is no accumulation of any partially folded intermediates essentially giving rise to bimodal distribution of conformations, i.e. only two distinguishable ensembles- unfolded and fully folded, at any point on the reaction coordinate. In type 0 scenario folding occurs in a downhill manner under native bias such that there is a continuous transition between ensembles having varying degrees of native structure. Thus, partially folded intermediates become accessible and populated states can be found even at any intermediate values of the reaction coordinate. However, in a global downhill

scenario predicted by a simple statistical mechanical model, a unimodal distribution of conformations results for any value of native bias (i.e. ranging from native-like to strongly destabilizing conditions)[7]. Type 0 and Type I represent the two extreme scenarios with cases in which barriers are marginally low (<3-4 RT) in between.

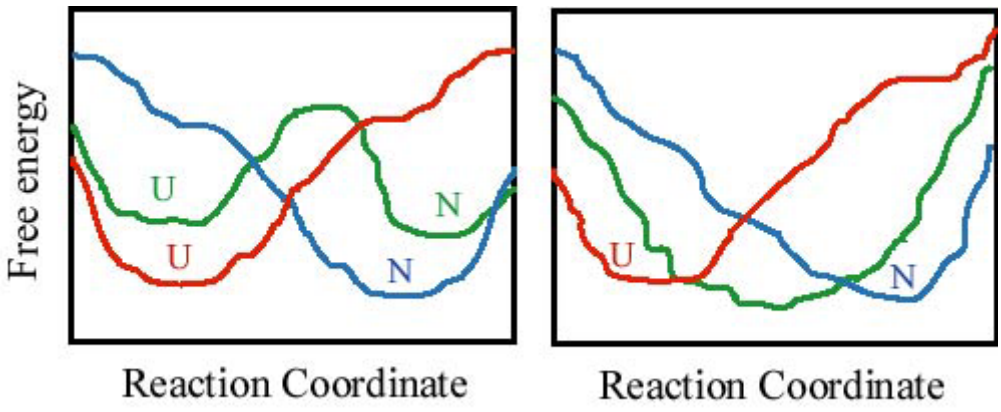


Figure 1.3 1-D free energy profiles with different folding scenarios

(Illustration of prediction by theoretical models)

(Left Panel) At high native bias (blue curve) there is complete absence of barrier (Type 0 scenario). As native bias decreases (green curve) a barrier appears (Type I). Free energy profile at highly destabilizing condition (red curve). (Right Panel) global downhill behavior: There is no barrier at any value of native bias, the single minimum just moves along the reaction coordinate (similarity to native state increases from left to right).

Energy landscape theory suggests that Kramer's-like diffusion on such low-dimensional free energy surfaces can be useful in describing folding kinetics[8]. The crucial question is whether 1-D free energy surface[a] is able to capture the kinetic and dynamic aspects of a folding reaction, or in other words, whether the selected single order parameter onto which the free energy surface is projected is sufficient to behave as the reaction coordinate. It is only through comparison with experiments that the diffusive approach for obtaining folding kinetics can be verified.

This approach has been proven successful in calculating kinetics in Monte Carlo cubic lattice simulations[9]. The rate of contact formation in unfolded peptides has also been estimated from diffusion on 1-D potential of mean force derived from end-to-end distance distribution[10]. However, prediction of protein folding rates from diffusion on 1-D free energy surfaces computed from coarse-grained features of three-dimensional structures has received only moderate success[11].

1.3 Macroscopic picture: experimental observation and theoretical analysis of various protein folding scenarios

Traditionally protein folding is investigated by bulk experiments in which the ensemble-averaged signal (spectroscopic or fluorescence) monitoring the changes in protein conformation due to perturbation by heat, acid or chemical denaturants is measured. The experimental results are usually analyzed with chemical mass action models and conclusions regarding the folding mechanisms are made based on some general criteria. For example when more than one distinct transition is observed in

[a]Although the definition of a 'SURFACE' requires at least two dimensions, here the term 'One-Dimensional free energy SURFACE' is commonly used that actually refers to one-dimensional free energy profile.

calorimetric or spectroscopic profiles it is the indication for either the presence of intermediates and more than two thermo-stable states in case of single domain proteins or sequential unfolding of individual units in case of multi-domain proteins[12,13]. A thermodynamically characterized multi-state folding process ensures multiphase/non-exponential kinetics, however the inverse is not always true. In some larger proteins (>100-120 residues) with single globular domains partially folded 'burst-phase' intermediates are observed to populate within the dead time (<5 ms) of the mixing instrument. Due to their poor characterization it is not clear whether these intermediates are distinct thermodynamic states or formed just transiently as a consequence of sudden exposure of the denatured ensemble to native conditions[14]. Several small monomeric proteins (<~100-120 residues) are observed to exhibit type I scenario in *in vitro* studies[15]. The folding process in these cases, implied to involve large barriers (several *RT*), are described as a first-order-like all-or-none transition and usually analyzed with a chemical two-state model. In the absence of partly folded intermediates in these cases the complexity involved in identifying the rate-limiting step is alleviated and the system can be sufficiently described with only two variables– the equilibrium constant and the relaxation rate constant that can be partitioned into folding and unfolding rate constants. To determine the folding mechanism it becomes necessary to characterize the activated ensemble at the top of the free energy barrier. In this respect protein engineering studies have been widely used[16] and the results are most generally interpreted in terms of structure of transition state ensemble, in spite of the robustness of the protein 3-D structure. In other words,

single or double mutations do not change the overall fold of the protein, but significantly affect its energetics and kinetics.

The general experimental criteria for classifying the folding mechanism as two-state are: (i) Sigmoidal changes in spectroscopic signals on denaturation and single peak in differential scanning calorimetric profile; (ii) Consistent local (far UV-CD[b], Fluorescence) and global (near UV-CD, FRET[b]) structural information provided by multiple probes; (iii) Isosbestic points[b] in CD spectra obtained under different denaturing conditions showing same signal intensity for denatured and fully folded conformations; (iv) Equivalence between the directly measured calorimetric enthalpy of transition with van't Hoff enthalpy obtained from fitting calorimetry profiles with a two-state assumption; (v) Single exponential relaxation kinetics; (vi) Linear dependence of the logarithm of relaxation rates on chemical denaturant concentration in concentration jump studies (V-shaped Chevron plots); (vii) Agreement between the sum of the slopes of the linear unfolding and refolding arms (m_f and m_u) and the slope of the transition region of the equilibrium chemical denaturation profile (m_{eq}); (viii) Equality of the ratio of the folding and unfolding rate constants to the equilibrium ratio of the population of unfolded and folded conformations. Significant deviations from these criteria are often interpreted as presence of intermediate states and failure to fit experimental data to a two-state model is overcome by employing a three state chemical model. These intermediates whose stability depends on experimental conditions do not seem to be obligatory in all the cases. Hence in order to probe them it is necessary to use a combination of techniques and inspect a wide range of

[b]UV-CD: Ultraviolet Circular Dichroism; FRET: Fluorescence resonance Energy transfer; Isosbestic point: A point corresponding to the wavelength at which two or more chemical species have the same molar ellipticity.

experimental conditions[17-19]. One very common signature for non-two-state behaviors is the non-linearity in the plots of logarithms of refolding (more commonly observed) and unfolding rate constants versus denaturant concentration (Chevron rollovers). Once again the general implication of a downward deviation (roll-down) of the refolding arm or curvatures in one or both arms of chevron has been the presence of intermediates[20]. It is now known that these features even in an apparent two-state kinetics can also result from any of the following effects: aggregation that is dependent on protein concentration, amino acid composition and solvent conditions[18,21]; dead time artifacts[22]; solvent effects on denatured ensemble[14]; sensitivity of native state stability to ionic strengths[23,24]; Hammond behavior, i.e. shifts in transition state ensemble due to changes in stabilities of native and denatured ensembles caused by mutations[25]; an additional slow phase arising from cis-trans proline isomerization[14]. In addition, coarse-grained versions of Go-like lattice and continuum (or off-lattice) models [c] have suggested that Chevron rollovers are universal features rather than exception[26]. Folding trajectories predicted by these models revealed that as folding conditions become more and more native-like the lifetimes of transiently populated intermediates increase. These intermediates are kinetic traps and escaping from them require barrier crossing, which impedes the conformational search in the final stages of folding. Although for two-state proteins the amino acid sequences are minimally frustrated (i.e. relatively very few destabilizing interactions as compared to stabilizing ones), few kinetic traps do exist

[c] Go-like models take into consideration only native interactions and neglect non-native ones. Lattice models are simplified representation of a protein chain in which each residue occupies a single point on a 3X3 or 5X5X5 lattice. Residue-based off-lattice models are relatively more realistic representation of polypeptide chains.

but escaping from them is much easier at experimental temperatures, resulting in linear Chevron plots. At lower temperatures, Chevron rollovers may appear even for apparent two-state proteins. Hence, failure to satisfy some empirical two-state criteria does not imply non-two-state behavior.

By the same token those proteins that do satisfy a few of the above-listed criteria specifically (i), (iii), (iv) and (v) should not necessarily be classified as two-state. Although the above criteria may be useful in distinguishing two-state from three-state proteins they need to be redefined to identify proteins with low or negligible barriers. In one case of downhill behaviors the free energy barrier disappears only under sufficiently high native bias. Once the native bias is decreased by increasing temperature type 0 scenario is replaced by type I scenario[27]. Such behaviors are experimentally observed in fast folding proteins (having low barriers) and their mutants. The other case is globally downhill, i.e. at any value of native bias there is absence of barrier[7]. A small protein domain referred to as BBL is confirmed to exhibit global downhill folding[28]. When two-state chemical model is employed to analyze data on BBL, good fits are obtained for differential scanning calorimetric profiles (DSC) as well as for sigmoidal curves from double perturbation thermodynamic experiments (using chemical and thermal denaturation)[29,30]. However, these fits do not hold any meaning (i.e. should not be taken to imply two-state folding for BBL) due to the unphysical crossing of the baselines for the native and unfolded states within the transition region. The base lines correspond to the enthalpy fluctuations in the native and unfolded ensembles. It has been shown empirically that the slope of the native baseline i.e. temperature dependence of the

native state heat capacity is proportional to protein size[31]. Higher values of the slope than that expected from the linear dependence on size show that the native ensemble has too high enthalpy fluctuations to let native ensemble be a relatively unique one. Furthermore, global fits to two-state model of CD data obtained from double perturbation studies (denaturation with urea and temperature) yield native baselines with such steep slopes that they cross the unfolded baseline at temperature within the transition region[30]. Also there is observation of cold denaturation with the shift in the maximum of CD signal as urea concentration is increased. One very important signature for downhill proteins is the inconsistency in the unfolding transitions when different structural features are observed. Probe-dependent kinetics has been observed in downhill folding mutant of lambda repressor by Gruebele[32]. In a recent high resolution proton NMR study Sadqi *et al*. has followed the unfolding of BBL by observing the behavior of individual atoms (amide protons, side chain protons and C-α protons) following heat denaturation[33]. The distribution of individual T_m's for each atomic unfolding transition (i.e. chemical shift as a function of temperature) has a standard deviation of 17 K around the average T_m of ~ 304 K unlike similar T_m's expected for a strictly all-or-none two-state behavior. Single exponential kinetics is usually assumed to imply presence of significant free energy barrier that separates two distinct populations. However, it has been shown that the decay of fluorescence energy transfer signal measuring chain dimension calculated even from a simple diffusive barrier-less model of hydrophobic collapse deviates only slightly from experimentally observed single exponential relaxation[34]. Given the combined effect of experimental noise, normalization procedures and artifacts of the instrument these

deviations are more than accounted for. Analytical theory has suggested multi-exponential or stretched exponential kinetics to be associated with barrier-less transitions[35-37]. In support of this, recently identified downhill folding proteins (engineered mutants of lambda repressor) are observed to follow stretched exponential relaxation[38]. Also, native-centric Monte Carlo dynamics model used by Kaya and Chan have predicted non-exponential kinetics under strongly native-like conditions[26]. In addition Kaya and Chan have found that their simplified atomic models predict single exponential behavior under weak native bias up to transition conditions and non-exponential relaxation at strongly native biases for both two state as well as for downhill folding proteins[39,40]. This suggests that the kinetic relaxation behavior cannot alone be used to discriminate between different folding mechanisms. Native-centric models have revealed that calorimetric criterion of $\Delta H_{vH}/\Delta H_{cal} \sim 1$ for two-state-ness is not fulfilled for downhill proteins and attempts to satisfy this criterion results in unreasonable baselines[40].

Undoubtedly the traditional way of analyzing thermodynamic and kinetic data is not appropriate for downhill folding proteins. New improved analytical tools are thus required that can explain a wide range of folding behaviors without prior assumption about the number of 'chemical' macrostates or any empirical criterion.

1.4 Scaling laws in protein folding

The mechanistic, equilibrium and kinetic properties of folding are expected to depend on protein size, sequence and topology as much as on external conditions such as pH, temperature, ionic strength, and presence of denaturing and viscogenic agents. However, determining the precise dependence of folding characteristics on these

factors require a large body of experimental data on proteins with wide range of lengths and architectures and studied under broad range of conditions with proper controls. Although this seems to be a far-fetched goal at present, some trends have already emerged from proteins investigated so far. Thermodynamic parameters namely enthalpy, entropy and heat capacity changes associated with unfolding transition obtained from calorimetric and spectroscopic studies have been compiled for around 50 proteins with size ranging from 50-350 residues[41]. Regression analyses of these energetic parameters with protein size (number of residues) have clearly demonstrated linear relationships. Lattice simulations and theoretical studies adopting polymer physics principles and scaling theory of spin glasses have predicted folding rates to show a power law dependence on protein length N of the form: $k_f \sim \exp(-c.N^{\beta})$ where c is a constant and exponent β may range from 1/2 to 1[8,42-44]. This length scaling has been confirmed by analyses performed by independent groups using a dataset of experimentally measured folding rates of ~54-57 proteins and peptides with N ~20-400 [45,46]. It has been shown that the correlation of natural logarithm of folding rates with protein length is indistinguishable for any value of 1/2, 2/3 or 1 for β. However, based on the values of prefactors obtained from linear fits to $ln(k_f)$ vs. N^{β} it has been suggested that β values of 1/2 and 2/3 are more reasonable for proteins. With an extended dataset containing 69 proteins Naganathan and Muñoz have found that the relation of logarithm of folding times with N shows a correlation coefficient of 0.74 when β =1/2 and reaches a maximum value of 0.78 as β tends to zero[47]. They have further pointed out that determining the value of β that gives a higher correlation coefficient is not appropriate given the limited size of the database and protein length

not being the only determinant of folding rates. The interesting observation, however, is the prediction of folding rates within ~1.1 orders of magnitude of experimentally determined ones knowing only the number of residues *N* for each protein.

Several structural descriptors that condense the main characteristics of native structure into a single quantity have been compared to experimentally measured folding rates. Using a dataset of ~30 two-state proteins Plaxco *et al*. have observed a strong anti-correlation (R ~-0.8) between the natural logarithm of folding rates and a descriptor of native structure (relative contact order, RCO)[48]. RCO is a measure of the average sequence separation between residues normalized by protein length *N*. ($RCO = \frac{1}{N.L}\sum^{N}\left|\Delta L_{ij}\right|$ where L is the total number of non-hydrogen atomic contacts an ΔL_{ij} is the number of residues separating a pair of contacting residues). This means that α-helical proteins in which residues are in close sequential proximity will have a lower value of RCO and fold faster compared to β proteins where more long-range contacts are made. In contrast to two-state proteins, no topology dependence has been observed for folding rates of three-state proteins[46]. Following the work of Plaxco *et al.* other structural parameters such as long range order (LRO)[49] obtained from the information on long range contacts and total contact distance, a combination of LRO and RCO[50] have been proposed as predictors of folding rates. Ivankov and Finkelstein have also found folding rates to be correlated with the effective chain length calculated using secondary structure information (number of helices and β-sheets and number of residues forming helices and β-sheets)[51].

1.5 Simple native-centric models

The above findings have suggested that native topology is sufficient to describe folding kinetics and have motivated the development of several physical models employing simple free energy functions with only Go-like interactions and disregarding non-native ones. These statistical mechanical models have basic features common to the earlier theoretical treatment by Zwanzig[52]. In Zwanzig's Ising-like model a protein's conformation is represented by a set of binary numbers: 0 or 1 signifying random coil ('incorrect') or native ('correct') states respectively for each residue. The 1-D free energy profile is essentially built with entropy derived from the degeneracy of the state and energy calculated as a linear function of the number of the incorrect residues with an additional energy gap for completely native conformation. This simplified approach has allowed for an analytical solution of the master equation and has been successful in describing the general thermodynamic and kinetic features of protein folding. However, as the sequence-dependent interactions are not accounted for explicitly, the model cannot be used directly to explain experimental data on real proteins. In the same spirit Muñoz and Eaton[11] developed a simple phenomenological model in which the interaction energy obtained from contact maps of individual proteins was compensated by the loss in configurational entropy of ordered residues. Using low–resolution features of native structures (number of native contacts) as the reaction coordinate Muñoz and Eaton obtained folding rates of 18 proteins by solving diffusion equation on a 1-D free energy surface. The iso-energetic pairwise interaction cost for each protein was adjusted to yield stabilities that matched experimental free energy of folding. An additional simplifying assumption was made

in which segments of native residues were limited to two to three as formation of native structure progressed. The barrier heights and folding rates calculated with the model agreed well with experimentally determined folding rates in the absence of denaturants with a correlation coefficient of 0.87. Concurrently similar models were developed by groups of Baker and Finkelstein[53-55]. Alm and Baker calculated the accessible surface areas for each residue that was assumed to be proportional to the interactions made by that residue[55]. Galzitskaya and Finkelstein described a network of kinetic pathways connecting conformations with chain links having varying number of unfolded residues and scaled the strength of each atomic contact to obtain equilibrium free energies at mid-transition ($\Delta G_{eq}=0$)[53]. While Alm and Baker derived the entropy cost for loop closure from off-lattice simulations Galzitskaya and Finkelstein obtained it from polymer physics principles. These groups employed recursive algorithms to determine transition state ensembles and identify the pathway having the lowest free-energy maximum. Comparison of theoretically calculated transition state free energies to experimentally obtained mid-transition rates Alm and Baker obtained a correlation coefficient of 0.67 for a dataset of 37 two-state proteins[56]. Finkelstein's group predicted mid-point rates with 59% correlation for 19 two-state proteins using a pre-factor of 10^8 s^{-1} and mid-transition barrier heights[57]. Moreover, distribution of phi values (measure of effect of mutations at residue-level on the free energy of transition state ensemble) obtained theoretically by the above-mentioned simple models were only in moderate agreement with experimentally measured ones[11,53,55].

In spite of the success in predicting folding rates of two-state proteins there are certain limitations associated with these models. Their rigid 1-D nature imposes the formation of native structure in a sequential manner. There are approximations regarding the choice of the reaction coordinate, level of detail in representing protein structures to obtain residue-residue contacts, and the number of allowed unfolded links or contiguous native segments to calculate the partition function. Moreover the free energy functions employed by these models are of limited accuracy requiring parameterization from experiments. Recently Henry and Eaton performed a combinatorial assessment in which the performance of each assumption used in such models was evaluated systematically[58]. They found that prediction of folding rates in the absence or presence of denaturant was insensitive to the choice of reaction coordinates (i.e. number of native residues or native contacts) consistent with the idea from analytical theory that such collective coordinates may be sufficient to describe protein folding kinetics. Surprisingly C-α representation of protein structure was found to perform better in predicting folding rates in water than all non-hydrogen heavy-atom description. Moreover, two entropic parameters – one for conformational entropy of residues in secondary structures and the other for residues in disordered loops were required to improve prediction over that using just one value of the parameter for all residues.

Indeed these simple models that only consider the trade-off between stabilizing interaction energy and destabilizing conformational entropy are sufficient to capture the general folding characteristics of two-state proteins. The question of interest is whether they can also reproduce global downhill folding scenarios with

their inherent two-state assumption (apparent from the method of evaluating equilibrium constants by considering populations on either side of a free energy barrier) and without adjusting parameters of energetic or entropic cost. Since proteins are marginally stable as a result of small differences between large numbers of interactions, neglecting any of them will affect the delicate interplay between the two opposing components of the free energy. Especially to reproduce experimentally observed effects of mutations that are very sensitive to energetics, it is necessary to develop more robust free energy function that includes contributions from hydrogen bonding, sequence-dependent potentials and precise entropic penalties for fixing backbone as well as side chains. In addition, the statistical mechanical approach of the above models makes them cumbersome to apply for direct analysis of experiments even with approximations about the number of contiguous native segments. Finally, for a more realistic description of protein folding it is also important to incorporate energetic frustration by including non-native interactions in the free energy functional. A recent NMR study probing the unfolding of individual protons has suggested that atomic contacts between a pair of residues defined by spatial distance obtained from three-dimensional structure need not contribute towards the interaction energy of the protein[33]. Hence, this provides a cautionary example against the use of native-centric models that typically consider pair-wise interactions of native structures to obtain protein folding energetics.

1.6 Folding dynamics and nature of free energy barriers

According to the transition state theory the rate coefficient is expressed as $k = k_o.exp(-\Delta G^{\#}/RT)$ where $\Delta G^{\#}$ is the folding barrier height and the pre-exponential factor $k_o \sim k_B T/h$ has the value of 6×10^{12} at 300 K or ~0.2 ps^{-1}. This value is appropriate for gas-phase chemical reactions involving small molecules. Such reactions are described by a single pathway with few intermediates along a potential energy surface on which all molecules pass through a unique transition state. As there is very little effect in the rest of the structure apart from where the covalent bond is formed or broken a single reaction coordinate is often sufficient for these reactions. In contrast, protein folding involves formation of only non-covalent interactions and global organization of structure accompanied by large loss in conformational entropy. In spite of these differences protein folding kinetics have been interpreted with transition state theory and assigned the same prefactor of 0.2 1/ps[59]. For a macromolecule like protein where several atomic coordinates need to be fixed this value of prefactor is highly unlikely. Lack of precise estimates of prefactor in protein folding has disallowed the estimation of absolute barrier heights and reliable temperature dependence of the activation parameters ($\Delta G^{\#}$, $\Delta H^{\#}$ and $\Delta S^{\#}$) from the measured folding rate constants. For protein folding reactions occurring in aqueous solutions Kramer's theory, that describes a chemical reaction as diffusional motion over a low-dimensional free energy surface is more suitable[60,61]. Kramer's rate expression is given by $k = D.exp(-\Delta G^{\#}/RT)$ where D, the effective diffusion coefficient on one dimension reflects dynamic motions and depends on protein sequence, reaction coordinate, solvent viscosity, temperature and roughness of the

multi-dimensional energy landscape . When the barriers are several k_BT's high D represents activated dynamics associated with barrier crossing. In case of negligible barriers D reflects purely diffusive dynamics and may provide an estimate of the folding speed limit. Studies on fast folding proteins and elementary events such as loop formation or secondary structure formation have provided estimates of diffusion coefficient (section 1.7). Another approach in obtaining diffusion coefficient from observed relaxation rates is to independently determine the barrier heights. Using the data on temperature dependence of rates and upper and lower bounds of empirical estimates of diffusion coefficient Akmal and Muñoz have analyzed the thermodynamic properties of six two-state proteins[62]. Their analysis has revealed that the folding barrier heights of these two-state proteins range from 6-12 RT at 298 K. Remarkably the ratio of activation thermodynamic parameters, $\Delta H^{\#}$ and $\Delta S^{\#}$ to the total change in folding enthalpy and entropy respectively were similar for all six proteins belonging to different structural classes (all α, all β and $\alpha-\beta$). The decay of entropy was consistently faster than the gain of stabilization energy for all proteins indicating an entropic factor at play for generating the barriers. Since both energy and entropy are large numbers even relatively small difference between them can give rise to large barriers of several RT's. Naganathan and Muñoz have recently obtained absolute barrier heights using the length scaling of thermodynamic parameters heat capacity change ΔC_p and ΔH at 333 K with the expression: $n_\sigma = \Delta H(333\text{K})/(\Delta C_p.RT^2)^{1/2}$ [47]. Barrier heights were calculated from the depth of a harmonic potential at n_σ standard deviations from the potential minimum. These barrier heights were consistent to those extracted from the differential scanning

calorimetry thermograms of proteins in an alternative approach[63,64]. Most importantly using these simple procedures Naganathan and Muñoz have shown that barrier heights obtained from thermodynamic information exhibited strong correlation with experimental folding rates. Moreover they identified many previously classified two-state proteins (implying barriers of several *RT*'s) to have only marginal barriers. The important implication of their work is that barrier heights of proteins can be used as a criterion to decide folding behaviors and that, on contrary to general belief, two-state approximation does not hold for all natural proteins.

1.7 Elementary events in protein folding

In protein folding studies it is difficult to segregate the formation of secondary and tertiary structures and hence determining the timescales and mechanisms of these events necessitates studying them in isolation, outside the context of protein. Understanding the factors that contribute to the stability and formation of secondary structural elements has provided important clues to the dynamic aspect of protein folding. α-helical peptides containing ~20-25 residues are found to fold in 200 ns - 1 μs timescale while peptides forming β hairpins take much longer up to 50 μs[65-69]. Analysis of kinetic experiments on α-helical and β-sheet peptides have suggested a rate of ~2-10 ns for peptide bond rotation[70,71]. Collapse triggered by laser T-jump and followed by FRET in an acid denatured protein domain (BBL) with 40 residues has been found to occur in 60 ns[72]. The reconfiguration dynamics associated with such non-specific collapse, where chain dimensions reduce without formation of native-like interactions, may likely reflect the motions occurring in the earliest stages of folding. In cytochrome c, however, collapse probed by Trp fluorescence has been

shown to take ~100 μs which was at least 3 orders of magnitude slower than estimates given by theoretical studies on homopolymer collapse[73]. This discrepancy is attributed to the formation of rather few but specific tertiary interactions in heteropolymers like proteins, more so in cytochrome c containing the heme group, that require breakage of already formed non-native interactions thereby slowing down dynamics. The rate of collapse may also be restricted by intrachain diffusive dynamics[74]. In unfolded cytochrome c formation of a loop of 50-60 residues long has been observed to take 35-40 μs[75]. A simple theory of diffusion controlled contact formation and predictions from the random walk chain model of Szabo, Schulten and Schulten suggested that the rate of loop formation scales with loop size n as $n^{-3/2}$. In proteins where typically n ranges between 6-10 the rate of end-to-end contact formation is expected to be in 1-3 μs timescale[75]. Since proteins cannot fold faster than the slowest elementary processes, this study provided the first estimate of the upper limit of protein folding rate. Recent studies of fast folding proteins and their engineered mutants have also given similar estimates of folding speed limit and hence of the effective diffusion coefficient [38,76,77].

1.7.1 Characteristics of helix-coil transition

Helix formation represents the simplest prototype of protein folding. Extensive studies in the last 60 years have resulted in a well-established theoretical description and a detailed thermodynamic characterization of helix-coil transition. α-helix formation is essentially described as a nucleation-elongation process in which at least four consecutive residues need to be fixed in helical conformations simultaneously to form the first helical turn followed by bi-directional propagation (see Figure 2.1)[78,79].

The nucleation step is difficult due to larger loss in conformation entropy compared to the increase in stabilization energy on formation of backbone interactions giving rise to a free energy barrier. Growing the existing helix by fixing additional residues is comparatively easier as a result of net gain in enthalpy. Various factors responsible for helix stability such as interactions of the charged groups with the helix macro-dipole, *i, i+4* hydrogen bonds, van der Waals interactions, dipole-dipole backbone interactions; and *i, i+3* and *i, i+4* side chain interactions; and the stabilizing effects arising from the N and C caps have now been identified[80-82]. Intrinsic preferences for helical conformation (i.e. nucleation (σ) and elongation (s) parameters of helix-coil theory) have been determined in free energy scales for each type of amino acid using host-guest studies[83]. Over the years the basic helix-coil theory proposed by Zimm and Bragg has been modified to include these sequence-dependent effects[84]. This has allowed the formulation of the AGADIR force-field capable of accurately predicting the helical content of peptides at any given temperature, pH and ionic strength[85-87].

The seminal theoretical treatment of helix-coil kinetics given by Schwarz[88] around forty years back proposed the relation between σ, s, the rate constant for helix propagation k_f and the mean relaxation time τ^*: $\tau^* = 1/(4\sigma + (s-1)^2 k_f)$ where s represents the degree of transition. At the mid-point of transition s~1, τ^* is maximum and equal to $(4\sigma k_f)^{-1}$. From the earlier experimental findings[89,90] τ^* was reported to be ~1 μs and k_f was estimated to be on the order of ~10^8 sec^{-1}. After the initial wave, studies on the kinetics of α-helix formation were halted for a long time. As a result, compared to its thermodynamics its kinetic aspects remained less well characterized. It is only in the last one decade that it received renewed interest mainly due to the

developments in the ultra-fast techniques and availability of short alanine-based peptides exhibiting considerable helical content in solution. This new generation of kinetic studies on helix formation have mainly employed laser-induced temperature perturbation techniques[91]. Using either simple polyalanine peptides (20-25 residues) or alanine-rich peptides with only single side-chain-side-chain interaction single exponential relaxation with time constants of ~10^6-$10^7 s^{-1}$ have been observed[66-68,92,93]. This timescale is 6 orders of magnitude faster than 100-millisecond estimate suggested for helix nucleation by denaturant-jump stopped flow CD experiments[94]. In order to explain this discrepancy it was argued that T-jump experiments probe only the local perturbations i.e. local formation and unwinding of helices that are much faster than helix nucleation involving the global folding/unfolding event. However, this argument is disputable because in T-jump experiments the equilibrium amplitudes are reached at the most within a few microseconds supporting the fact that there are no events occurring slower than microsecond timescale. Since the 100-millisecond relaxation was never reproduced by other research groups (in tryptophan fluorescence stopped flow studies[91]) and never observed in previous studies of helix-coil kinetics, it can possibly be the result of an artifact.

Improvements in T-jump instrumentation including reduction in dead time and the use of more protein-like sequences have permitted the resolution of an additional fast phase with lifetime of tens of nanoseconds[67,93]. Statistical mechanical models based on helix-coil theory predicting biphasic relaxation have been successful in explaining these results[95]. However, molecular dynamic simulations on alanine penta-peptides have suggested that helix formation takes place via barrier-less

conformation diffusion search[96,97]. This description is not in consensus with helix-coil theory that predicts a nucleation barrier. In support of the diffusive search model, laser T-jump studies on peptides with isotopic labels on carbonyl carbon atoms in different regions have revealed complex kinetic behavior[66,93,98]. Apparent relaxation times of helix formation have been found to depend on the magnitude of perturbation i.e. size of the T-jump and also on the specific region of the peptide probed. Moreover stretched exponential time courses have been reported for each peptide irrespective of the position of the labels or temperatures before or after the jump. This controversy regarding the mechanism of helix formation has been addressed in the current work (see Chapter 2).

1.8 Scope of the present work

The above sections provide a brief review of the progress made in understanding equilibrium, kinetic and dynamic properties of protein folding, the limitations of current analytical procedures and the gap between theory and experiments while interpreting folding mechanisms. The present work focuses on formulation of simple models of protein folding that are compatible with established theory as well as empirically observed scaling laws; and application of these models in analyzing available experimental data on alternative folding behaviors. This manucript is broadly organized into two main segments. The first segment that comprises Chapters 2 and 3 concentrates on providing a physical basis for the complex relaxation behaviors observed in kinetic studies of α-helical peptides. Specifically the following questions are addressed:

i. Can a kinetic nucleation-elongation theory explain the T-jump size- and probe-dependent relaxation kinetics observed in helix-coil transition?
ii. Can the kinetics obtained from a detailed model be reproduced from Kramer's- like diffusion on a 1-D free energy surface of α-helix formation?

Chapters 4, 5 and 6 deal with the objectives of the second segment that are listed as follows:

i. Formulation of a 1-D free energy surface model of protein folding
ii. Prediction of protein folding rates using the 1-D free energy surface model
iii. Estimation of conformational entropy from statistical analysis of protein structure database for sequence-dependent parameterization of the model
iv. Analysis of protein folding experiments with 1-D free energy surface model

For clarity, model description and results pertaining to each segment are separately discussed in individual chapters.

Chapter 2: Analysis of T-jump experiments on ^{13}C-labeled peptides with a detailed kinetic model of α-helix formation

2.1 Introduction

Gai and co-workers investigated helix formation in analogous peptides having the same sequence Ac-YGSPEAAAKAAAAKAAAA-r-NH_2 but ^{13}C labeled at carbonyls of alanine residues either in the N-terminal, middle or C-terminal regions[93]. Using laser-induced T-jump the peptides were subjected to a sudden increase in temperature and the relaxation to the new equilibrium at higher temperature was followed by Fourier transformed infrared spectroscopy (FTIR). Gai and co-workers observed that relaxation kinetics of peptides labeled in different regions after a T-jump of 10 K to a final temperature of 288 K were dissimilar. The relaxation of peptides labeled in the C-terminus region was faster than those of peptides labeled at the N-terminus or middle regions. This observation is consistent with one of the predictions of helix-coil theory that helices with intermediate lengths show helicity concentrated in the central region with ends frayed. In other words, the probability of forming helices in the middle of the peptide is greater than at the termini. Surprisingly the apparent relaxation times of the N-terminally labeled peptides were very close to those of the peptides labeled in the middle region when the differences in the signal were normalized. Also different relaxation kinetics resulted for peptides labeled in the middle region when probed at different frequencies[98]. Furthermore, subjecting the middle labeled peptides at different initial temperatures to the same final temperature, Gai and coworkers found that the relaxation rates were linearly dependent on the

magnitude of the T-jump. A T-jump of ~14 K resulted in a relaxation that was ~1.5 times faster than a 4 K jump to the same final temperature. The time courses of each labeled peptide were non-exponential and fitted to stretched exponentials with β (measure of deviation from single exponential, β=1) between 0.7 and 0.85. Unlike previous T-jump experiments that used simple alanine-based sequences with residues having very similar intrinsic helical propensities, Gai and co-workers investigated more heterogeneous peptides and revealed for the first time that even simple short α-helical peptides could exhibit such complex behaviors.

These results were interpreted in terms of a conformational diffusive search model describing helix formation as a downhill diffusion process in the coil region of the phase space. And hence they were not considered compatible with nucleation-elongation theory that predicts a free energy barrier separating helical and coil ensembles. The conformational diffusive search description suggested originally from MD simulations of alanine penta-peptides[96,97] provides only an anecdotal picture and fails to explain the observed probe-dependent kinetics. In a recent comprehensive all-atom MD study using global distributed computing Sorin and Pande performed a quantitative assessment of the AMBER force fields generally used in simulating helical peptides[99]. The AMBER-94 variant, the force field used earlier by Hummer and co-workers in the MD simulation of alanine penta-peptides was found to overstabilize the helical conformations and hence predict smaller to negligible barriers.

In order to analyze the kinetic experiments performed by Gai and coworkers an improved version of the nucleation-elongation model is formulated. This model

explicitly takes into account amino acid sequence dependence and allows for helix breaking and merging. The aforementioned results of T-jump kinetics are explained using simple 1-D free energy projections. To directly compare kinetics simulated by the model with that seen in experiments, FTIR signals are calculated from time dependent probabilities generated by the model and amide I spectra represented as Gaussian curves.

2.2 Model and Methods

2.2.1 Description of the equilibrium model

The fundamental features of the model are similar to the earlier statistical mechanical models of α-helix and β-sheet formation. The basic conformational unit in the model is the peptide bond. Each i^{th} peptide bond can assume one of the two states: helical if the flanking dihedral angles ϕ_{i+1}, ψ_i have α-helical values; or coil for any other values of the dihedral angles. The coil state is the reference having statistical weight of 1. For simplicity both ϕ_{i+1}, ψ_i are assumed to rotate simultaneously. Fixing a pair of dihedral angles in α-helical conformation accompanies loss in conformational entropy. More loss in conformational entropy occurs when dihedral angles of several successive residues (typically 4-5) are fixed. This unfavorable and rate-limiting process is helix nucleation and gives rise to a free energy barrier. As the helical segment increases to a particular length compensating backbone interactions such as van der Waals, dipole-dipole and hydrogen bonds between the carbonyl oxygen of i^{th} residue and amide hydrogen of $i+4^{th}$ residue are formed. From this point on, each subsequent hydrogen bond is realized by fixing just one more pair of dihedral angles

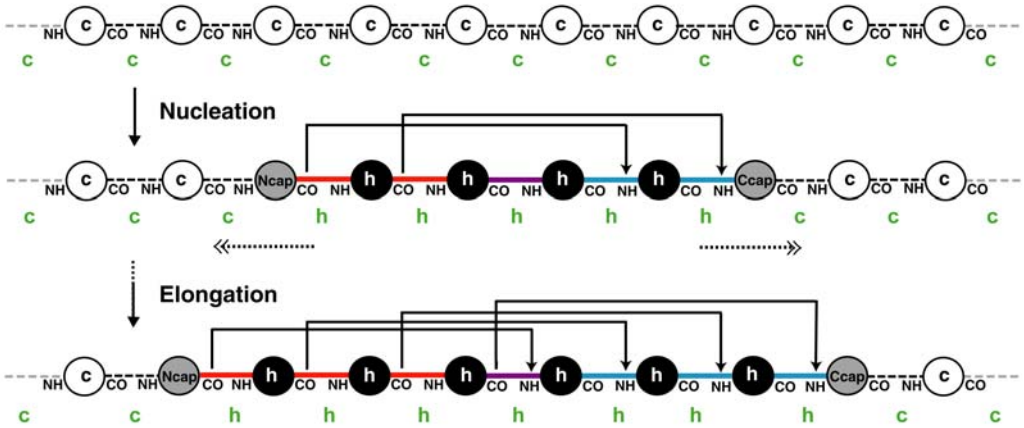


Figure 2.1 The nucleation-elongation model of α-helix formation

A peptide segment undergoing helix initiation and bi-directional propagation along with conformational assignments to residues (as in AGADIR) and peptide bonds (as in this model) is shown here. Each residue represented by its C-α atoms (circles) in either coiled (letter 'c') or helical (letter 'h') conformation or forming N- and C-caps (gray circles) are shown. The letters in green indicate the conformation of each peptide bond. Peptide bonds with only carbonyl hydrogen bonds (red), only amide hydrogen bonds (blue), both carbonyl and amide hydrogen bonds (purple) and no hydrogen bonds (dashed lines) are shown. An illustration of formation of helix nucleus with 5 consecutive peptide bonds in helical conformation (not necessarily flipped in a simultaneous fashion) and hydrogen bonds between the carbonyl group of residue *i* and amide group of residue *i+4* (bent black arrows) is also given.

resulting in the net gain of stabilizing backbone interactions. This process leads to elongation that may occur in either directions of the nucleated helix (Figure 2.1). The other favorable interactions responsible for holding the helix together include side chain interactions between *i, i+3* and *i, i+4* residues; helix capping effects and electrostatic interactions of charged residues with the helix macro-dipole. Amino acid sequence-dependent free energy contributions from all these interactions are directly obtained from the empirically derived parameters of the AGADIR algorithm based on helix-coil theory[81,85,86]. However, in AGADIR the conformational unit is residue rather than peptide bond in this model. Due to this difference the minimal helical unit of AGADIR comprising of 6 residues: 4 helical residues plus the N- and C- caps is equivalent to a helical nucleus of five consecutive peptide bonds in this model. This mapping allows the assignment of the same mean enthalpic contribution of AGADIR, $w_{bb} = \exp(-\Delta G_{bb} / RT)$ where ΔG_{bb} is the sum of backbone interactions, to helix nucleation in this model. The statistical weight for fixing any peptide bond has only entropic contributions and its value depends on the intrinsic propensities of amino acid residues (from AGADIR) flanking the peptide bond, $w_{in} = \exp(-(\Delta G_{in,i} + \Delta G_{in,i+1})/(2RT))$. Hence the statistical weight for every helical peptide bond added to the nucleus is the product of w_{bb} and w_{in}. The N- and C-cap weights (w_n and w_c) arise from residues immediately preceding the first helical peptide bond and just after the last helical peptide bonds respectively. In AGADIR any helical segments having lengths less than helix nucleus (6 residues) are not considered explicitly because of their low probabilities. In order to provide a detailed kinetic description of helix formation the current model includes all the short

segments with one to four helical peptide bonds and allows for their kinetic connections with other helical species. The statistical weights of these short helical segments include contributions only from N- and C- caps and intrinsic helical preferences. For example, statistical weight of a helical segment is given by $w_n.(w_{in})^h.(w_{bb})^{h-4}.w_c$ if h>=5 and by $w_n.(w_{in})^h.w_c$ for h<5. With binary states for each peptide bond there can be 2^N possible combinations or species for peptide of length N+1. The peptide analyzed experimentally by Gai and co-workers is 19 residues long with the non-natural D-Arg as the C-cap and ends protected (sequence: Ac-YGSPEAAAKAAAAKAAAA-r-NH_2). In the model D-Arg is replaced by Gly, the best-known natural C-cap[100], in addition to placing one Gly residue at each end to account for acetylation and amidation at N- and C-terminals respectively. These substitutions result in a 21-residue peptide (20 peptide bonds). One major improvement in this model as compared to helix-coil models used to analyze previous equilibrium and kinetic experiments is the introduction of double sequence approximation (DSA). Unlike earlier models that employed single sequence approximation, i.e. allowing helix breaking and forming only from the ends of a helical segment, this model permits helix breaking in the middle of a helical segment and merging of two helical segments. Since helix nucleation is energetically unfavorable not allowing more than two helix initiation sites is a good enough approximation for a peptide length of 21 residues (n=20). This is also confirmed by stochastic kinetic simulations involving 2^{20} conformations in which species having more than two helical segments are only transiently populated with half life of <400 ps. Using DSA the number of possible species drastically reduces from 2^{20} to

6196 including coil conformation. The partition function in DSA is given by $Q = 1 + \left(\sum_{i=1}^{n} \sum_{j=1}^{n-i+1} w_{ij} \left(1 + \sum_{p=1}^{n-i-j} \sum_{q=i+j+1}^{n-p+1} w_{pq} \right) \right)$ where w_{ij} and w_{pq} are statistical weights of two helical segments having i and p helical peptide bonds starting at position j and q respectively.

2.2.2 Modeling FTIR amide I band spectra

Amide I band spectra corresponds to normal modes of vibration mainly arising from the stretching of the C=O bond. However, vibrations due to stretching of N-H bond also contributes significantly up to ~25% [101,102]. To take this into account basis spectra of peptide bonds with variable number of hydrogen bonds involving either C=O or N-H or both are generated by modeling them as Gaussian or Lorentzian curves. Figure 2.1 shows the classification of peptide bonds based on the hydrogen bonding pattern and absence or presence of ^{13}C labels on C=O. Each peptide bond chromophore is assigned parameters that describe the characteristics of its Gaussian/Lorentzian curve. To reproduce temperature-dependent amide I spectra measured by Gai and co-workers amide I spectra are calculated at the same experimental temperatures as weighted average of basis spectra of all kinds of peptide bonds.

2.2.3 Description of the kinetic model

In the model rotation of a peptide bond from coil to helical (on rate) or from helical to coil (off rate) conformations constitute elementary kinetic steps. For each conversion an elementary transition state is assumed in which entropy is lost due to the fixing of

the peptide bond in helical angles but no interactions are realized yet. The species can be kinetically connected to only those other species that have one more or one less helical peptide bond. For example species such as ---cccchhhhccc--- can be converted to ---cccchhhhhcc--- or ---cchchhhhccc--- by a single flip but not to ---ccchhhhhhc--- or ---chhchhhhccc---. Similarly species with two helical segments such as ---cccchhhhhcccchhhccc--- can be connected to species like ---ccchhhhhhcccchhhccc--- or ---cccchhhhhcccchhhhcc--- but neither to ---cccchhchhcccchhhccc--- nor ---ccchhhhhhcccchhhhcc---.

The on rate is expressed as $k_{on} = k_o.w_{in}$ where k_o is the pre-exponential factor that defines the rate of the peptide bond rotation in the model and varies with 1/T in the same manner as the temperature dependence of viscosity of water. k_o is an adjustable parameter. The off rates are obtained by detailed balance, $k_{off} = k_{on}.(w_{h+1}/w_h)$ where w_h and w_{h+1} are statistical weights of species differing by one helical peptide bond. As given below the set of master equations (Equations 2.1-2.3) is built by using on and off rates for each transition:

for i = 1

$$\frac{dP_{0,0,0,0}}{dt} = \sum_{j=1}^{n}\left(k_{off}^{j}P_{1,j,0,0} - k_{on}^{j}P_{0,0,0,0}\right) \quad \text{.....(2.1)}$$

Equation 2.1 represents the inter-conversion of species with a single helical segment comprising of a single helical peptide bond with the fully coil species. The superscript on k_{off} and k_{on} rates is the peptide bond number that is undergoing conversion. The subscripts on the probabilities of the species, i.e. P, are the indices for the lengths (indices 1 and 3) and positions (indices 2 and 4) of the two helical segments.

for 1 < i < n ; j > 1; i + j - 1 < n

$$\frac{dP_{i,j,0,0}}{dt} = \underbrace{\left(\sum_{q=i+j+1}^{n} \left(k_{off}^{q} P_{i,j,1,q} - k_{on}^{q} P_{i,j,0,0} \right) \right)}_{\text{I}} + \underbrace{\left(\sum_{q=1}^{j-2} \left(k_{off}^{q} P_{1,q,i,j} - k_{on}^{q} P_{0,0,i,j} \right) \right)}_{\text{II}} +$$

$$\underbrace{\left(k_{off}^{i+j} P_{i+1,j,0,0} - k_{on}^{i+j} P_{i,j,0,0} \right)}_{\text{III}} + \underbrace{\left(k_{off}^{j-1} P_{i+1,j-1,0,0} - k_{on}^{j-1} P_{i,j,0,0} \right)}_{\text{IV}} +$$

$$\underbrace{\left(k_{on}^{j} P_{i-1,j+1,0,0} - k_{off}^{j} P_{i,j,0,0} \right) + \left(k_{on}^{i+j-1} P_{i-1,j,0,0} - k_{off}^{i+j-1} P_{i,j,0,0} \right)}_{\text{V}} +$$

$$\underbrace{\sum_{m=1;p=i-2;b=j+1}^{m=i-2;p=1;b=i+j-2} \left(k_{on}^{b} P_{m,j,p,j+p-1} - k_{off}^{b} P_{i,j,0,0} \right)}_{\text{VI}} \quad \text{.....(2.2)}$$

for boundary conditions, i.e. for $i + j - 1 = n$ *omit terms I and III; and for* $j = 1$ *omit terms II and IV.*
for $i < 3$, *omit term VI.*

Equation 2.2 represents the inter-conversion of species with single helical segment with species having two helical segments- the second segment with one helical peptide bond on the right of the first helical segment (Term I), on the left (Term II); with species having a single helical segment with one more helical peptide bond to the right (Term III), to the left (Term IV); with species having a single helical segment with one less helical peptide bond (Term V); and with species forming two helical segments as a result of breaking a single segment

for 0 < i = n ; 0 < p = n; i + j + 1 < q; j > 1; p + q - 1 < n

$$\frac{dP_{i,j,p,q}}{dt} = \underbrace{\left(k_{off}^{i+j}P_{i+1,j,p,q} - k_{on}^{i+j}P_{i,j,p,q}\right)}_{\text{I}} + \underbrace{\left(k_{off}^{j-1}P_{i+1,j-1,p,q} - k_{on}^{j-1}P_{i,j,p,q}\right)}_{\text{II}} +$$
$$\underbrace{\left(k_{off}^{p+q}P_{i,j,p+1,q} - k_{on}^{p+q}P_{i,j,p,q}\right)}_{\text{III}} + \underbrace{\left(k_{off}^{q-1}P_{i,j,p+1,q-1} - k_{on}^{q-1}P_{i,j,p,q}\right)}_{\text{IV}} +$$
$$\underbrace{\left(k_{on}^{j}P_{i-1,j+1,p,q} - k_{off}^{j}P_{i,j,p,q}\right) + \left(k_{on}^{i+j-1}P_{i-1,j,p,q} - k_{off}^{i+j-1}P_{i,j,p,q}\right)}_{\text{V}} +$$
$$\underbrace{\left(k_{on}^{q}P_{i,j,p-1,q+1} - k_{off}^{q}P_{i,j,p,q}\right) + \left(k_{on}^{p+q-1}P_{i,j,p-1,q} - k_{off}^{p+q-1}P_{i,j,p,q}\right)}_{\text{VI}} \quad \text{.....(2.3)}$$

for cases in which two helical segments merge, i.e. $i + j + 1 = q$, *omit terms I and IV and add* $\left(k_{off}^{i+j}P_{i+p,j,0,0} - k_{on}^{i+j}P_{i,j,p,q}\right)$

for boundary conditions, i.e. for $j = 1$ *omit term II; and for* $p + q - 1 = n$ *omit term III.*

for i = 1 omit term V and add $\left(k_{on}^{j}P_{0,0,p,q} - k_{off}^{j}P_{i,j,p,q}\right)$

for p = 1 omit term VI and add $\left(k_{on}^{q}P_{i,j,0,0} - k_{off}^{j}P_{i,j,p,q}\right)$

Equation 2.3 represents the inter-conversion of species with two helical segments with those species having one more helical peptide bond on the right (Term I) and the left (Term II) of the first helical segment, on the right (Term III) and left (Term IV) of the second helical segment; with species having one less helical peptide bond on either side of the first helical (Term V) and second helical (Term VI) segments.

The resulting sparse rate matrix is solved numerically using standard differential equation solver routines for stiff problems. These calculations are performed with the CVODE package provided by Eric Henry at NIH[103]. This package implements an iterative algorithm for solving stiff differential equations and sparse linear systems. Relaxation kinetics following laser T-jumps is simulated by integrating master

equation at the final temperature using equilibrium probabilities calculated at initial temperatures of the jump. To compare the relaxation kinetics calculated by the model to the one observed experimentally time-dependent FTIR signals are calculated. Most of the calculations are performed with Matlab 6.5 and Microsoft Visual C++.

2.3 Results and Discussion

2.3.1 Comparison between the equilibrium behavior of α-helical peptides observed in experiments and that predicted by the model

Experimentally, the equilibrium unfolding of non-labeled peptides and peptides labeled at N-terminus (positions 6 to 8), in the middle region (positions 10 to 13) and at C-terminus (positions 15-18) was probed by FTIR. The amide I spectra of these peptides measured by Gai and co-workers at various temperatures[93] are shown in the upper panel of Figures 2.2. In case of non-labeled (^{12}C) peptides the amide I band shows a shift in frequency from ~1635 cm^{-1} ~1650 cm^{-1} and a decrease in intensity as a result of thermal melting. The amide I band spectra of ^{13}C- labeled peptides (panels B-D of Figure 2.2) exhibit an additional peak at ~1600 cm^{-1}. The difference spectra calculated by subtracting the spectrum at the lowest temperature are shown in the upper panel of Figure 2.3. The loss in intensity in the amide I band is accompanied by the increase in the intensity of a positive spectral feature at higher wavenumbers. Interestingly, the effects of temperature on amide I spectra of peptides labeled in different regions are quite dissimilar. For peptides labeled in the middle region a sharper ^{13}C peak of very high intensity is observed while labeling at the N-terminus region results in a less intense ^{13}C peak. On the other hand, alanines labeled at the C-

terminus region produce a broad ^{13}C shoulder of very low intensity. These differences in spectral features allows the independent monitoring of helix melting in non-labeled and labeled peptides as well as provides information about the helical content in the selected regions of the peptide as a function of temperature.

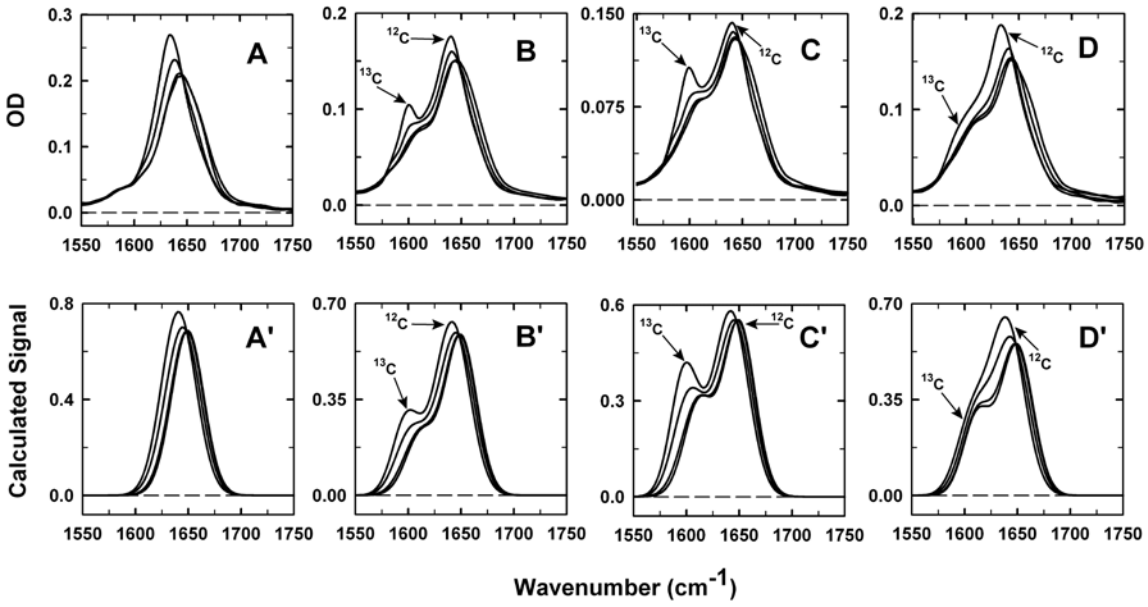


Figure 2.2 Equilibrium amide I band spectra of non-labeled and labeled peptides as a function of temperature

Panels A (non-labeled), B (N-terminus labeled), C (Middle labeled), D (C-terminus labeled) show the experimental amide I spectra obtained by Gai and coworkers at temperatures of ~276, 303, 330 and 348 K. Panels A'-D' show the corresponding amide I spectra calculated by the equilibrium model at the same temperatures. In each panel the peak with the highest intensity refers to the lowest temperature.

To compare these experimental results with the model calculations amide I basis spectra are simulated as Gaussian curves for peptide bonds with different types and number of hydrogen bonds. The basis spectra of coil peptide bonds as well as of helical peptide bonds (in helical segments having length <5) with no hydrogen bonds are modeled as Gaussian curves with a maximum at 1650 cm^{-1}. While basis spectra of helical peptide bonds with only hydrogen bonded carbonyls are represented as Gaussian curves with the maximum shifted to 1636 cm^{-1}. For peptide bonds with singly hydrogen bonded carbonyl and amino groups the basis spectra are modeled as Gaussian curves with maximum at 1639 cm^{-1} and 1646 cm^{-1} respectively. The basis spectra of ^{13}C labeled peptide bonds are similarly modeled as those of their non-labeled counterparts except with the width being narrower and the maxima shifted by ~38 wavenumbers. The strength of the transition dipole that reflects the IR-absorbing intensity is also increased in the same proportion for all isotopically labeled peptide bonds with different hydrogen bonding status. The spectral parameters for each kind of peptide bond are listed in Table 2.1. An increased entropic stabilization of the coil ensemble results due to the inclusion of conformations having shorter non-hydrogen-bonded helical segments. This effect is balanced by using a higher mean enthalpic contribution per peptide bond in each elongation step than that used in AGADIR[87]. The amide I spectra are then obtained from the basis spectra of different spectral groups and temperature-dependent probabilities of 6196 species calculated using -1.04 *kcal/mol* for the mean enthalpic contribution. The theoretical amide I spectra and the difference spectra with reference to the lowest temperature are shown in the lower panels of Figures 2.2 and 2.3. In any species when proline occurs at the first position

of a helical segment, it is assigned the same intrinsic propensity value as that for glycine. Since the amino groups of first three residues of a helix never participate in backbone hydrogen bond formation, presence of a proline at the beginning of the helix does not affect helix formation and the overall helical content as previously found by experiments[104]. The Gaussian representations of the amide I band spectra in Figure 2.2 (A'-D') show similar changes in relative intensity with temperature as those observed in experimental spectra.

Table 2.1 Spectral parameters used in modeling amide I band spectra

	Peptide bonds	**Mean μ (cm^{-1})**	**Standard Deviation, σ (cm^{-1})**	**Relative Intensity**
Non-labeled (^{12}C)	HB C=O	1639	15	1.2
	HB NH	1646	15	1.2
	HB C=O and NH	1636	15	1.0
	helical NHB	1650	15	1.3
	coil	1650	15	1.3
Labeled (^{13}C)	HB C=O	1601	13	1.8 x 1.2
	HB NH	1608	13	1.8 x 1.2
	HB C=O and NH	1598	13	1.8
	helical NHB	1612	13	1.8 x 1.3
	coil	1612	13	1.8 x 1.3

HB C=O: helical peptide bonds with only C=O hydrogen bonded (red bonds in Figure 2.1); HB NH: helical peptide bonds with only NH hydrogen bonded (blue bonds in Figure 2.1); HB C=O and NH: helical peptide bonds with both C=O and NH hydrogen bonded (purple bonds in Figure 2.1); helical NHB: helical peptide bonds with no hydrogen bonds; coil: peptide bonds in coil conformation.

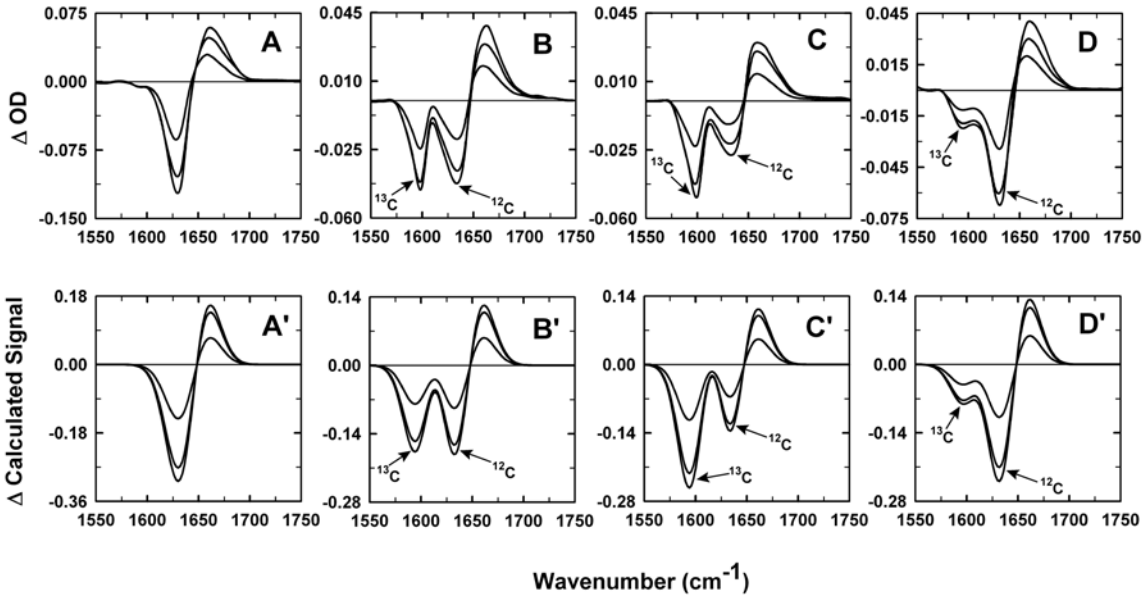


Figure 2.3 Difference amide I band spectra of non-labeled and labeled peptides as a function of temperature

Panels A (non-labeled), B (N-terminus labeled), C (Middle labeled), D (C-terminus labeled) show the difference spectra obtained by Gai and coworkers with reference to the lowest temperatures (276 K). Panels A'-D' show the corresponding theoretical difference spectra. In each panel the peak with the highest intensity refers to the lowest temperature.

It can be clearly seen from Figure 2.3 that the ratio of the maximum and the minima and the relative intensity of the ^{12}C and ^{13}C peaks observed in difference spectra obtained from experiments are reproduced in the theoretical difference spectra. Similar results are obtained when amide I band spectra are modeled as Lorentzian curves or a different set of spectral parameters are used. Although Lorentzian curves better represent the shape of the amide I spectra, the difference spectra calculated

with Lorentzian functions appear more dissimilar to the experimental difference spectra. The calculated equilibrium melting transition (Figure 2.4) show a T_m of ~293 K for the peptides considered here. Although no results of any experiments have been reported for these peptides to directly compare the T_m, the theoretical estimate is close to the T_m of 289 K suggested from two-state fits of far UV-CD data. The distribution of helical probability along the peptide sequence obtained from the calculations is shown in the inset to Figure 2.4. The helix content is maximal in the central region of the peptide and decreases towards the ends. The dip in the helix probability at the third peptide bond is due to the presence of serine in the fourth position followed by proline that act as a helix stop signal. There is an increase in the helical probability at positions at the N-terminal before serine because short helical segments comprising of one or two non-hydrogen bonded peptide bonds can be formed. The extent of fraying in the C-terminal region is larger than that seen at the N-terminus, which is also evident from the intensity of the ^{13}C peaks of peptides labeled in the C-terminal

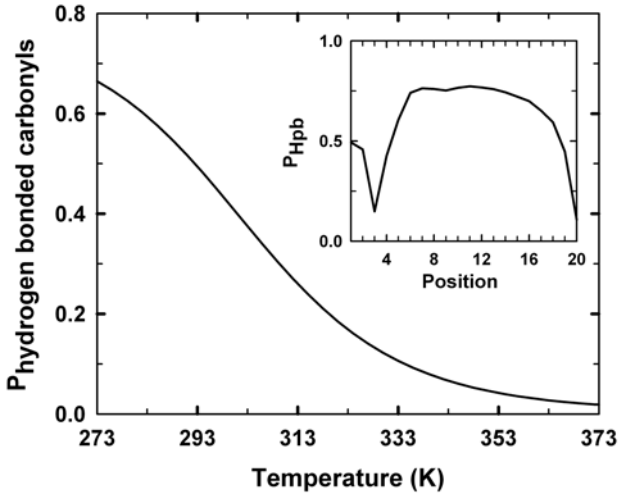


Figure 2.4 Theoretical equilibrium thermal transition

The probability of hydrogen bonded carbonyls are plotted against temperature. The inset shows the probability of finding each peptide bond in helical dihedral angles at the T_m (293K).

region. This is because out of the four labeled carbonyls at the C-terminus, the hydrogen bonds of only two carbonyls are satisfied. Unlike the N-terminus amino groups the side-chain-backbone hydrogen bonds are not favored at the C-terminus. The fraying effect at the C-terminus is, however, diminished to a small extent due to the presence of a strong C-cap.

The general spectral features of the equilibrium FTIR spectra namely the shifts in the frequency of the ^{13}C labeled peptides and the decrease in the amide I band intensities with increase in temperature are successfully reproduced by the equilibrium model. The model also predicts the characteristic end fraying effect of helix-coil transition.

2.3.2 Comparison between T-jump relaxation kinetics of α-helical peptides observed in experiments and those predicted by the model

Gai and co-workers monitored the relaxation kinetics of the isotopically labeled peptides at 1600 cm^{-1} after perturbations induced by a temperature jump of 10 K. They found that the peptides labeled at the N-terminus and the middle region showed identical relaxation kinetics when the signals are normalized to the same scale. However, relaxation kinetics of the C-terminus labeled peptides was found to be relatively faster (inset to Figure 2.5A). These relaxation traces were fitted to a stretched exponential function along with a so -called 'instantaneous' component that is not resolved due to the limitation of the response time of the instrument. The instantaneous component is generally assumed to arise from the temperature induced shifts in the equilibrium IR spectra along with some contributions from the actual

helix-coil transition (i.e. local relaxations occurring on sub-nanosecond timescale). Figure 2.5A shows the time courses predicted by the model after a simulation of a T-jump from 278 K to 288 K. The theoretical time courses also exhibit a faster relaxation of peptides labeled in the C-terminus and very similar decays for the N-terminus and middle labeled peptides. When plotted on a logarithmic scale (Figure 2.5B) the time courses show biphasic behavior. The reason of the C-terminus-labeled peptides having a faster relaxation also becomes apparent. Both the N-terminus and middle labeled peptides have similar ratios of amplitudes for the fast and slow phases while the C-terminus labeled peptides show a relatively faster fast phase with larger amplitude. This indicates that in case of peptides labeled at the C-terminus there is large amount of helix fraying resulting in the fast phase having a relatively larger contribution to the overall relaxation as compared to that of peptides labeled in other regions. Larger amplitudes and shorter times of fast phase arise from the greater amount of local perturbations at the C-terminal. A similar fraying effect should also be expected for the peptides labeled at the N-terminus. However, the presence of residues having low helical propensities results in slower rates of elongation at the N-terminus of the peptide making the fast phase considerably slower than that of C-terminus labeled peptides. But the presence of a strong capping motif (SPE) at the N-terminus stabilizes the helix and does not allow any modification to the slow phase i.e. the global melting of the helix. In earlier kinetic studies peptides containing only alanine, lysine and arginine were investigated, all of which have high intrinsic helical propensities. The high helical propensities give rise to faster propagation/de-propagation rates that ultimately result in a fast phase with very low relaxation times

(i.e. 10 ns). Hence the fast phase was often not resolved due to the detection limit of the T-jump instrumentation employed in those studies. As mentioned above the peptide used by Gai and coworkers and considered here has a heterogeneous sequence with residues in the N-terminal region having low intrinsic propensities and strong capping motifs. Due to this the propagation rates at the N-terminus are slower

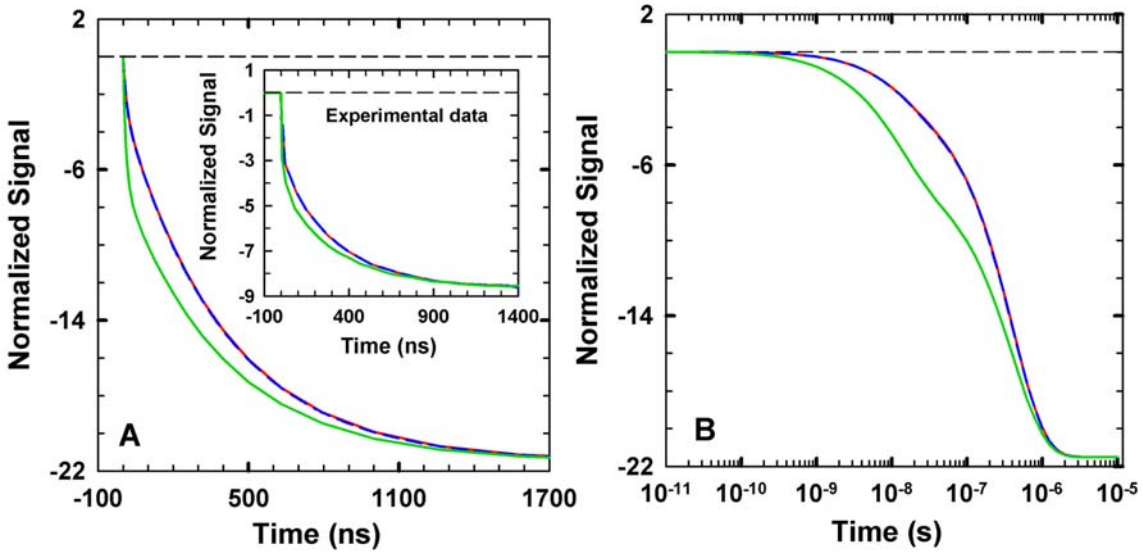


Figure 2.5 Relaxation kinetics observed at selected regions of the peptide

(A) Theoretical relaxation kinetics for N-terminus labeled (blue dashed line), Middle labeled (red) and C-terminus labeled peptides (green) following a T-jump from 278-288 K at the observation frequency of 1600 cm^{-1}. Following Gai and coworkers the time courses are normalized to 0 at time t=0. The scaling factors are 0.72 for middle labeled and 1.6 for C-terminus labeled peptides. The inset shows the original experimental data obtained by Gai and coworkers. (B) The relaxation kinetics shown in panel A are plotted on a logarithmic timescale. In both the panels the change in signal is scaled by 10^3.

than other regions of the peptide. This results in a fast phase that becomes sufficiently slow to get resolved in the kinetic experiments. Thus, it is the intricate balance between the sequence effects and the phenomenon of end fraying that gives rise to changes in the relative amplitudes of the fast and slow phases.

Similarly relaxation kinetics of middle labeled peptides obtained by Gai and coworkers in another study were also reported to be non-exponential and show marked differences depending on the probing frequency[98]. The results of T-jump simulations performed with the model for middle labeled and non-labeled peptides at different frequencies are shown in Figure 2.6. It can be seen from Figure 2.6B where the data is displayed on a logarithmic scale that the times courses are bi-exponential and the differences between them is the result of the changes in the relative amplitudes of the fast and slow phases. The amplitude of the fast phase increases from ~13% of the total amplitude at 1600 cm^{-1} to almost ~26% at 1635 cm^{-1}. Furthermore, Gai and coworkers reported that the relaxation times of the middle labeled peptides seem to depend on the magnitude of the T-jump. The non-exponential relaxation kinetics were fitted to stretched exponential functions with an instantaneous component. For T-jump sizes ranging from ~4 K to ~15 K to the same final temperature of 288 K the β values were found to vary between 0.75 to 0.85 whereas the instantaneous component contributed from ~15% to 30% to the full amplitude. The relaxation times exhibited a linear relation with the T-jump size with a slope of ~-10 ns K^{-1} and an intercept of ~390ns (Inset to Figure 2.7A).

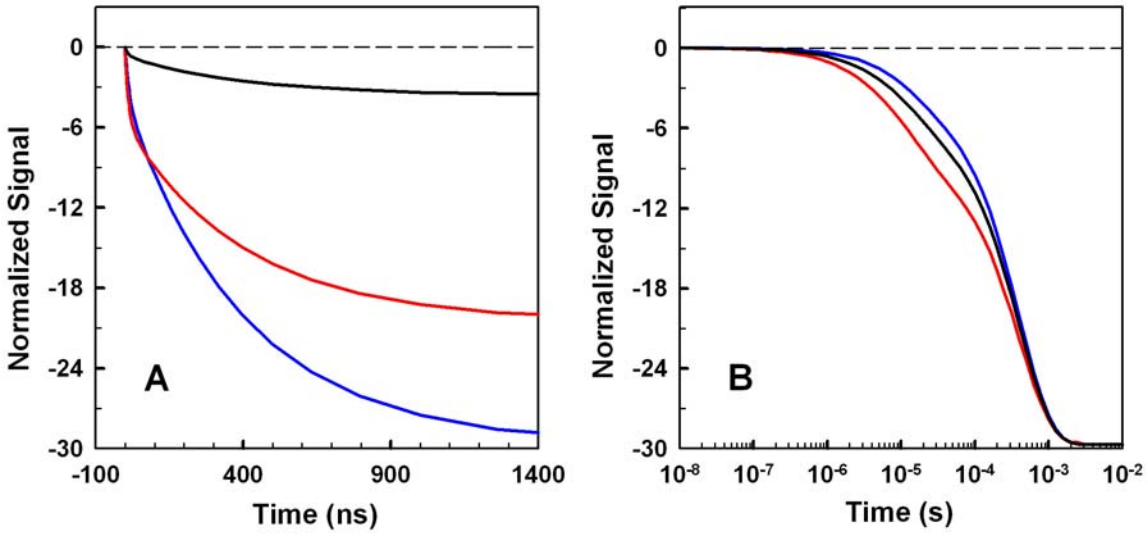


Figure 2.6 Relaxation kinetics observed at different probing frequencies

(A) Theoretical relaxation kinetics of peptides labeled in the middle region shown at 1600 cm^{-1} (blue) and 1635 cm^{-1} (red). Relaxation kinetics calculated at 1600 cm^{-1} for non-labeled peptides is shown in black. The theoretical time courses are calculated for a T-jump of ~10K to a final temperature of 293 K. The signals are normalized to 0 at time *t*=0. (B) The same relaxation kinetics shown in panel A are plotted on a logarithmic scale with scaling factors of 1.45 for peptides probed at 1635 cm^{-1} and 8.2 for non-labeled peptides probed at 1600 cm^{-1}. The signals are normalized to 0 at time t=0 and to the signal at 10 ms of middle-labeled trace at 1600 cm^{-1}. The change in signal is scaled by 10^3 in all cases.

To determine whether the kinetic model described here can reproduce this perturbation size-dependent kinetics, the same T-jumps as in experiments are simulated using equilibrium population calculated at the initial temperatures. The relaxation traces obtained at 288 K after starting from different initial temperatures

are shown in Figure 2.7A. In accordance with the experimental observation the apparent relaxation kinetics are found to become faster as the T-jump size increases. Again, the time courses plotted on the logarithmic timescale reveal biphasic relaxation and differences in the ratio of fast and slow phase amplitudes as well as in the ratio of relaxation times. To compare with experimental results, the apparent relaxation times are calculated by fitting the calculated kinetic traces to stretched exponential with β values of ~0.7. However, these apparent relaxation times yields a much lower slope of ~-2 ns K^{-1} when plotted against the difference between initial and final temperatures. When apparent relaxation times are calculated for a 40 K jump (blue trace in Figure 2.7A), they continue to decrease linearly with T-jump size. There are two possibilities for the discrepancy in the temperature dependence of the experimental and theoretical apparent relaxation rates: if the magnitude of the T-jumps in experiments is larger than that reported or the AGADIR parameters used in the model underestimate the effects of temperature on helix-coil transition. To investigate the latter possibility the calculations were repeated under conditions of higher helix stability by using a mean enthalpic contribution of –1.18 kcal/mol. The slope of the apparent relaxation times now increases and becomes very similar to that obtained from experiments. However, increasing the strength of the mean enthalpic contribution raises the T_m of the peptides from ~293 to 305 K due to which the final temperature of the T-jumps i.e. 288 K no longer falls in the transition region. Also increased helix stability does not reproduce the equilibrium behavior and other kinetic results correctly. These calculations reveal that in order to match the experimentally

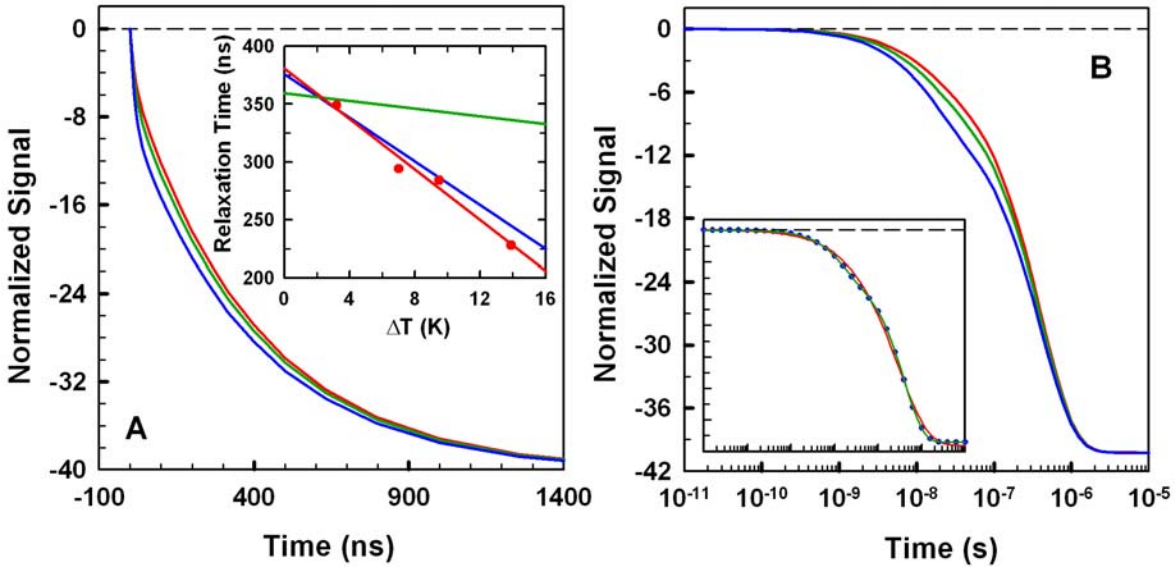


Figure 2.7 Relaxation kinetics of middle labeled peptides after T-jumps of different sizes

(A) Theoretical relaxation kinetics after a T-jump to a final temperature of 288 K from 285 K (red), 273 K (green), 248 K (blue). All the time courses are normalized to 0 at time t=0 and to the signal at 10 μs of the green trace (i.e. after a jump of 273 to 288 K). The scaling factors are 0.62 for red trace and 3.96 for the blue trace. The inset shows the dependence of relaxation times on the size of the T-jump. The experimental relaxation times are shown as red circles with a linear fit through the data (red line). The green line shows the dependence of apparent relaxation times on the magnitude of T-jump obtained from theoretical calculations using a mean enthalpic contribution of –1.04 kcal/mol. The blue line shows the dependence of apparent relaxation times on T-jump size obtained from the calculations when an increased mean enthalpic contribution of –1.18 kcal/mol is used. (B) The relaxation kinetics shown in (A) are plotted on a logarithmic timescale. In both (A) and (B) the change in signals is scaled by 10^3. The inset shows the calculated relaxation kinetics

(blue circles) with stretched exponential fit (red) with β=0.64 and double exponential fit (green).

observed slope of relaxation times the relative amplitude of the slow phase should be less than ~40%. Since kinetic traces are fitted to stretched exponential the apparent relaxation times are dependent on whether the fast or the slow phase dominates at the time at which the signal decays to $A_{final} - (A_{final} - A_{initial})/e$. In the experiments the instantaneous component that arises from the temperature dependence of the amide I spectrum will increase with the magnitude of the T-jump. Hence the amplitudes of the instantaneous component and the partially resolved fast phase tend to decrease the relative amplitude of the slow phase as the T-jump size increases. This will eventually result in an overestimation of the temperature dependence (i.e. increased slope) of the experimental relaxation rates on the magnitude of the T-jump.

The above results clearly demonstrate that the detailed kinetic model based on nucleation elongation theory is able to reproduce the experimentally observed dependence of relaxation rates on the specific sites probed, on the observation frequency as well as on the magnitude of the perturbation.

2.3.3 Analysis of the observed complex kinetics in helical peptides using 1-D free energy surface

As seen in the above section the complexities in helix-coil kinetics are consequences of the phenomenon of helix fraying and the heterogeneity of the sequence under consideration. Using a detailed microscopic model that involves solving more than 6000 differential equations allows understanding helix-coil kinetics on a quantitative

level. By projecting the free energy surface on a simple reaction coordinate, i.e. number of helical peptide bonds, the physical origin of observed complexity in helix-coil kinetics can be clearly understood. The 1-D free energy profile (Figure 2.8 A) consists of a barrier (~3.5 *RT*) separating two broad basins corresponding to two ensembles – one with coil conformations and very shorts helices (< 5 peptide bonds) and another with long stretches of one or more helices (with lengths > 5). Inset to Figure 2.8A shows that these basic features are preserved even in a free energy profile generated by a model in which the double sequence approximation is relaxed (i.e. allowing for than two helical segments to exist simultaneously on a peptide molecule). The features of 1-D free energy profile are very similar to the two-dimensional (2-D) energy landscape produced by atomistic simulations and projected onto two order parameters (helical content and radius of gyration) for a 21-residue peptide. The 2-D free energy landscape also shows a small free energy barrier separating the helical and the coil basins, each of which contain a diverse population of microstates with different helical content and radii of gyration.

Perturbations induced by temperature have two effects on the free energy surface. At first the helical basin is shifted upwards due to the decrease in stability, which is accompanied with the increase in the population of the coil ensemble (Figure 2.8B). Secondly at higher temperatures the helical well is shifted towards lower values of reaction coordinate as a result of the change in the distribution of helical lengths and number (Figure 2.8B). One of the predictions of helix-coil theory is that larger helices are formed at the expense of shorter ones during the course of the transition. In other words elongation of already existing helices is more preferable

than helix initiation at new sites. Formation of shorter helices from longer ones is a much faster process than formation of helices from coil conformations that requires crossing the nucleation barrier. Hence according to the nucleation-elongation mechanism one should expect to see two processes well separated in time in a T-jump relaxation experiment – the barrier crossing event i.e. equilibration between coil and helical ensembles and re-equilibration within the helical well between helices of varying lengths. Indeed these two events manifest in a biphasic relaxation as discussed in previous section and seen in inset to Figure 2.7B. The relaxations within the helical well are purely diffusive and hence the relaxation time of the fast phase will be proportional to the decrease in the average helical length. Since the free energy barrier is small the slow phase, which corresponds to relaxation between the helical and coil ensembles, also has a diffusive component. Due to this any changes in the average helical length affects the slow phase to a small extent. This picture of helix-coil transition is somewhat similar to the conformation diffusion process suggested by Gai and coworkers. However, they describe diffusion to occur in the coil basin with barrier-less transition into the helical region while in the above picture diffusion is occurring in the helical basin that is separated from the coil region by a small barrier.

For peptides labeled isotopically in different regions, the changes in their free energy surfaces due to increase in temperature are the same (Figure 2.8C). However, the decay of the signal at 1600 cm^{-1} strongly depends on the position of the labels. Since the extent of helix fraying is less in the N-terminal and middle regions, the rate of signal decay with respect to the average helical length is less for the labels placed

in these regions as compared to the more drastic decay of the C-terminal labels. Helix fraying is maximal in the C-terminal region and hence the C-terminal labels are more sensitive to the melting of long helices. When the changes in the probabilities at 278 K and 288 K (the T-jump temperatures) are weighted by the signal (Figure 2.8D) the largest differences are seen for peptides labeled in the C-terminal region, especially in the helical region indicating local relaxations. In accordance with this, the relaxation of the peptides labeled in the C-terminal region is accompanied by a large change in signal and thus exhibit a fast phase with larger relative amplitude (seen in Figure 2.5B). The relative magnitude of the fast phase amplitude is related to the relative height of the positive shoulder at higher values of the reaction coordinate while its relaxation time is proportional to the weighted distance between the negative peak and the maximum in the positive shoulder. In Figure 2.8D the height of the positive peak is related to the relative amplitude of the slow phase.

Figure 2.8E shows the differences in the probabilities at the initial and final temperatures of the T-jump for peptides probed in the middle region. The observation of faster apparent relaxation kinetics with the increase in the magnitude of the T-jump size, at first glance, appears counterintuitive as one would expect longer relaxation times as the difference between the initial and final temperatures becomes greater because the displacement between the distributions is largest when the initial temperature is lowest. The changes in the probability distribution after the T-jumps show greater intensities for both the negative helical peak and the positive peak in the coil region as the size of the T-jump increases. When the difference between the initial and final temperature (ΔT) is small the flux of molecules from the helical to the

coil basin is reflected in the negative peak and a nearly equal increase in the magnitude of the positive feature in the coil region. However, as ΔT increases, an increase in the positive coil peak intensity is not compensated by an equal increase in the helical negative peak. Instead a positive shoulder appears in the helical region with increasing intensity. The magnitude of the positive shoulder reflects the amount of redistribution in helical lengths that takes place after the T-jump. Since redistribution of helical lengths occurs by the fast process of helix propagation/de-propagation that does not require crossing the barrier, greater the magnitude of the positive shoulder greater is the amplitude of the fast phase. Hence, for greater ΔT there is a relative increase in the fast phase amplitude arising from increased local motions. The redistribution of probabilities after T-jumps when weighted by the change in signal at the probing frequency for the middle labeled peptides shows that the negative helical feature is narrower for lower initial temperatures (Figure 2.8F).
For local relaxation around the minima (average helical length) lesser displacement is required for lower ΔT and hence shorter relaxation times. Thus, the different redistribution of helical lengths at various ΔT results in subtle changes in the relative amplitudes and relaxation times of the fast and the slow phases and indirectly in shorter apparent relaxation times with increasing ΔT.

Hence, analysis of kinetic experiments on α-helical peptides using simple 1-D free energy projection demonstrate that nucleation-elongation theory is a valid mechanistic description for α-helix formation and the observed complex kinetics are inherent to helix-coil transition.

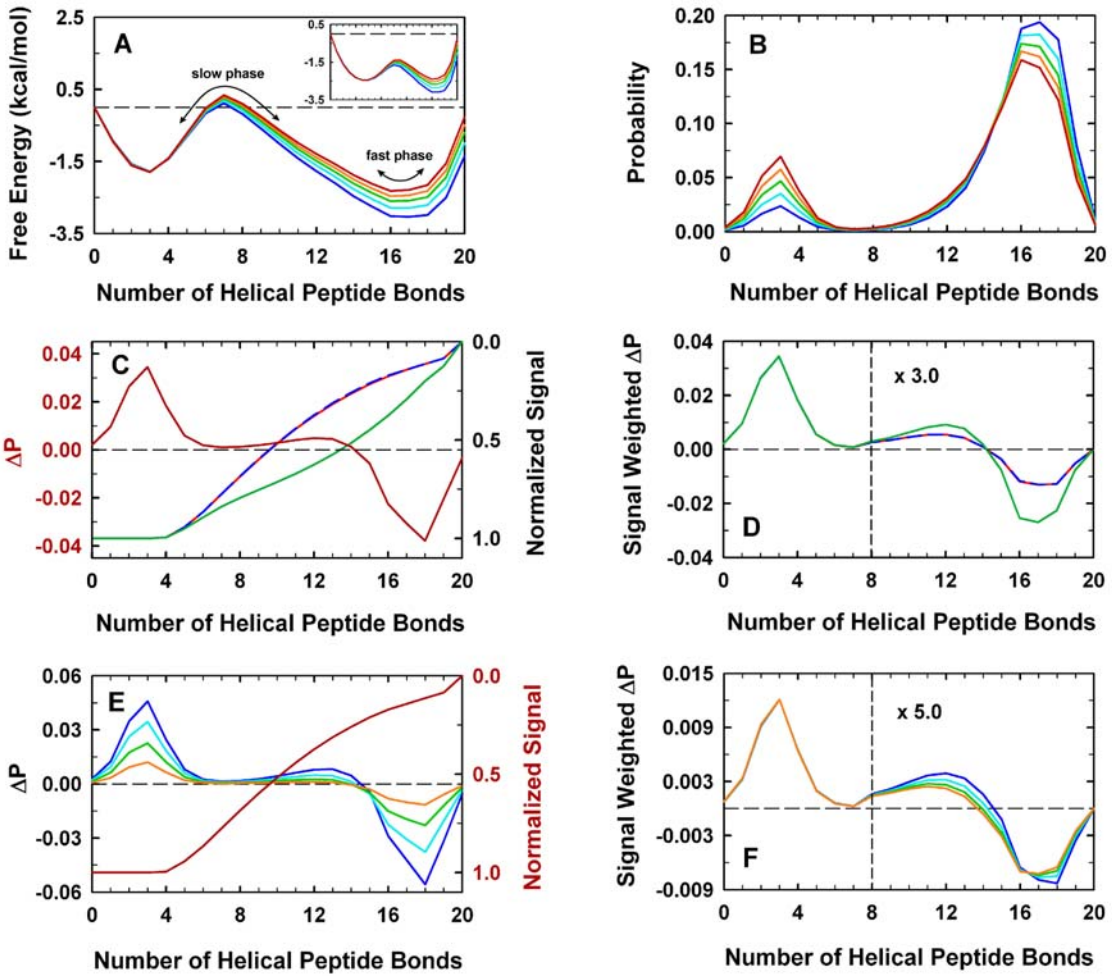


Figure 2.8 Characteristics of α-helix formation explained with 1-D projections of free energy surface

(A) Free energy projection as a function of number of helical peptide bonds at 273 K (blue), 278 K (cyan), 282 K (green), 285 K (orange), 288 K (red). The inset shows the free energy profile generated by the model with multiple sequence approximation. (B) Distribution of probabilities at the same temperatures with the same color code as in (A). (C). Redistribution of probabilities after a T-jump of 278 K-288 K i.e. P_{288}-P_{278} (left scale, dark red) and change in signal at 1600 cm^{-1} for N-terminal (blue dashed line), middle (red) and C-terminal (green) labels i.e. difference between the

signal as a function of helical peptide bonds and the signal of full length helix with 20 peptide bonds (right scale). (D) Redistribution of probabilities as shown in (C) but weighted by the change in signal for each labeled peptide. The same color code is maintained as in (C). The signal weighted probabilities are scaled by a factor of 3 beyond 8 helical peptide bonds to visualize the differences at higher magnification (E) Redistribution of probabilities after T-jumps of various sizes i.e. P_{288}-P_{273} (blue), P_{288}-P_{278} (cyan), P_{282}-P_{278} (green), P_{285}-P_{278} (orange) (left scale) and change in signal at 1600 cm^{-1} for the middle labeled peptides as shown in (C) (right scale, dark red). (F) Redistribution of probabilities after T-jumps of various sizes as shown in (E) but weighted by the change in signal at 1600 cm^{-1} for the middle labeled peptides. The signal weighted probabilities are scaled by a factor of 5 beyond 8 helical peptide bonds to visualize the differences at higher magnification.

2.3.4 Investigation of helix nucleation with the detailed kinetic model

As mentioned in the earlier sections, the complexity of helix-coil kinetics arises mainly from the balance between the favorability of nucleation and propagation in different regions of the peptide. The timescales of helix nucleation obtained from T-jump experiments[67,70,93,105] are at least 6 orders of magnitude higher than that suggested by pH-jump stopped flow experiments[94].To investigate whether the model can explain this discrepancy the nucleation process in the model is modified in different ways to simulate the same set of T-jump experiments mentioned above. One way of modifying helix nucleation is to alter the cooperativity of the helix-coil transition. This is achieved by changing the entropic cost of fixing peptide bonds in helical angles, which is compensated by the changes in the mean enthalpic contribution to keep the T_m of the peptides constant. When several combinations of the entropic cost and the mean enthalpic contributions are used results viz. site-dependent, probe-dependent and perturbation size-dependent kinetics, similar to those in previous sections are obtained. This required readjusting the intrinsic rate of peptide bond rotation for each calculation to reproduce the experimental scale. In an alternative approach, the size of the nucleus is altered from five peptide bonds used in the model discussed above. The size of the nucleus controls the relative timescales of the fast and slow phases. When a nucleus of seven peptide bonds is allowed to form the slow phase becomes slower by ~100 fold than the fast phase (as compared to ~10 fold in the model described in the previous sections and seen in experiments). On the other hand, a nucleus of three peptide bonds produces the two phases that tend to cluster together. This result is similar to the one obtained when helix is pre-nucleated

and only allowed to propagate (i.e. no cost in entropy considered for residues involved in the nucleus, the statistical weight for a pre-nucleated species is then given by $w_n.(w_{in})^{h-s_n}.(w_{bb})^{h-s_n+1}.w_c$, where s_n is the size of the nucleus). It is found that the nucleus of five peptide bonds that corresponds to the nucleus involving four residues and N- and C-caps used in AGADIR and for all calculations of previous sections is optimal. These results demonstrate that the timescale of helix nucleation depends on how the different energetic contributions are compensated with each other. Even with the 100-fold slower slow phase for a peptide with a longer nucleus the timescale of helix nucleation is ~tens of μs, i.e. far smaller than the 100-millisecond relaxation observed in stopped flow CD studies. These results thus support T-jump experiments suggesting the global helix folding/unfolding event to take place in the microsecond timescale and the possibility of artifactual effects in denaturant-jump stopped flow studies.

Chapter 3: Calculation of helix-coil kinetics as diffusion on 1-D free energy surface

3.1 Introduction

As discussed in Chapter 2, the kinetic nucleation-elongation model for α-helix formation has been thoroughly tested for quantitative analysis of experiments. In this chapter time courses are generated from diffusive kinetics on a 1-D mean force potential calculated as a function of number of helical peptide bonds. The issue of whether this approach can reproduce the probe- and T-jump size-dependent kinetics observed in experiments and predicted by the detailed model is investigated here. Furthermore, the physical basis of length dependence of relaxation kinetics is also explained.

3.2 Model and Methods

3.2.1 Calculation of 1-D free energy functional

A potential of mean force as a function of the number of helical peptide bonds (h) is calculated using double sequence approximation as follows:

$$\left.\begin{aligned} F(0) &= -RT\ln(1) \\ F(h) &= -RT\ln\left(\sum_{i=1}^{h}\sum_{j=1}^{n+1-i} w_{ij}\left(1+\sum_{q=i+j+1}^{n+1-p} w_{pq}\right)\right) \end{aligned}\right\} \quad \text{..... (3.1)}$$

where n is the total number of helical peptide bonds, w_{ij} and w_{pq} are the statistical weights of helical segments having lengths i and p (p=h-i) and starting at positions j and q respectively. The statistical weights are obtained from the equilibrium

AGADIR model based on helix-coil theory. For a peptide with 21 residues *h* takes the value from 0 to *n*=20 resulting in a 1-D free energy profile (a discrete vector *F* with *n+1* points). With double sequence approximation there are a total of 6196 conformations having single and double helical segments.

All the calculations in this work are performed using Matlab 6.5, Microsoft Visual C++ and the CVODE program provided by Eric Henry[103].

3.2.2 The diffusion model

Lapidus *et al*. numerically solved the 1-D diffusion equation given by Szabo, Schulten and Schulten by approximating it to rate equations that describe time-dependent probabilities of the to and fro passage of the molecules along the reaction coordinate[10]. Using the same method here the following equation is numerically solved.

$$\frac{dx}{dt} = Rx \qquad \text{..... (3.2)}$$

x is a vector of *n+1* probabilities (nearest neighbor along the reaction coordinate) and *R* is the rate matrix with dimension (n+1)×(n+1). The elements of *R* are given by

$$2R_{i,i+1} = D_i \frac{p_i}{p_{i+1}} + D_{i+1} \quad \text{(upper diagonal)}$$

$$2R_{i+1,i} = D_i + D_{i+1} \frac{p_{i+1}}{p_i} \quad \text{(lower diagonal)}$$

$$R_{1,1} = -R_{2,1} \quad \text{(diagonal elements)}$$

$$R_{i,i} = -R_{i+1,i} - R_{i-1,i} \quad (2 \le i < n)$$

$$R_{n+1,n+1} = -R_{n,n+1}$$

where $p_i = p_i(h_i) = \exp(-F(h_i)/RT)/Q$ and $D_i = D(h_i)/(\Delta h)^2$. The rate matrix is diagonalized to obtain eigenvalues and eigenvectors. Relaxation kinetics following a

T-jump is simulated as diffusion on 1-D free energy surface computed at the final temperature after the jump. The amplitude corresponding to each eigenvector is obtained using probability distribution at the initial temperature. The diffusion coefficient *D* is assumed to be independent of the position along the reaction coordinate.

3.3 Results and Discussion

3.3.1 Comparison between predictions of 1-D diffusive model and detailed kinetic model: Site-specific relaxation kinetics of α-helical peptides

Free energy surface for the peptide sequence Ac-YGSPEAAAKAAAAKAAAA-r-NH_2 obtained from Equation 3.1 as a function of number of helical peptide bonds *H* is shown in Figure 3.1A. The set of peptides having the above sequence are ^{13}C labeled at the carbonyls of alanines either at the N-terminus, middle or C-terminus region. When observed experimentally after subjecting all of the three labeled peptides to the T-jump of 10 K to the same final temperature of 288 K, relaxation kinetics of C-terminally labeled peptide was faster than the N-terminally or middle labeled peptides. As mentioned in Chapter 2 and shown in Figure 3.1C (colored lines) this kinetic behavior is successfully reproduced with the detailed kinetic model. Relaxation kinetics calculated as diffusion on the free energy surface at the final temperature and weighted by the signal decay of each respective labeled peptide (Figure 3.1B) are as dotted lines in Figure 3.1C. The results of the diffusive model are in agreement with the relaxation kinetics calculated from the detailed model for N-terminally and middle labeled peptides. However, for peptide labeled in the C-

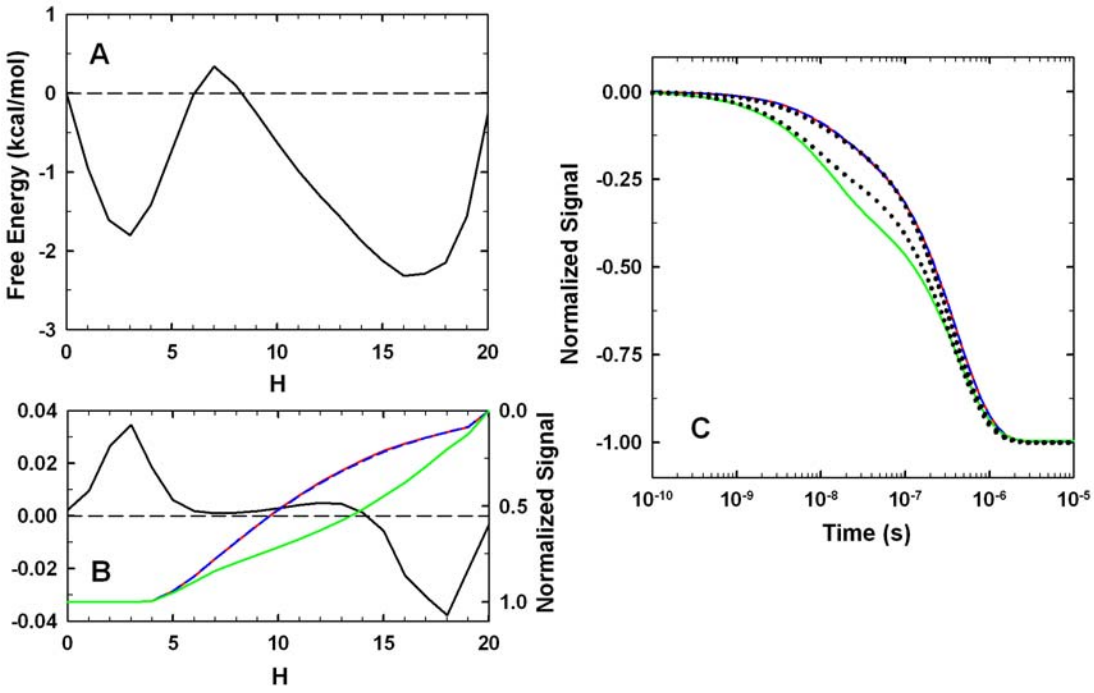


Figure 3.1 Comparison between predictions of 1-D diffusive model and detailed kinetic model: Site-specific relaxation kinetics of α-helical peptides

(A) 1-D Free energy surface at 288 K as a function of number of helical peptide bonds *H*. (B) Redistribution of probabilities after a T-jump of 278 K-288 K i.e. P_{288}-P_{278} (left scale, black) and signal decay of amide I band calculated with the detailed model at 1600 cm^{-1} for N-terminal (blue dashed line), middle (red) and C-terminal (green) labels as a function of *H* (Same as Figure 2.8C). (C) Relaxation kinetics calculated with the detailed kinetic model for a T -jump from 278 to 288 K for peptides labeled at the N-terminus (blue dashed line), middle (red), or C-terminus (green) region. The relaxation kinetics calculated as diffusion on free energy surface in (A) are shown as dotted lines. To facilitate comparison the signals are normalized.

terminus, the discrepancy between the two calculations arises mainly due to underestimation of the fast phase amplitude by the diffusive model. The reason for this deviation is the rough approach used to calculate the signal decay along with the heterogeneous nature of the peptide. At each value of H there are conformations with the same number of helical peptide bonds. However, the probability distribution of these conformations is not uniform due to the helix fraying effect and the range of intrinsic helical propensities of the amino acid residues in the hetero-peptide used by Gai and coworkers. In addition amino acid residues with low helical propensities are clustered in the N-terminal region. At $H=H_{max}$ there is only one species possible with probability of 1 but as H decreases the probability distribution becomes more bell-shaped. Hence the time dependent probability distribution at each value of H will not be constant along the peptide sequence. The time courses resulting from the detailed model show the time evolution of the weighted signal. For the diffusive model it is required to know the decay of signal for each labeled peptide as a function of H, which is not easy to calculate especially for properties not directly related to the average number of helical peptide bonds. An approximate solution is obtained by representing the group of conformations at each H as a microcanonical ensemble. It is the intrinsic error in assuming microcanonical ensembles for a heteropolymer that is likely to cause the deviation in the prediction of relaxation kinetics of C-terminally labeled peptides by the diffusive model from that of the detailed model. As opposed to labeled alanines at the C-terminal extreme the alanines labeled in the N-terminal region are preceded by strong capping motifs due to which the fraying effect is more pronounced in the C-terminal region (see inset to Figure 2.3). There are more

contributions of local motions in the relaxation kinetics of the peptides labeled in the C-terminal region thus giving rise to relatively larger amplitude of the fast phase. Despite the approximate calculation of the signal decay the relaxation behaviors of N-terminally and middle labeled peptides predicted by the diffusive model are in remarkable agreement with those calculated by the detailed model.

3.3.2 Comparison between predictions of 1-D diffusive model and detailed kinetic model: T-jump size-dependent relaxation kinetics of α-helical peptides

T-jump of 3 K and 20 K are simulated for the middle labeled peptides of Gai and coworkers by the detailed model and the diffusion model. The probability of hydrogen bonded carbonyls as a function of temperature is shown in Figure 3.2A. The two initial and the final temperatures of the T-jump are also indicates. Figure 3.2B shows the changes in the probabilities occurring as a result of 3 K and 20 K T-jumps. The relaxation kinetics predicted by the detailed model (continuous lines in Figure 3.2C) exhibit bi-exponential behavior as previously mentioned. Relaxation time courses calculated as diffusion on the free energy surface at the final temperature (Figure 3.1A) are also biphasic with similar ratio of the relative amplitudes of the fast and the slow phases. There are small discrepancies between the detailed and the approximate time courses at the beginning of the fast phase (i.e. ~10 ns) and the end of the slow phase. These discrepancies produce an error of ~10% in the calculation of the apparent relaxation times. However the overall agreement between the exact calculation and the diffusion model is very encouraging and provides support for the use of 1-D free energy surfaces in interpreting helix-coil kinetics.

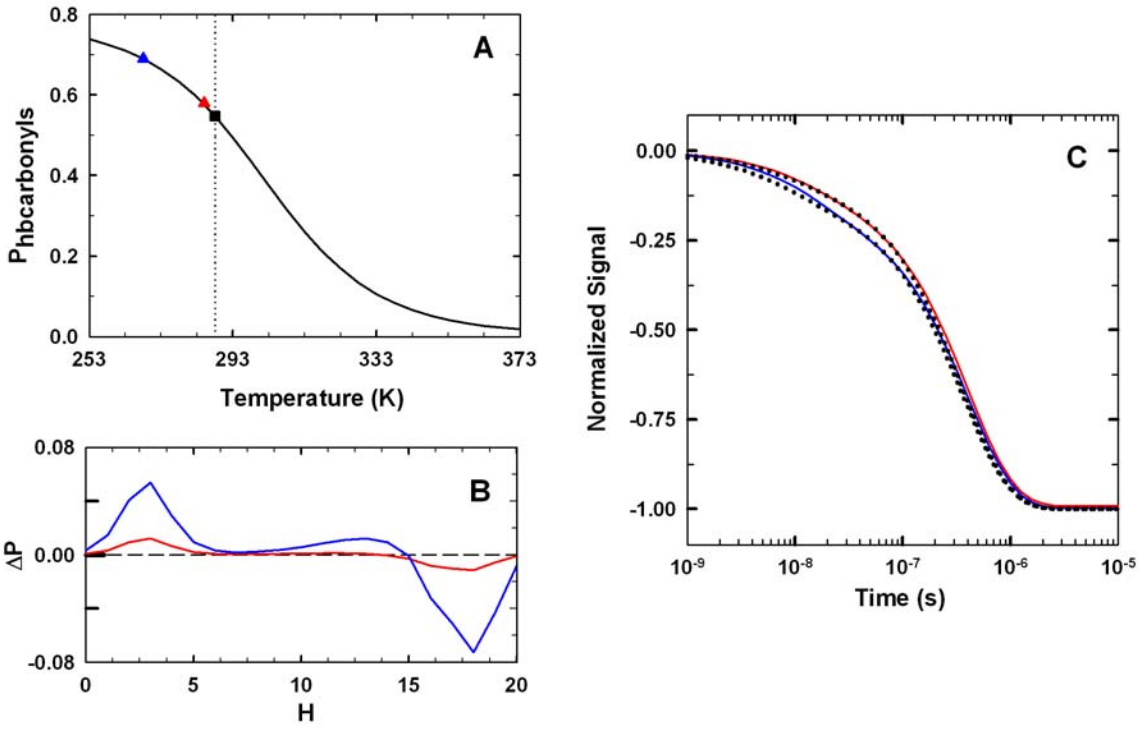


Figure 3.2 Comparison between predictions of 1-D diffusive model and detailed kinetic model: T-jump size-dependent relaxation kinetics of α-helical peptides

(A) Probability of hydrogen bonded carbonyls as a function of temperature. The black square indicates the probability at the final temperature (288 K) after the T-jump whereas the blue (268 K) and red triangles (285 K) indicate the probabilities at the initial temperatures after the T-jumps. (B) Redistribution of probabilities (i.e. P_{288}-P_{268} (blue) and P_{288}-P_{285} (red)) after T-jumps of 20 K and 3 K. (C) Relaxation kinetics calculated with the detailed model after T-jumps of 20 K (blue) and 3 K (red). The relaxation kinetics calculated as diffusion on free energy surface of Figure 3.2B are shown as dotted lines. For comparison the signals are normalized.

3.3.3 Comparison between predictions of 1-D diffusive model and detailed kinetic model: Length dependence of relaxation kinetics

Polymer physics theories have predicted relaxation times to scale with the size of the protein. To investigate whether relaxation times helical peptides follow the same length dependence peptides with varying numbers of repeating units are used in the model calculations (i.e. Ac-YGG$(KAAAA)_n$G-NH_2). These peptides have similar sequence to the one used in previous calculations in which SPE takes the place of GKA and D-Arg in place of G at the C-terminal. Relaxation kinetics for the above peptides with n = 2,3,5,and 7 are calculated by the detailed kinetic and the diffusion model. Comparison between the two set of calculations is carried out at two conditions: by simulating T-jumps of the same magnitude to the same final temperature for all peptides and to the apparent T_m of each peptide. In Figure 3.3A the probability of hydrogen-bonded carbonyls as a function of temperature depicts the theoretical thermal denaturation of the peptides. The conditions before and after a 20 K jump to a final temperature of 303 K are also indicated. It can be seen from Figure 3.3B that the free energy surfaces at 303 K of the peptides with lengths 16 (n=2), 21 (n=3), 31(n=5), 41(n=7) residues differ significantly in their stability. The peptide with n=2 has its helical minimum at ~11 helical peptide bonds and shows marginal stability at 303 K while the 7-repeat peptide has its helical minimum at H~38 and stabilized by ~10 RT over the coil minimum. The exact and approximate time courses calculated by the detailed and diffusive models respectively predict biphasic relaxation for all the peptides for T-jumps of 20 K to the final temperatures of 303 K.

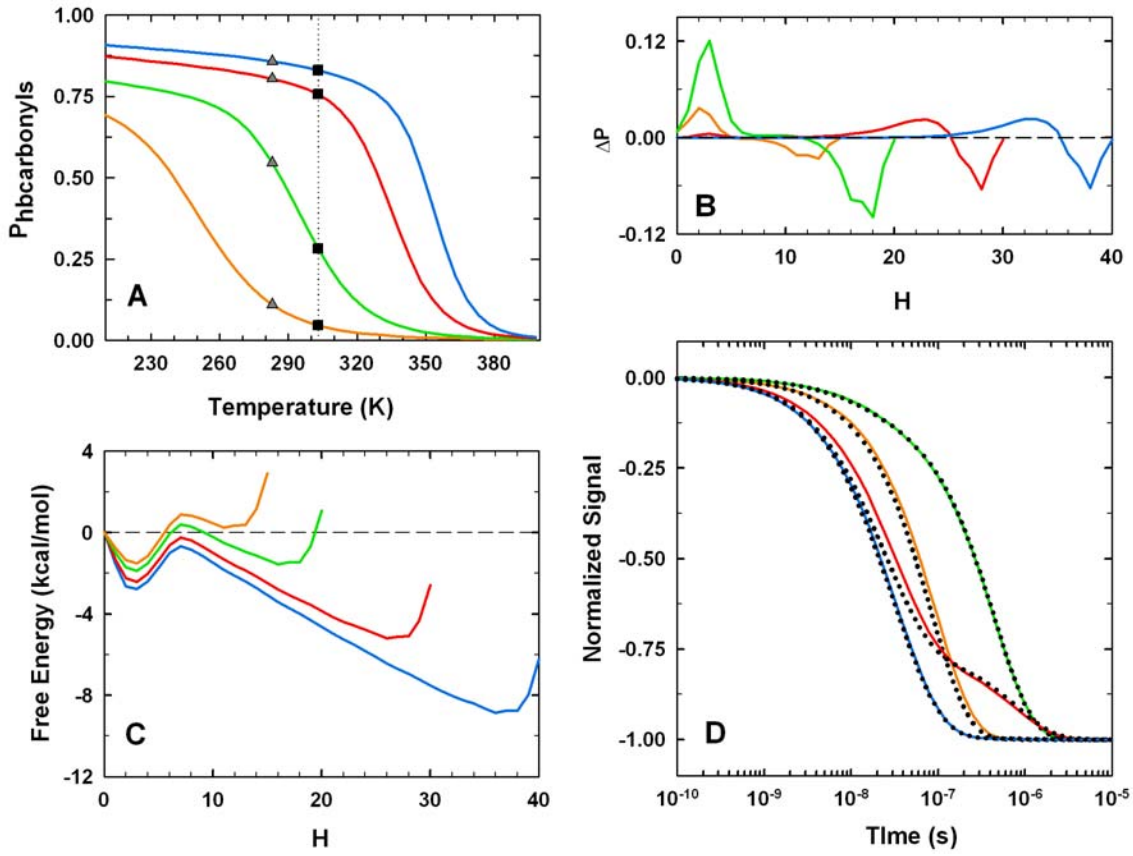


Figure 3.3 Length dependence of relaxation kinetics after T-jump of same size to the same final temperature

(A) Probability of hydrogen bonded carbonyls as a function of temperature calculated for peptides with sequence Ac-YGG(KAAAA)$_n$G-NH$_2$ with n=2 (orange), n=3 (green), n=5 (red), n=7 (blue). The squares indicate the probability at the final temperature and gray triangles indicate the initial temperatures for the four peptides (B) Redistribution of probabilities as a function of H for the four peptides after a 20 K T-jump to the same final temperature of 303 K. (C) Free energy surfaces of the four peptides at 303 K (D) Relaxation kinetics calculated with the detailed model (colored lines) for the four peptides after a 20 K T-jump and relaxation kinetics calculated as

diffusion on respective free energy surfaces in (C) are shown as dotted lines. The color coding in all other panels is the same as (A).

For the T-jumps at 303 K the apparent relaxation becomes faster as the length of the peptide increases. However, peptide with n=2 does not follow this rule and shows faster apparent relaxation than for peptide with n=3. These results can be explained on the basis of stability of the peptide at the final temperature and probability distributions at the initial and final temperature. If a hot T-jump is exerted at a final temperature that is much below the apparent T_m of the peptide, dynamics is dominated by the redistribution of helical lengths. In this case redistribution of probabilities shows very little intensity of the positive peak in the coil region as compared to the intensities of the helical peaks (for n=5 and 7 in Figure 3.3C). As a result, relaxation is dominated by the fast phase and the apparent relaxation is increasingly faster. For a hot T-jump much above the T_m in case of peptide with n=2, the peptide is marginally stable such that the unfolding barrier is very small and there is very little change in the flux. This makes the slow phase relatively faster and gives rise to a faster apparent relaxation. For the peptide with n=3, the initial and final temperatures fall before and after its apparent T_m respectively such that there is a large flux of molecules crossing the barrier resulting in the slowest relaxation. This is evident from the increased intensity of the positive peak in the coil region and negative peak in the helical region in Figure 3.3C. The agreement between the time courses calculated with the detailed model and the diffusion model is clearly

noticeable. The diffusion model reproduces the ratio of the relative amplitudes as well as of the relaxation times of the fast and the slow phases.

Figure 3.4 shows the calculations carried out by the detailed and the diffusive model for T-jumps simulated to the apparent T_m of each peptide. The apparent T_m*'s* correspond to the isostability conditions at which the flux of molecules crossing the barrier is similar for all the peptides (Figure 3.4B and C). Under this condition the height of the nucleation barrier increases with the length of the peptide, which gets reflected in progressively slower apparent relaxation for longer peptides with higher T_m's (Figure 3.4D). The time courses generated from the diffusion model are in remarkable agreement with those obtained from exact calculation. The relaxation kinetics of all peptides are predicted to be biphasic with similar ratio of the relative amplitudes of the fast and the slow phases.

All the above calculations using the detailed model are performed with a temperature independent pre-exponential (k_o) of 2.5 X 10^8 s^{-1} at 1 centepoise. The diffusive kinetic calculations are carried out with a constant diffusion coefficient of 0.57 X 10^9 s^{-1} for all the peptides with different sequences and lengths.

As seen from the above sections and Chapter 2 1-D free energy surfaces generated from nucleation-elongation models and empirical force fields are sufficiently accurate to explain the complexities observed in kinetic experiments. Here, it is shown that diffusion on such highly simplified 1-D free energy surfaces can reproduce all the kinetic behaviors viz. dependence of relaxation times on T-jump size, specific region probed and chain length with adequate accuracy as those predicted by the detailed kinetic model. From the diffusive model the diffusion

coefficient corresponds to a timescale of ~2 nanoseconds. This is in very good agreement to the timescale of ~4 nanoseconds for the elementary peptide bond rotation obtained from the detailed kinetic model. For both the models the

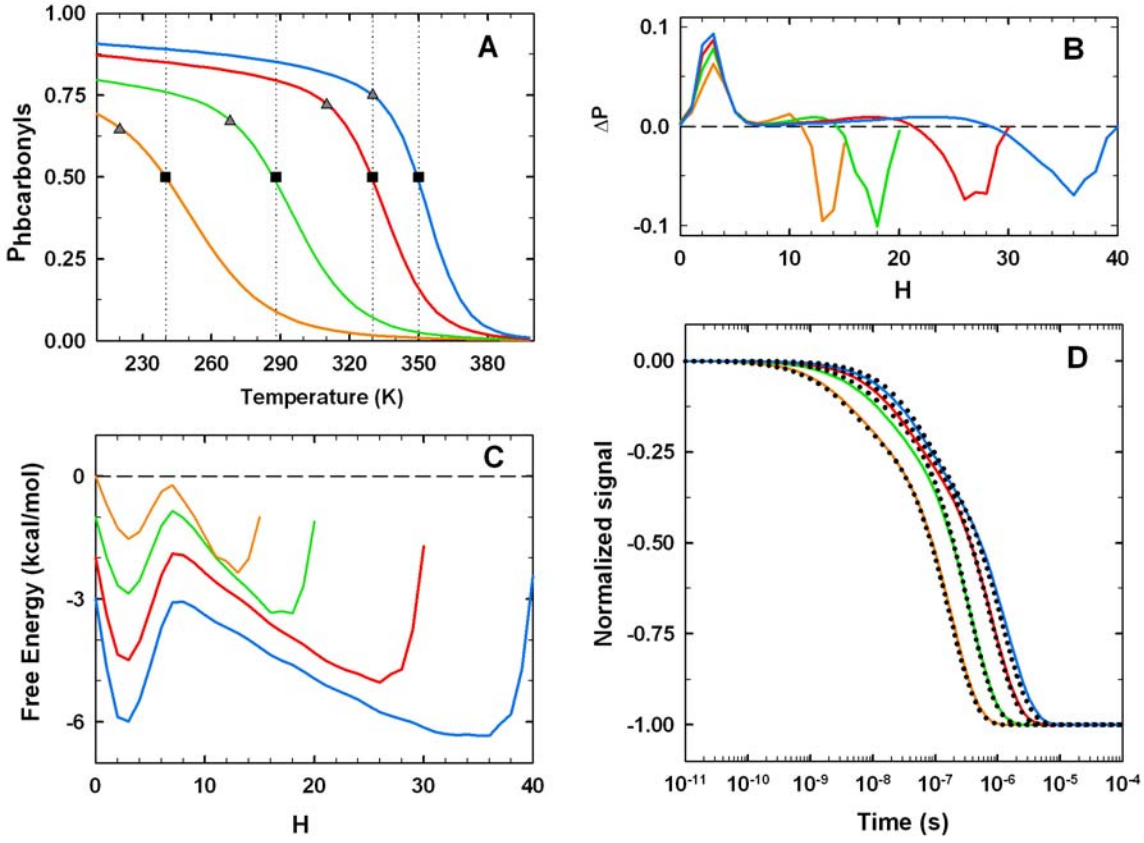


Figure 3.4 Length dependence of relaxation kinetics after T-jump of same size to different apparent T_m's

All the panels are the same as in Figure 3.3 except that a 20 K T-jump is simulated to the final temperature corresponding to the apparent T_m of each peptide. The color coding is maintained. The free energy surfaces in (C) are shifted on the y-axis for clarity.

temperature dependence of the pre-exponential factor and the diffusion coefficient is derived from the temperature dependence of viscosity of water. It is remarkable that the timescales for the dynamic motions obtained from two different approaches - one that involves solving several thousand differential equations and the other that models kinetics as diffusion on 1-D free energy surface, are of the same order. This demonstrates that the number of helical peptide bonds, *H*, is a robust reaction coordinate for α-helix formation. From the above calculations it is shown that a common diffusion coefficient is sufficient for peptides differing in amino acid sequence or lengths. This implies that the 1-D free energy surface is able to capture the roughness arising from the differences in the amino acid sequence. If α-helix formation is supposed to be a sequential process then the dependence of diffusion coefficient on the reaction coordinate should be expected. However, the diffusion coefficient used in the above calculations is independent of *H*. This supports the description in which α-helix forms in a parallel process where nucleation can occur at several sites followed by multi-step propagation. The above results also demonstrate that using 1-D free energy surfaces can tremendously simplify the analysis of complex kinetic experiments of α-helix formation. Furthermore, they provide strong encouragement to use 1-D free energy surface approach for the analysis of protein folding experiments.

Chapter 4: Formulation of a mean field 1-D free energy surface model of protein folding

4.1 Introduction

Analysis of available experimental data on folding of several small single-domain proteins has unraveled some general trends in thermodynamic and kinetic behaviors[15,69]. The amino acid sequences of natural single domain proteins seem to be selected to have sufficiently low energetic frustration. As a consequence these proteins fold fast with rates spanning from microseconds to seconds. Their folding rates can be largely determined from length as well as from gross topological features[45,106,107]. The equilibrium unfolding properties characterized by $\Delta H_{F\text{-}U}$, $\Delta S_{F\text{-}U}$ and ΔC_p exhibit linear scaling with protein length (Figure 4.1). In order to explore the physical basis of the connection between the experimental quantities and the inherent properties of protein- its length and structure, it is necessary to gain a quantitative understanding of the folding process. Towards this end, 1-D free energy surfaces can serve as a good starting point to investigate the interplay between dynamic, energetic and structural contributions[106]. Moreover, the presence or absence of barriers on 1-D free energy profiles can help in distinguishing a range of folding regimes. The success achieved in explaining the underpinnings of helix-coil kinetics provides support and encouragement for using 1-D free energy functional for analyzing protein folding experiments.

However, the simplified projection on 1-D should account for, though implicitly, the complexities of the multi-dimensional free energy landscape. Besides,

for protein folding, lack of a precise force field and detailed knowledge of entropic factors makes the formulation of an adequate 1-D free energy profiles with statistical mechanics very challenging. Here, this problem is addressed by employing a mean field approach to derive an approximate 1-D free energy functional from the combination of simple mathematical functions that model the evolution of stabilization energy/enthalpy and entropy as folding progresses. The model is suitable for describing folding behaviors ranging from two-state to completely downhill. However, the current model does not address folding regimes involving three-states.

4.2 Model and Methods

4.2.1 Description of thermodynamics

Earlier Zwanzig-like models have used discrete parameters such as number/fraction of incorrect residues or ordered residues as reaction coordinates[11,52,53,55]. The use of these quantities has facilitated the calculation of conformational entropy just by combinatorial counting. Here, free energy is expressed in terms of a quantity called 'nativeness' (n) that is, to some degree, a continuous version of Zwanzig's parameter $(N\text{-}S)/N$ (where N is the total number of residues and S is the number of incorrect residues). Nativeness is defined as the average probability of finding a residue in native conformations. This definition fits the mean field description of the model and permits the calculation of conformational entropy in a straightforward manner.

Conformational entropy as a function of n ($\Delta S^{conf}(n)$) is simply the entropy of mixing obtained from Gibb's theorem:

$$\Delta S_{res}^{conf}(0) = \Delta S_{res}^{n=0} = S_{res}^{n=0} - S_{res}^{n=1} \quad \text{..... (4.1)}$$

$$\Delta S_{res}^{conf}(n) = -R\left[n\ln(n) + (1-n)\ln(1-n)\right] + n\Delta S_{res}^{n=1} + (1-n)\Delta S_{res}^{n=0} \; for\; n > 0 \quad \text{..... (4.2)}$$

$$\Delta S^{conf}(n) = N\Delta S_{res}^{conf}(n) \quad \text{..... (4.3)}$$

where $\Delta S_{res}^{n=0}$ is the cost in conformational entropy of fixing a residue from all non-native to completely native conformations. Since n=1 is the reference state, $\Delta S_{res}^{n=1}$=0. The total conformational entropy $\Delta S^{conf}(n)$ is obtained by scaling $\Delta S_{res}^{conf}(n)$ to the total number of residues N. The stabilization energy of folding ΔH^0 is assumed to be an exponential function of n:

$$\Delta H^0(n) = \Delta H_{res} N\left[1 + \left(\exp(k_{\Delta H} n) - 1\right) \Big/ \left(1 - \exp(k_{\Delta H})\right)\right] \quad \text{..... (4.4)}$$

where ΔH_{res}^0 is the stabilization energy per residue . Free energy can then be expressed as

$$\Delta G(n) = \Delta H^0(n) - T\Delta S^{conf}(n) \quad \text{..... (4.5)}$$

4.2.2 Modeling temperature effects

To simulate thermal denaturation experiments temperature effects on solvation entropy are modeled by including a heat capacity function that also decays exponentially with n:

$$\Delta C_p(n) = \Delta C_{p,res} N\left[1 + \left(\exp(k_{\Delta C_p} n) - 1\right) \Big/ \left(1 - \exp(k_{\Delta C_p})\right)\right] \quad \text{..... (4.6)}$$

where $\Delta C_{p,res}$ is the change in the heat capacity of folding per residue. The total entropy can then be expressed as:

$$\Delta S(T,n) = \Delta S^{conf}(n) + \Delta C_p(n)\ln(T/385) \quad \text{..... (4.7)}$$

At 385 K the polar and apolar solvation terms are counterbalanced so that the total entropy change of unfolding ($\Delta S_{U\text{-}F}$) mainly reflects ΔS^{conf}. 385 K is also the convergence temperature suggested by Robertson and Murphy for $\Delta S_{U\text{-}F}$ obtained from DSC data of 53 proteins[41]. At convergence temperature the $\Delta S_{U\text{-}F}$ normalized with respect to size is expected to approach a single value for all proteins. Although for the protein dataset used by Robertson and Murphy no clear convergence behavior was observed the correlation coefficients between $\Delta S_{U\text{-}F}$ at various temperatures and number of protein residues *N* plotted as a function of temperature reaches an asymptotic value around 385 K (Figure 4.1D). This supports the assumption that $\Delta S_{U\text{-}F}$ at 385 K primarily corresponds to the change in conformational entropy. Figure 4.1 shows the size scaling behavior of thermodynamic parameters using the dataset of Robertson and Murphy.

The total change in enthalpy as a function of *n* and temperature is then determined with mid-point temperature as the reference as follows:

$$\Delta H(T,n) = \Delta H^0(n) + \Delta C_p(n)(T - T_m) \quad \text{..... (4.8)}$$

The exponents ($k_{\Delta H}$, $k_{\Delta C_p}$) of the ΔC_p and ΔH functions controls their curvatures and thereby the values at the top of the barrier, which partitions ΔC_p and ΔH into their respective activation values of folding and unfolding for two state systems.

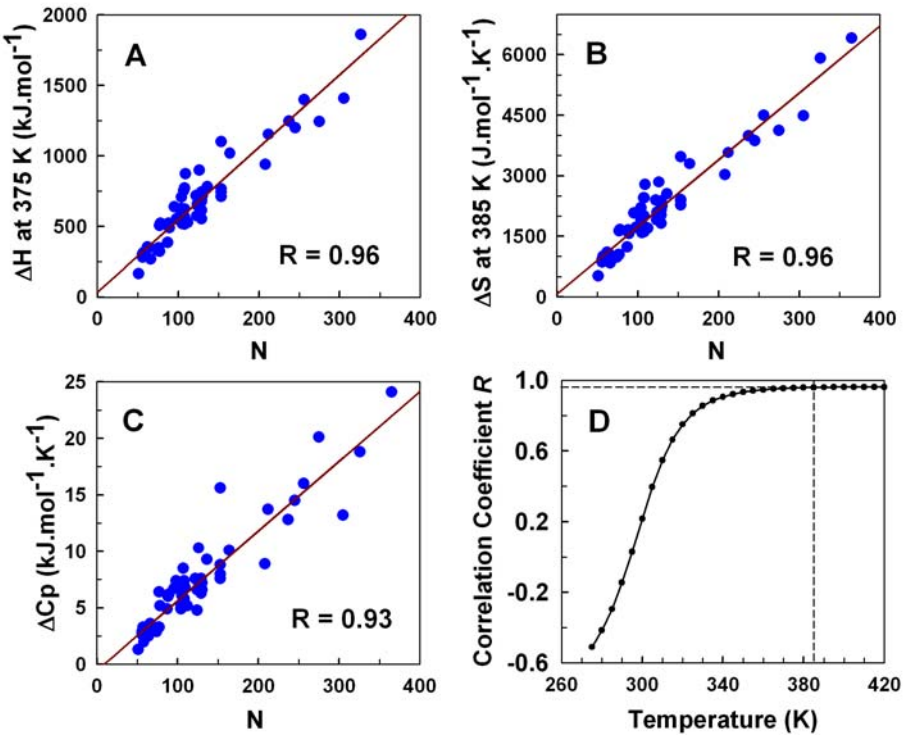


Figure 4.1: Correlation of thermodynamic parameters with protein size (number of residues *N*). (Reproduced using the data in ref. 41)

Linear regression is shown by solid red line. For (A) ΔH of unfolding at 102°C vs. N slope and intercept of 5.13 $kJ.mol^{-1}.res^{-1}$ and 35 $kJ.mol^{-1}.K^{-1}.res^{-1}$ respectively is obtained; for (B) ΔS of unfolding at 112°C vs. N, slope and intercept are 16.6 $J.mol^{-1}.K^{-1}.res^{-1}$and 75 $J.mol^{-1}.K^{-1}$ respectively; for (C) heat capacity changes ΔC_p vs. N slope and intercept are 0.062 $kJ.mol^{-1}.K^{-1}.res^{-1}$ and –0.5412 $kJ.mol^{-1}.K^{-1}$. Respective correlation coefficients are indicated in the individual plots. (D) Correlation coefficient of ΔS of unfolding vs. N as a function of temperature. No significant improvement in correlation coefficient is obtained beyond 385 K supporting the concept of convergence temperature.

Free energy as a function of n and temperature is directly obtained from Equations 4.7 and 4.8:

$$\Delta G(T,n) = \Delta H(T,n) - T\Delta S(T,n) \quad \text{..... (4.9)}$$

The above description of temperature dependence of folding is in accordance with that used earlier for thermal denaturation experiments. The magnitude of the free energy barrier is determined solely by the exponent ($k_{\Delta H}$) of the stabilization energy function without any adjustment in the total entropy or enthalpy.

4.2.3 Modeling chemical denaturation effects

Destabilization caused by chemical denaturation is assumed to be linearly dependent on the concentration of chemical denaturant [d]. The free energy functional is expressed as

$$\Delta G(d,n) = \Delta H^0(n) - T\Delta S(n) - mE_d(d - d_m) \quad \text{..... (4.10)}$$

where $\Delta H^0(n)$ is the stabilization energy at the experimental folding temperature (Equation 4.4) and $\Delta S(n)$ is obtained from Equation 4.3. E_d sets the scale while m describes the dependence of destabilization free energy on nativeness.

$$m = 1 - \left[(1+C)\left(n^j / \left(n^j + C \right) \right) \right] \quad \text{..... (4.11)}$$

Here, C and j are adjustable parameters and m takes the values from 1 to 0 as n goes from 0 to 1. Using Equation 4.11 allows the division of the destabilization effect between the folding and unfolding side of the barrier in a ratio that is consistent with the experimental estimate of m_f/m_{eq} for two-state proteins.

4.2.4 Calculation of free energy barrier heights and folding rates

For calculating folding rates at chemical mid-point (where T_m is not known) or in absence of chemical denaturant the following equation is used (combining Equations 4.4 and 4.6).

$$\Delta G(T,n) = \Delta H^0(n) - T\Delta S(T,n) \quad \text{..... (4.12)}$$

For a two-state scenario folding (and unfolding) barrier heights can be obtained, by definition, from the difference in the free energies between the unfolded (and folded) minima and the top of the barrier. However, folding and unfolding barrier heights calculated in this manner will not be in complete conformity to each other at mid-transition due to the differences in the curvature of the unfolded and native basins. Moreover, the above method cannot be applied to free energy profiles lacking barriers. Hence, a general protocol is followed for all proteins in which a dividing line is set along the reaction coordinate (n_d) at n=0.75. By adjusting only ΔH^0_{res} equal populations on either side of the dividing line are obtained for mid-transition condition. The transition state is defined as the region having a width (w) of 0.12 around the dividing line. Barrier heights are then obtained from the ratio of the integrated probability of the unfolded (and folded) state and the transition state ensemble, i.e. the folding barrier height $= -RT \ln \left(\frac{\int_{i(n>n_d-w/2)}^{i(n=n_d+w/2)} \exp(-\Delta G_i / RT)}{\int_{i(n=0)}^{i(n=n_d-w/2)} \exp(-\Delta G_i / RT)} \right)$. The position of the dividing line and the width of the transition state region are calibrated according to the typical shapes of the free energy surface and to obtain maximum

agreement between folding and unfolding barriers heights as well as between populations on either side of the barrier (i.e. $P_U = \int_{i(n=0)}^{i(n=n_d)} \frac{\exp(-\Delta G_i / RT)}{\int_{i(n=0)}^{i(n=1)} \exp(-\Delta G_i / RT)}$;

$P_F = \int_{i(n>n_d)}^{i(n=1)} \frac{\exp(-\Delta G_i / RT)}{\int_{i(n=0)}^{i(n=1)} \exp(-\Delta G_i / RT)}$). For calculating barrier heights at conditions other than mid-point, populations are adjusted on either side of the dividing line such that $-RTln(P_U / P_F)$ matches experimental stabilities, i.e. unfolding free energies. Folding rates are then calculated using Kramer-like expression $k = D\exp\left(-\Delta G^{\#-U} / RT\right)$ where the effective diffusion coefficient D is expressed as

$$D(T) = k_0 \exp\left(\frac{-E_{a,res}N}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right) \quad \text{..... (4.13)}$$

Here $k_o \propto 1/N$ and assumed to be temperature independent while the reference temperature T_0=298 K. The effects of temperature-dependent changes in solvent viscosity and roughness of the landscape are implicitly accounted for and lodged in the activation energy per residue ($E_{a,res}$). A value of 1 *kJ/mol* is used for $E_{a,res}$ estimated from the analysis of temperature dependence of relaxation rates of microsecond-folding proteins (for more details see Chapter 6).

4.2.5 Modeling DSC profiles and Chevron Plots

From Equation 4.9 free energy profile is generated at the desired mid-point temperature by allowing ΔH^0_{res} to float such that $-RTln(P_U/P_F)$=0. Using the modified

Gibbs-Helmholtz equation free energy profiles at various temperatures are built by propagating the temperature effects from the thermal mid-point:

$$\Delta G(T,n) = \Delta H(T,n) - T\Delta S(T,n)$$
$$\Delta H(T,n) = \Delta H(T_m,n) + \Delta C_p(T - T_m)$$
$$\Delta S(T,n) = \Delta S(T_m,n) + \Delta C_p \ln(T/T_m)$$

.....(4.14)

The excess heat capacity as a function of temperature is obtained from temperature-dependent probabilities:

$$\Delta C_p^{excess}(T) = \frac{\langle \Delta H^2 \rangle - \langle \Delta H \rangle^2}{RT^2} \quad \text{..... (4.15)}$$

where $\langle \Delta H \rangle = \sum_{i(n=0)}^{i(n=1)} \Delta H_i(T,n)\, p_i(T,n)$; $p_i(T,n) = \dfrac{\exp(-\Delta G_i(T,n)/RT)}{\sum_{i(n=0)}^{i(n=1)} \exp(-\Delta G_i(T,n)/RT)}$

The DSC profile can be then generated from excess heat capacity and chemical baseline as follows:

$$\langle \Delta C_p(T) \rangle = \Delta C_p^{excess}(T) + \sum_{i(n=0)}^{i(n=1)} \Delta C_{p_i}(n)\, p_i(T,n) \quad \text{..... (4.16)}$$

To simulate the dependence of relaxation rates on the concentration of chemical denaturant (i.e. Chevron plot), the same approach as described in Chapter 3 is taken. Relaxation kinetics following a chemical-jump is modeled as diffusive kinetics on free energy surfaces resulting from Equation 4.10 using a constant effective diffusion coefficient.

4.2.6 Inclusion of size, structure and sequence effects

The exponents $k_{\Delta H}$ and $k_{\Delta C_p}$ can be directly related to protein size by the expressions:

$$k_{\Delta H} = c_{\Delta H}.B^{X/N}; \quad k_{\Delta C_p} = c_{\Delta C_p}.B^{X/N} \quad \text{..... (4.17)}$$

where X=1; and B, $c_{\Delta H}$ and $c_{\Delta C_p}$ are adjustable parameters. The effects of structure are incorporated by putting X=ΔL, the number of residues separating a pair of residues in contact. A contact is defined between two residues if any of their atoms are within a specified spatial distance or if a backbone hydrogen bond is formed between their C=O and N-H groups. Backbone hydrogen bonds are calculated from protein three-dimensional structures (protons are added using the WHAT IF program in case of crystal structures) using the same geometrical considerations and parameters as those utilized by Kortemme *et al*.[108] in deriving an orientation-dependent hydrogen bonding potential. Residue-residue contacts are considered with different atomic representations: only C_α-C_α pairs, only C_β-C_β pairs, all non-hydrogen heavy atoms, only side chain heavy atoms, center of masses of side chain heavy atoms.

Furthermore, the effects of amino acid sequence can be incorporated into the model in a very straightforward manner. The cost of fixing the backbone and side chain of a residue in a particular native dihedral angle space varies for different amino acids. The estimation of sequence-dependent conformational entropies from protein structure statistics is discussed in the next chapter. In the bare bones version of the model a mean value for $\Delta S_{res}^{n=0}$ is used. This value can be replaced by individual sequence-dependent conformational entropies resulting in a range of conformational entropy functionals for each residue. To obtain the total conformational entropy

functional for each protein, the individual decays for each residue are averaged and then scaled by protein size. As a preliminary approach empirically derived residue-residue contact potentials from protein structures are used to include sequence-dependent energetics in the model[109,110]. This is easily achieved by knowing the specific amino acids involved in a contact and scaling the decay of individual contacts (with respect to *n*) with respective interaction energies obtained from the matrix of pair-wise contact energies.

4.3 Protein Database Used in the Analysis

4.3.1 Selection criteria

Folding data on chemical as well as thermal denaturation of proteins from stopped-flow, ultra-fast mixing and T-jump relaxation studies is considered here. Proteins clearly confirmed as three-state proteins from both equilibrium and kinetic experiments under a range of conditions (pH, ionic strength, buffer) such as lysozyme, myoglobin, barnase, barstar, ribonucleases, etc. are excluded. Additionally, proteins containing heme groups (cytochromes), tandem repeats (ankyrin repeat) or disulfide linkages are not included. All single domain two-state proteins having lengths less than 130 residues for which folding/unfolding data is available with the exception of VlsE (Variable major protein-Like Sequence, Expressed, *B. burgdorferi*) having 341 residues and those proteins for which experimental data is not good enough; and naturally fast folding proteins as well as those designed to fold fast are selected. The protein dataset selected for the present

analysis (Table 4.1) is the most extensive one used so far even after the aforementioned exclusions.

The best way to compare folding rates of proteins is by using those values measured at their respective folding temperatures, around neutral pH and in the absence of any chemical denaturant or salt. For proteins studied under a range of temperatures in pure water, such rates are directly available. But a majority of proteins reported in the literature are investigated by chemical jump experiments for which rate constants (k_f and k_u) in water are estimated by linear extrapolation from conditions of higher denaturant concentrations. k_f and k_u are generally fitting parameters and thus their values are highly dependable on the fitting procedures used. The reported uncertainties for these rate constants are usually underestimated and related to only fitting errors. Comparison of k_f's predicted from free energy profiles require one more empirical parameter – folding stability for each protein, which is very sensitive to experimental conditions (temperature, pH, ionic strength, buffer) making their estimates quite error-prone. In order to avoid these uncertainties, here, folding rates are compared at isostability conditions (i.e. zero stability at mid-point transition produced either by chemical or thermal denaturation). Using mid-point rates (k_m) for comparison has dual advantages: first experimental estimates of k_m have relatively less errors than k_f and k_u and secondly the precision of k_m's predicted from the model is not affected by the *ad hoc* model procedures for calculating populations on either side of the barrier. Additionally this allows the inclusion of fast folding proteins that have been studied only at their mid-point temperatures and proteins such as Colicin-binding bacterial immunity protein 7 that show deviations from two-state behavior

only under highly native conditions but not at mid-point transition. In case of availability of folding rates at both chemical and thermal mid-point points for a single protein, two entries per protein are added.

For proteins that exhibit an additional slow phase due to cis/trans proline isomerization, only the fast phase is considered here.

Table 4.1 Proteins/Protein domains used in the analysis

	Class	Protein	Species	PDB code	Experimental Method	L_{PDB}	L_{EXP}
1. BBL	α	E3-binding domain of dihydro-lipoamide succinyl transferase	Escherichia coli	2CYU	NMR	39	40
2. BBL (H166W)	α	BBL pseudo wild type	Escherichia coli	2BTH	NMR	45	45
3. E3BD	α	E3-binding domain of dihydro-lipoamide acetyl transferase	Bacillus stearothermophilus	1EBD	X-ray	41	41
4. E3BD (F166W)	α	E3BD pseudo wild type	Bacillus stearothermophilus	1W4E	NMR	45	45
5. POB	α	E3-binding domain of dihydro-lipoamide succinyl transferase	Pyrobaculum aerophilum	1W4J	NMR	51	51
6. EngHD	α	Engrailed HomeoDomain	Drosophila melanogaster	1ENH	X-ray	54	54
7. hTRF1	α	DNA-binding domain of human telomeric protein	Homo sapiens	1ITY	NMR	67	67
8. hRAP1	α	Human RAP1 Myb domain	Homo sapiens	1FEX	NMR	59	59
9. c-Myb	α	c-Myb transforming protein	Mus musculus	1GUU	X-ray	50	50
10. FSD	α	Full Sequence Design-1	-	1FME	NMR	28	28
11. Trp Cage	α	Tryptophan cage	-	1L2Y	NMR	20	20
12. α-3D	α	Designed protein α-3D	-	2A3D	NMR	73	73
13. BdpA	α	B-domain of protein A (Y15W)	Staphylococcus aureus	1SS1	NMR	60	60
14.Villin-HP35 (N27H)	α	Headpiece subdomain of F-actin binding protein villin	Gallus gallus	1VII	NMR-average	35	35
15. $\lambda_{6\text{-}85}$	α	Monomeric N-terminal domain of Lambda Repressor	Bacteriophage lambda	1LMB	X-ray	80	80
16. ACBP	α	Acyl CoA binding protein	Bos taurus	2ABD	NMR	86	86
17. Im9	α	E Colicin binding Immunity Protein 9	Escherichia coli	1IMQ	NMR	86	86
18. Im7	α	E Colicin binding Immunity Protein 7	Escherichia coli	1AYI	X-ray	86	86 (94)

Table 4.1 Proteins/Protein domains used in the analysis (continued)

	Class	Protein	Species	PDB code	Experimental Method	L_{PDB}	L_{EXP}
19. Pin WW	β	Mitotic Rotamase Pin1	Homo sapiens	1PIN	X-ray	34	34
20. YAP65	β	Yes Kinase Associated protein 65	Homo sapiens	1K9Q	NMR	40	40
21. WW Prototype	β	Designed WW prototype	-	1E0M	NMR	37	37
22. FBP28 (W30A)	β	Formin Binding Protein	Mus musculus	1E0L	NMR	37	37
23. α-Spectrin SH3	β	α-Spectrin SH3 domain	Gallus gallus	1SHG	X-ray	57	62
24. Fyn SH3	β	Fyn proto-oncogene tyrosine kinase SH3 domain	Homo sapiens	1SHF	X-ray	59	67
25. Src SH3	β	c-Src tyrosine kinase SH3 domain	Homo sapiens	1FMK	X-ray	56	57
26. PI3K SH3	β	Phosphatidyl inositol-3-Kinase SH3 domain	Bos taurus	1PNJ	NMR-average	86	90
27. ABP1 SH3	β	Actin Binding Protein1 SH3 domain	Saccharomyces cerevisiae	1JO8	X-ray	58	68
28. Sso7d (Y34W)	αβ	DNA binding protein Sso7d	Sulfolobus solfataricus	1BF4	X-ray	63	63
29. CspB-Bs	β	Cold shock protein	Bacillus subtilis	1CSP	X-ray	67	67
30. CspB-Bc	β	Cold shock protein	Bacillus caldolyticus	1C9O	X-ray	66	66
31. CspB-Tm	β	Cold shock protein	Thermotoga maritima	1G6P	NMR	66	66
32. CspA	β	Cold shock protein	Escherichia coli	1MJC	X-ray	69	69
33. Fibronectin	β	9^{th} Fibronectin type III Domain of Fibronectin	Homo sapiens	1FNF	X-ray	90	90
34. Tenascin	β	3^{rd} Fibronectin type III Domain of Tenascin	Homo sapiens	1TEN	X-ray	90	90
35. TI27	β	Repeat 27 of Titin	Homo sapiens	1TIT	NMR-average	89	89

Table 4.1 Proteins/Protein domains used in the analysis (continued)

	Class	Protein	Species	PDB code	Experimental Method	L_{PDB}	L_{EXP}
36. Twitchin	β	Twitchin	Caenorhabditis elegans	1WIT	NMR-average	93	93
37. Tendamistat	β	Tendamistat	Streptomyces tendae	2AIT	NMR	74	74
38. GPW	αβ	Viral Protein	-	1HYW	NMR	58	61
39. mAcP	αβ	Muscle Acyl Phosphatase	Equus caballus	1APS	NMR	98	98
40. ctAcP	αβ	Common type Acyl Phosphatase	Bos taurus	2ACY	X-ray	98	98
41. CI2	αβ	Chymotrypsin Inhibitor 2	Hordeum vulgare	1COA	X-ray	64	64
42. C-PTL9	αβ	C-terminal domain of Ribosomal Protein L9	Bacillus stearothermophilus	1DIV	X-ray	92	92
43. N-PTL9	αβ	N-terminal domain of Ribosomal Protein L9	Bacillus stearothermophilus	1DIV	X-ray	56	56
44. Protein G	αβ	Immunoglobulin binding domain B1 of Protein G	Streptococcus Lancefield Group G	1PGB	X-ray	56	56
45. Protein L	αβ	Immunoglobulin binding domain B1 of Protein L	Peptostreptococcus magnus	1HZ6	X-ray	62	62
46. Ubiquitin	αβ	Ubiquitin	Homo sapiens	1UBQ	X-ray	76	76
47. ADAh2	αβ	Activation domain of Procarboxypeptidase A2	Homo sapiens	1AYE	X-ray	80	80
48. U1A	αβ	Spliceosomal protein U1A	Homo sapiens	1URN	X-ray	96	102
49. S6	αβ	Ribosomal Protein S6	Thermus thermophilus	1RIS	X-ray	97	101
50. FKBP12	αβ	FK506 Binding Protein	Homo sapiens	1FKB	X-ray	107	107
51. Hpr	αβ	Histidine containing phosphocarrier protein	Escherichia coli	1POH	X-ray	85	85
52. Villin14T	αβ	Actin severing domain Villin 14T	Gallus gallus	2VIK	NMR-average	126	126
53. RafRBD	αβ	Ras-Binding Domain of c-Raf1	Homo sapiens	1RFA	NMR	78	80

Table 4.1 Proteins/Protein domains used in the analysis (continued)

	Class	Protein	Species	PDB code	Experimental Method	L_{PDB}	L_{EXP}
54. Prb (K5I/K39V)	α	GA module of Albumin-binding domain	Peptostreptococcus magnus	1PRB	NMR-average	47	47
55. BBA5	α	Designed protein BBA5	-	1T8J	NMR-average	23	23

L_{PDB}: Number of residues reported in the PDB files and used for contact order calculations

L_{EXP}: Length of experimental construct

PDB structures as suggested by respective investigators and/or that closely match the experimental construct are chosen. Proteins for which both crystallographic and solution structures are available PDB files are chosen with the preferential order: X-ray > NMR-energy minimized average > NMR-multimodels (the chosen model number is specified in the following remarks)

Remarks are numbered according to the serial number of the protein. References to kinetic data are indicated .

1. Naphthyl-Alanine at the N-terminus is missing in the NMR structure. Atomic coordinates for residues 2-40 are reported in the pdb file. 1st model taken from 20 structures reported. [Personal Communication V. Muñoz]
2. Model 1 chosen from 20 NMR structures.[111]
3. Residues 130-170 of Chain C taken.[112]
4. Model 1 chosen from 20 NMR structures.[111]
5. Model 1 chosen from 20 NMR structures.[111]
6. No remarks.[113,114]
7. 2nd Model chosen from 25 NMR structures.[114]
8. 1st Model chosen from 25 NMR structures.[114]
9. No remarks.[114]
10. 34 structures reported, 1st model taken. [Personal Communication V. Muñoz]
11. 38 structures reported, 1st model taken.[115]
12. 1 solution structure reported.[116]
13. NMR structure corresponds to the characterized Y15W mutant.[117]
14. PDB file is for wild type protein with residues 41-76 whereas folding of N27H mutant studied.[118]
15. Segment 3, residues 6-85 taken.[119]
16. 29 structures reported, 1st taken.[120]
17. No remarks.[120,121]
18. Number in the parentheses corresponds to the length of the experimental construct that contains N-terminal His6 tag and a short unstructured C-terminal tail. However in this analysis the number of residues reported in the crystal structure is used. [120,121]
19. Residues 6-39 of Chain A.[122]
20. 20 structures reported, 1st model taken.[123]

21. 20 structures reported, 1st model taken.[123]
22. 10 structures reported, 1st model taken. The structure is for wild type whereas W30A mutant studied experimentally.[123]
23. In the X-ray structure residues 1-5 are unstructured. Atomic coordinates are reported for residues 6-62.[124]
24. The characterized construct had an N-terminal tail of residues 'GS' and a C-terminal tail of 'EFIVTD' residues that are not present in the pdb file.[125]
25. Residues 85-140.[126]
26. The extra N-terminal 'GS' and C-terminal 'WNSS' residues are present in the construct. In the NMR structure the C-terminal tail is not reported.[127]
27. The construct has unstructured N- and C-terminal tails.[120]
28. Residues 2-64.[128]
29. No remarks.[129]
30. No remarks.[130]
31. Out of 7 conformers submitted, 1st model taken.[130]
32. No remarks.[131]
33. Residues 1327-1416.[132]
34. Residues 802-891.[133]
35. No remarks.[134]
36. No remarks.[134]
37. 9 NMR structures reported, 1st model taken.[135]
38. 15 structures reported, 1st model taken. The construct has 'RRRG' C-terminal tail that is missing in the structure file. The PDB file has an extra Met at the N-terminal with Val at the 2nd position in the construct is replaced by Thr. [Personal Communication V.Muñoz]
39. 5 structures reported, 1st model taken.[136]
40. No remarks.[137]
41. PDB code 1coa is for mutant I76V while wild-type is characterized experimentally.[138,139]
42. Residues 58-149 taken.[140]
43. Residues 1-56 taken.[141]
44. No remarks.[142]
45. Residues 3-64 of PDB file, (residues 11-72 of the protein sequence). The PDB code 1hz6 is for the mutant Y47W.[143]
46. No remarks.[120]
47. Chain A, Residues 4A-99A is the pro-segment. Since 34B and 34C belong to the pro-segment, they are renumbered as 35,36,following which residues 35A-42A are renumbered as 37-44. Due to discontinuity 42A is followed by 47A. Residues 47A-82A are renumbered as 45-80.[144]
48. PDB code 1urn has mutationsY31H and Q36R. Residues 2-97 of chain A taken (1st N-terminal Met and last 5 C-terminal residues Lys-Gly-Thr-Phe-Val are missing in X-ray structure). Experimental construct is a F56W mutant with 102 residues.[145]
49. The last four residues Leu-Ala-Asn-Ala are missing from the C-terminus in the PDB file.[25]
50. No remarks.[146]; 51. No remarks.[147]; 52. No remarks.[148];
53. Residues 55-132.[120]; 54. Residues 7-53 reported for wild-type protein.[149]; 55. Contains a D-Pro at position 4.[150]

4.4 Results and Discussion

4.4.1 Particularities of the 1-D free energy surface model

As discussed in Section 4.2 the model incorporates the empirical size scaling behavior of thermodynamic properties of proteins. Hence, the total magnitude of the various contributions to the free energy surface can be determined from only knowing the length of the protein and the mean per-residue values of the thermodynamic parameters (i.e. ΔH^{0}_{res}, $\Delta S^{n=0}_{res}$ and $\Delta C_{p,res}$). However, the extent to which stabilization energy is gained and the total entropy is lost at any intermediate stage in folding is controlled by the values of the exponents of the enthalpy and heat capacity functionals ($k_{\Delta H}$ and $k_{\Delta C_p}$).

For most model calculations $\Delta S^{n=0}_{res}$ is fixed to 16.5 $J.mol^{-1}.K^{-1}.res^{-1}$. This value of $\Delta S^{n=0}_{res}$ translates into 17.6 $J.mol^{-1}.K^{-1}.res^{-1}$ for the entropic cost of folding, which is obtained from the maximum difference in the conformational entropy functional or calculated from Equation 5.9. This value for entropic cost per residue upon folding is similar to the estimate of 17.4 $J.mol^{-1}.K^{-1}.res^{-1}$ for the same at 385 K obtained from thermodynamic data of 53 proteins by Robertson and Murphy[41]. The protein database employed in the present analysis includes several fast folding proteins that have much shorter lengths than the smallest protein (with 56 residues) in Robertson and Murphy's dataset. Linear scaling with size using $\Delta C_{p,res}$=0.058 $kJ.mol^{-1}.K^{-1}.res^{-1}$ (as suggested by Robertson and Murphy) significantly overestimates ΔC_p values for these proteins. When DSC data for a set of proteins including two-

state as well as fast folding and downhill proteins is subjected to a global fitting using the simple model described here a value of 0.05 $kJ.mol^{-1}.K^{-1}.res^{-1}$ is obtained for $\Delta C_{p,res}$. Hence, in all the calculations mentioned in this chapter this value of $\Delta C_{p,res}$ is used.

At a particular temperature both the shape and magnitude of the conformational entropy (green curve in Figure 4.2A) function is determined by the value of $\Delta S_{res}^{n=0}$. A larger value of $\Delta S_{res}^{n=0}$ will render a curve with narrower width and its maximum shifted towards n=0. In Figure 4.2A the total change in enthalpy and heat capacity calculated with $k_{\Delta H}$=1.5 and $k_{\Delta C_p}$=3 are shown for a 65-residue protein. In case of enthalpy functional a lower value of $k_{\Delta H}$ gives rise to a shallower decay of stabilization energy. When total entropic contributions are constant increasing $k_{\Delta H}$ will result in steeper enthalpy functionals such that at any intermediate value of nativeness, for e.g. n=0.7, the % gain in stabilization energy becomes increasingly smaller ultimately manifesting into increasingly larger barriers. The effect of the change in $k_{\Delta H}$ on free energy barriers can be seen in Figure 4.2C where free energy surfaces are compared at the same temperature and isostability conditions. This clearly shows that modification of a single parameter $k_{\Delta H}$ is sufficient to alter barrier heights. The total entropy has two opposing components: the temperature-independent conformational entropy and temperature-dependent solvation entropy (Equation 4.7). It can be seen from Figure 4.2B that the major effect of temperature in the total entropy arises from the changes in solvation entropy at different temperatures.

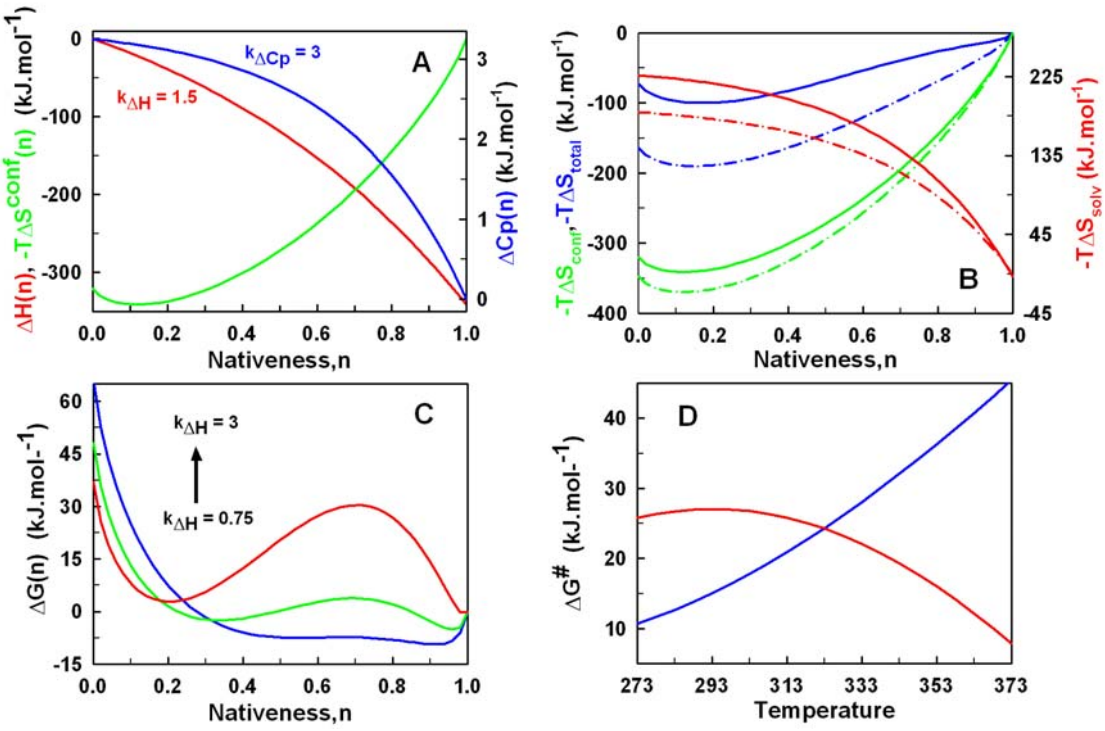


Figure 4.2 Functionals used in generating 1-D free energy surface and temperature dependence of free energy barrier heights

(A) (Left axis) Enthalpic (red, blue) and entropic (green) contributions to free energy; (Right axis) Total change in heat capacity. (B) Dissection of (left axis) total entropy (blue) at 298 K (solid lines) and 323 K (dashed dotted lines): conformational entropy (green) and (right axis) solvation entropy (red). (C) Free energy profiles showing negligible barrier (blue), marginal barrier (~3 *RT*) (green) and large barrier (~12 *RT*). (D) Dependence of folding (blue) and unfolding (red) activation free energy on temperature.

There are very little changes in the total conformational entropy between 298 K and 323 K. At a given temperature the curvature and magnitude of the solvation term depends on $k_{\Delta C_p}$ and the value of $\Delta C_{p,res}$. Figure 4.2D shows the temperature dependence of folding and unfolding barrier heights for a protein having a T_m of 323 K. The downward curvature for folding barrier heights and upward curvature for unfolding barrier heights are consequences of heat capacity of the transition state being intermediate between unfolded and native state (red curve in Figure 4.2A, where the value at n=0.7 corresponds to top of the barrier). This effect has also been seen in refolding experiments where the folding rate constants increases with temperature, passes through a maximum and then decreases.

4.4.2 Simulation of DSC and chemical denaturation experiments

As seen from Figure 4.3C the 1-D free energy surface model is able to reproduce the entire range of folding regimes from two-state with barriers of ~12 RT to marginal barriers of ~3RT to completely barrier-less. Since the model incorporates the effects of thermal and chemical denaturation, it can be directly applied for the analysis thermodynamic and kinetic experiments. Figure 4.3A shows a simulated DSC experiment for a protein having 65 residues, a T_m =323 K and folding free energy barrier of ~9 RT. The sharpness of the transition expected for a two-state system and the characteristic sigmoidal baseline reflecting heat capacity changes associated unfolding transition is reproduced in this DSC profile. Equilibrium denaturation profile calculated for a protein with 65 residues and having a chemical midpoint (d_m) at ~4.5 M is shown in Figure 4.3B. This profile exhibits sigmoidal decay typically

observed in experiments. Relaxation kinetics following perturbation of free energy surfaces generated from Equation 4.10 are shown in Figure 4.3C. The hypothetical protein considered here has a free energy barrier of ~16 RT at chemical midpoint suggesting a two-state system. And consistent with the two-state criterion all the

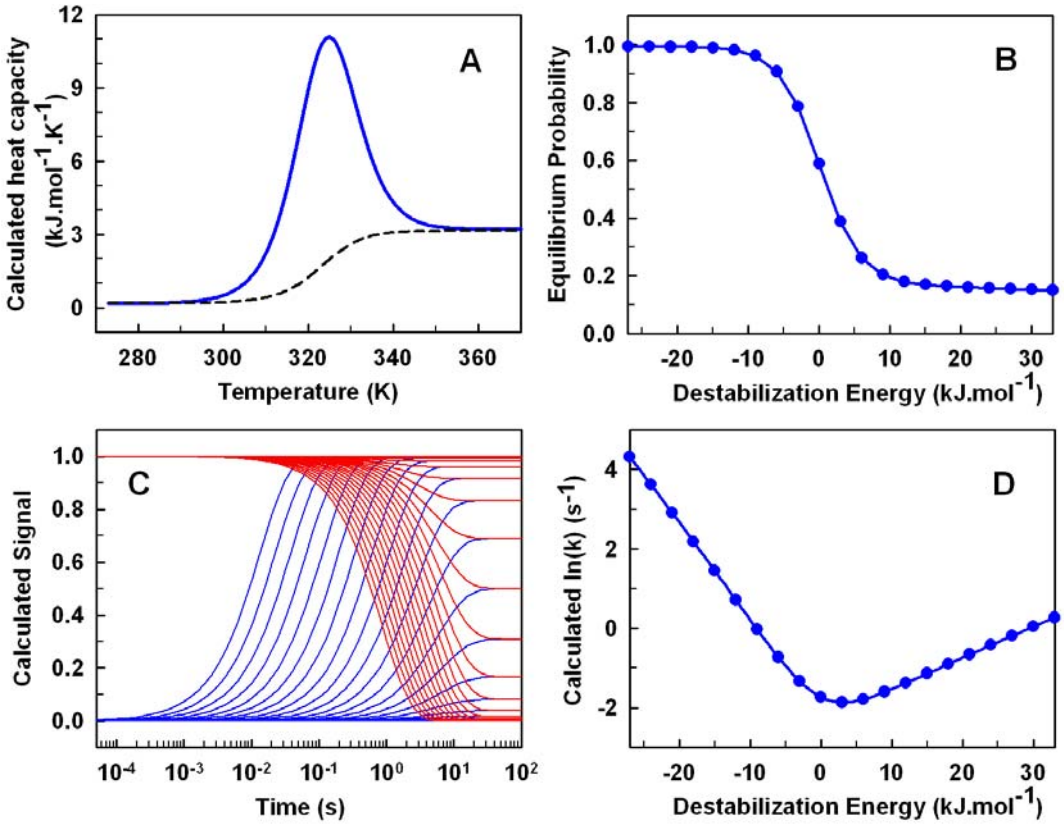


Figure 4.3 Simulations of thermal and chemical denaturation experiments

(A) Differential Scanning Calorimetry profile. Chemical base line is shown as dashed line; (B) Equilibrium probability at 298 K as a function of destabilization energy ($E_d(d\text{-}d_m)$ from Equation 4.10); (C) Relaxation traces after various chemical-jumps starting from highly destabilizing (blue) and stabilizing (red) conditions; (D) Relaxation rate constants obtained from (C) as a function of destabilization energy, i.e. Chevron Plot. Solid lines in (B) and (D) are guide to the eye.

relaxation traces in Figure 4.3C can be fitted to a single exponential function. In order to simulate chemical denaturation experiments, a arbitrary concentration range of 0-10 M and a scaling factor for destabilization energy E_d=6 $kJ.mol^{-1}$ are chosen. Rather than assigning an *ad hoc* stability (i.e. ΔG_{eq}) to the protein, destabilization effects are propagated from the mid-point $d=d_m$ (ΔG_{eq}=0). Although, the value of destabilization energy (i.e. $E_d\,(d\text{-}d_m)$) equals zero at $d=d_m$, it does not implicate high native bias, i.e. 0M (as d_m=4.5M). Instead of arbitrarily fixing destabilization energy, expressing it as $E_d\,(d\text{-}d_m)$ provides convenience in fitting experimental Chevron plots knowing the range of d_m and m_{eq} (which directly relates to E_d) from equilibrium chemical denaturation experiments. Relaxation rates obtained from the kinetic traces are plotted with respect to d_m resulting in the characteristic V-shaped Chevron plot (Figure 4.3D). Using j=2 and C=0.4 in Equation 4.11 generates the function shown in Figure 4.4 that partitions the destabilization energy between the folding and unfolding side of the barrier in a manner such that the ratio of the slopes of the folding and unfolding arm of the chevron is ¾:¼. This ratio is consistent with that found from kinetic denaturation experiments of two-state proteins. The destabilization energy (i.e. d_m) corresponding to the minimum in the Chevron plot closely agrees with that from the equilibrium denaturation profile at which the signal decays by 50%. The conformity in the value of d_m obtained from equilibrium and kinetic chemical denaturation experiments is also one of the signatures of two-state proteins. Moreover, the dependence of destabilization energy on nativeness described by Equation 4.11 translates into a linear change in the macroscopic unfolding free energy (i.e. $-RTln(P_U/P_F)$) similar to that observed in experiments (inset, Figure 4.4).

These results show that the model is able to simulate and reproduce experimentally observed features in DSC profiles and Chevron plots and thus can be used for direct fitting and analysis of empirical data. Unlike two-state analysis, the advantage of using this model is that no *a priori* assumptions are made for the number of macro-states.

Moreover, the parameters of the model are consistent with empirical estimates of thermodynamic quantities. At any given temperature the equilibrium population ratio can be modulated by adjusting just ΔH_{res}^{0} while barrier heights can be modified by the exponent of the enthalpy functional. Furthermore, the simplicity of the model facilitates the incorporation of the effects of protein size, 3-D structures and energetics by detailed parameterization of $k_{\Delta H}$ and $k_{\Delta C_p}$. In Equation 4.17 when X is 1, $k \propto 1/N$, suggesting that the curvatures of the exponential functionals decrease with protein size.

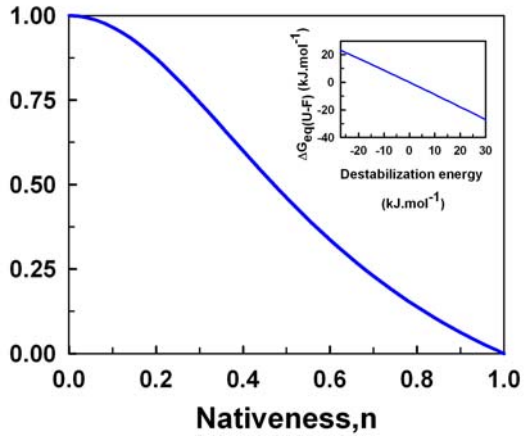


Figure 4.4 Dependence of destabilization energy on nativeness

The function generated from Equation 4.11. Inset shows the change in protein stability, i.e. $\Delta G_{eq(U\text{-}F)}$ from from 23 to –30 $kJmol^{-1}$as a result of destabilization caused by chemical denaturation.

However, this does not necessarily mean that larger barriers result for smaller proteins because the barrier height depends upon how the decrease in total entropy compensates the decay in total stabilization energy, both of which scale linearly with size.

4.4.3 Prediction of folding rates at mid-transition

Folding rates at chemical and/or thermal denaturation midpoints of 53 proteins listed in Table A1 are considered for prediction (with the exclusion of Prb and Bba5). Comparison between experimental midpoint rates ((i.e. $k_m/2$) and those predicted by the model using size-scaling of $k_{\Delta H}$ and $k_{\Delta C_p}$ are shown in Figure 4.5A. The predicted rates span the same range as the experimental rates with which they show a correlation coefficient (R) of ~0.9 (*p-value* <2.4e-20). The prediction resulting from using just a single input i.e. protein size is quite remarkable with a mean discrepancy between the calculated and experimental rates of less than one order of magnitude (i.e. a factor of 8 or ~1/9 of the dynamic range). However, this version of the model predicts almost identical rates for proteins having very similar sizes and folding temperatures (note the horizontal pattern of data points of some proteins in Figure 4.5A) suggesting that at the level of individual proteins it is important to add effects arising from other factors, i.e. differences in protein topologies and energetic effects arising from protein sequences.

Contact maps (Figure A1 in appendix) generated from 3-D structures (PDB filenames are listed in Table 4.1) reflect the sequence separation between contacting residues. Replacing X in Equation 4.17 with the sequence separation for all atomic contacts between residues allows understanding the role of structures in determining

barrier heights and hence folding rates. The mean evolution of stabilization energy (and heat capacity changes) with respect to *n* is obtained by averaging the decays of individual contacts. For two proteins of same size and compared at the same folding temperature, the one with a larger average sequence separation will have higher mean values of $k_{\Delta H}$ and $k_{\Delta C_p}$, and therefore will have a larger barrier. (Figure 4.2A). This is consistent with the idea that more complex topologies fold more slowly. Figure 4.5B shows the prediction of rates when contacts between Cα-Cα atoms at distances less than 0.6nm are included in $k_{\Delta H}$ and $k_{\Delta C_p}$ in combination with the dependence on size. The correlation between experimental and predicted rates shows improvement (R =0.93, *p-value* <1.2e-24) with the mean discrepancy decreasing by ~13% (factor of 6). Interestingly, maximum improvement is seen in α-proteins for which the mean discrepancy between calculated and experimental rates decreases by ~24% of that obtained with size-scaling followed by β-proteins with a decrease of ~20% in mean discrepancy. The prediction of rates for α-β proteins, on the other hand, becomes worse with an increase in the mean discrepancy of ~8% (see table 4.2). This suggests that the mutual effects of protein size and structure appear to work in opposite direction for some proteins, which is not surprising when several other factors are not accounted for.

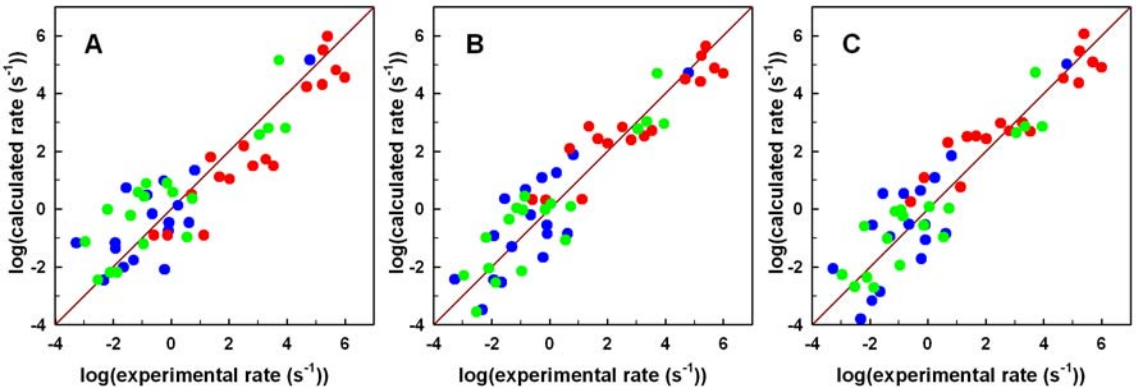


Figure 4.5 Prediction of mid-point folding rates with 1-D free energy surface model

Calculations performed with the model including (A) Size-scaling; (B) Sequence separation between Cα-Cα contacts of ≤0.6 nm; (C) Cα-Cα contacts of ≤0.6 nm energy weighted according to Miyazawa-Jernigan potential and sequence-dependent conformational entropies. The data points of the three main structural classes are shown in different colors: all-α (red), all-β (green) and α–β (blue).

Next sequence-specific details are added by replacing the mean value of $\Delta S_{res}^{n=0}$ with conformational entropies for each residue (listed in Table 5.5) and weighing residue-residue contacts according to the empirical force-field developed by Miyazawa and Jernigan[110]. The comparison of calculated and experimental rates in Figure 4.5B show noticeable changes but the correlation (R = 0.93, *p-value*<1.3e-24) is essentially the same as in Figure 4.5B. But the mean discrepancy is seen to increase slightly by ~2% of that when non-energy weighted contacts and mean value of $\Delta S_{res}^{n=0}$ are used (Figure 4.5B). In this case, β-proteins lead the race by showing a decrease of ~8% in mean discrepancy between their theoretical and empirical rates. The rate

prediction for α-proteins deteriorates by ~0.7% even though the mean discrepancy within the α class is always lower than β-proteins. For α-β proteins the mean discrepancy further decreases by ~12%. A likely explanation for the weak performance of the Miyazawa-Jernigan potential in predicting rates is that it is derived from the statistical survey of large number of 3-D structures. The interaction energy between two contacting residues is assumed to be proportional to the frequency of their occurrence in a structural database. By this method the sequence-specific energetic details at the level of individual atomic contacts are lost due to averaging. Empirical potentials like the Miyazawa-Jernigan one may not provide sufficiently detailed description of protein energetics and hence may not be adequate in reproducing the folding rates of individual proteins.

It is hard to conclude with certainty about the effects of including structure and sequence details on the predictive power of the model given the limited dataset and the statistically insignificant differences in rate prediction. As it can be seen from Table 4.2 the different versions of the model predict rates, on an average, within a factor of 10. Improving from a factor of 8 to a factor of 6 should be inconsiderable. However, it is the trend that is important to notice here. From these results it is clear that protein length is the primary determinant of folding rates with protein structure and sequence playing the secondary role. When the details of structure or sequence-dependent energetics are not properly modeled, their inter-related effects on the folding kinetics are hard to reconcile.

The dataset used in the present analysis involves a large majority of proteins whose folding properties are investigated experimentally. Examination of the dataset

shows that there are almost an equal number of representatives from each structural class (19 all α, 18 all β and 17 $\alpha\beta$ proteins). The problem, however, is that 65% of the proteins in the dataset belong to just a handful number of sets of homologous proteins. These proteins are grouped into different scaffolds based on the classification of both CATH[151] and SCOP [152]databases (shown in Table A2 in Appendix). The grouping in each scaffold is performed with the criteria that homologous proteins should also belong to the same fold, super-family, family, architecture and topology. By doing this the size of the dataset effectively reduces from 54 to 11 (19 singleton entries are excluded). Analysis of these 11 structural scaffolds helps to better discern the structure and sequence effects from the consequences of size-scaling. After performing global fitting of mid-point rates for all proteins as discussed above, the average folding rates for each scaffold are compared as shown in Figure 4.6. The improvement in rate prediction from using length-dependent (Figure 4.6A) to structure- and sequence-dependent (Figure 4.6B) exponents in the model now becomes more apparent. The mean discrepancy between experimental and theoretical average rates decreases by as large as ~43% (i.e. from Panel A to Panel B) for 11 scaffolds comprising of 35 proteins. While the mean discrepancy for all 54 proteins shows a reduction of only ~11% when structural and energetic details are added to the model. The reason for this drastic decrease is partly due to sheer averaging effects. The mean discrepancy between the average experimental and predicted rates is as low as 0.442 even if any 4 proteins are randomly chosen for each fold in the length calculation and 0.371 when sequence and structure are involved.

Table 4.2 Summary of results: Prediction of mid-point folding rates with 1-D free energy surface model

Model versions	$c_{\Delta H}$	$c_{\Delta C_p}$	**R**	$\overline{\left\|\log(k_m^{\exp}) - \log(k_m^{pred})\right\|}$ **(54 proteins)**	$\overline{\left\|\log(\overline{k_m^{\exp}}) - \log(\overline{k_m^{pred}})\right\|}$ **(11 scaffolds, 35 proteins)**
A. Size	1.441	3.794	0.899	**0.922**	**0.729**
				α 0.906 β 0.991 αβ 0.866	
B. Cα-Cα contacts	1.784	3.043	0.932	**0.805**	**0.461**
				α 0.693 β 0.795 αβ 0.939	
C. Energy weighted contacts + Sequence dependent conformational entropies	1.876	3.234	0.932	**0.820**	**0.412**
				α 0.698 β 0.734 αβ 1.05	

Fitting the chemical and thermal mid-point rates of 54 proteins is performed by fixing mean values of 16.5 $J.mol^{-1}.K^{-1}.res^{-1}$ and 0.05 $kJ.mol^{-1}.K^{-1}.res^{-1}$ for $\Delta S_{res}^{n=0}$ and $\Delta C_{p,res}$ respectively in versions **A** and **B** of the model. The only adjustable parameters are the coefficients in the expression $k_{\Delta H} = c_{\Delta H}.B^{X/N}$; $k_{\Delta C_p} = c_{\Delta C_p}.B^{X/N}$. The fitted values of these parameters are shown in Table 4.2. For calculations performed with version **C** of the model the mean value of $\Delta S_{res}^{n=0}$ is replaced with sequence-dependent conformational entropy cost per residue estimated from statistical analysis of protein structure database (Table 5.5). For all the above calculations B=2 and the pre-exponential $k_o=10^7/N$. The mean discrepancy is given by the average of the absolute difference between predicted (pred) and experimental (exp) mid-point folding rates and expressed in *log* units.

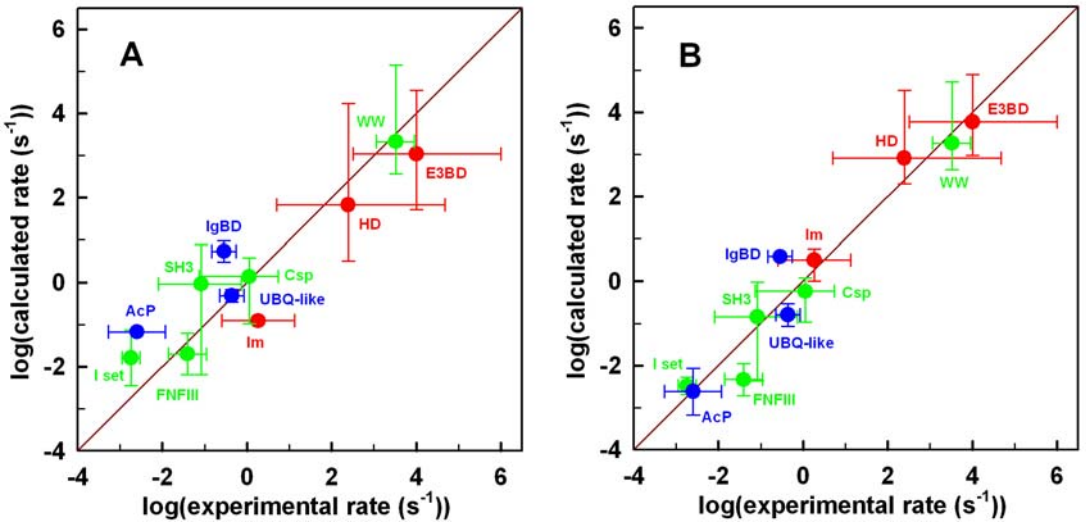


Figure 4.6 Comparison of average mid-point folding rates of structural scaffolds from prediction and experiments

Calculations performed with the model including (A) Size-scaling; (B) Cα-Cα contacts of ≤0.6 nm energy weighted according to Miyazawa-Jernigan potential and sequence-dependent conformational entropies. The data points represent the average mid-point folding rates for 11 different scaffolds: E3BD: E3 bonding domain (Peripheral Subunit Binding Domain of 2-oxo acid dehydrogenase complex); HD: Homeodomain; Im: Immunity proteins; WW: WW domains; Csp: Cold shock proteins; SH3: src-homology 3; FNFIII: Fibronectin type III; I set: Immunoglobin-like; IgBD: Immunoglobulin binding domain; UBQ-like: Ubiquitin-like; AcP: Acyl phosphatase. The color-coding for α, β, and α-β classes is the same as in Figure 4.5. The horizontal and vertical error bars correspond respectively to the range in experimental and theoretical folding rates spanned by members of individual scaffold.

As mentioned earlier the present model accounts for the differences in folding rates of proteins arising from the variation in the folding temperatures (298 K for chemical to ~350 K for thermal denaturation experiments). However, the model cannot reproduce the differences in folding rates measured in a range of solvent conditions (pH, ionic strength or buffer composition); or resulting from the use of urea or guanidinum salts as denaturing agents; or from temperature or denaturant dependence of viscosity that has a significant effect especially at mid-point conditions.

From the standpoint of model calculations the errors in predicted rates can arise due to the use of X-ray structures determined at different crystallization conditions or the choice of any one model from multiple NMR structures over an average structure. In addition, the use of wild-type protein 3D-structures for theoretical predictions when structures of protein constructs characterized experimentally are not available can also lead to discrepancies between calculated and experimental rates. The effect of grouping the proteins into scaffolds is that the errors in the measurement of folding rates of the member proteins are averaged out. And hence the performance of the models incorporating different degrees of details can be judged more clearly.

In Figure 4.6 vertical bars show the range of rates predicted by the model for each scaffold whereas horizontal bars correspond to the deviations in experimental rates of proteins within each scaffold from the mean. A perfect prediction can be said to have resulted for proteins belonging to a scaffold when the vertical bar shows the same proportion of deviation around the mean as the horizontal bar and the average predicted rate falls on the diagonal (one-to-one correspondence line). Since scaffolds

HD and E3BD include proteins with folding rates measured at chemical mid-point (~298 K) as well as temperature mid-point (~340 K) they show a larger spread in experimental rates. A similar and proportional spread in predicted rates of these scaffolds in Figure 4.6A points out to the ability of the model in accounting for temperature effects. It is interesting to note that for HD and E3BD proteins along with improvements in the prediction of average rate the spread around it also reduces when structure and sequence-specific details are added to the model. The effects of protein length and temperature are overridden by those of structure and sequence. This result is not surprising given the very similar topology (see the similar contact maps in Figure A1 in appendix) and high sequence similarities of proteins in these scaffolds. On the contrary the effects of structure and sequence are beneficial for the prediction of proteins belonging to WW, I set and AcP scaffolds (see the reduction in the length of the vertical error bar). Noticeable improvements in the prediction of average rates upon using structure and sequence can also be seen in case of SH3 and Im. Interestingly, for Csp, FNFIII and UBQ-like protein length seems to be sufficient information for reproduction of their rates. Adding structure and sequence details only worsens the prediction of their rates. For IgBD, on the other hand, neither length nor structure and sequence can successfully reproduce the folding rates. In order to quantitatively analyze the contributions coming from length, structure and sequence, a large body of experimental data on folding kinetics is required for proteins including representatives from a large number of scaffolds each having several members (here, the average number of proteins per scaffold is only 4).

In addition, for comparative analyses between theoretical predictions and experimental data on a quantitative level, irrespective of the size of the database used for comparison, precise estimation of experimental errors is necessary. This has not been possible due to the absence of a general consensus for using a standard set of conditions and reporting data among researchers within the protein folding community, until recently. Various research groups have made a combined effort to obtain the variability across and within laboratories by studying the kinetics of the same protein[153]. Using their data on the wild type and seven mutants of Fyn SH3 a mean error of ~0.26 (*log* units) (i.e. average standard deviation, corresponds to ~45% error) in mid-point folding rates is obtained.

In order to estimate errors in folding rates due to fitting kinetic data to two-state analysis, here, the experimental Chevron plots of 34 proteins are reproduced by digitization. The Chevron plot of each protein is then fitted to a two-state model and random noise with the same magnitude as the standard error of the fit (i.e. standard deviation of the difference between the best fit and original data) is added to the fitted curve. Next, the newly generated Chevron curve is subjected to two-state fit in the same manner as the original one. This procedure is repeated 50 times for each protein and from the distribution of each fitted parameter (k_f, k_u, m_f, m_u from which k_m is calculated) the associated standard deviation is calculated. This exercise yields a mean error of ~0.02 (*log* units) (i.e. average standard deviation, corresponds to ~5% error) in k_m for 34 proteins.

For some proteins folding has been investigated at a wide range of pH. The variation in k_m due to changes in pH obtained for spectrin SH3[154] and C-terminal

domain of protein L9[140] ranges from ~0.2-~0.7 (*log* units) (i.e. corresponding to 60-80% discrepancy). This may provide a rough estimate of the errors involved while comparing folding rates of different proteins not obtained at a common pH.

Propagation of the errors mentioned above gives a crude estimate of ~0.3-0.75 (*log* units) which is perhaps the upper limit. The reliability of this estimate (obtained from whatever very limited data available) is highly questionable because teasing out the contributions of different sources of error requires a large number of systematic and controlled studies (where variables are changed one at a time to observe their individual effects) and there are possibilities of certain errors canceling out. Nevertheless, it is assuring at present that the variation of experimental mid-point rates within one order of magnitude is very similar to the predictive power of the simple model used here. Such precision in reproduction of rates is helpful only in studying the general properties of folding. Even if the variation between predicted and experimental rates were randomly distributed around the average, the mean discrepancy would be 0.8. This suggests that, although the mean discrepancy decreases from 0.922 to 0.8 when structural details are included, there is still much room for improvement in predicting rates at the level of individual proteins.

To obtain an estimate of errors in prediction of rates from the use of different 3-D structures, 16 different X-ray structures of lysozyme (including those for mutants and those in which lysozyme forms complex with other moieties) are obtained. The mid-point folding rates of lysozyme predicted by the model (described here) using the information of residue-residue contacts (Cα-Cα) obtained from individual structures shows a deviation of ~0.05 (*log* units). Similarly the mean variation in mid-point rates

predicted from structural information derived from multiple NMR structures is evaluated to be ~ 0.06 (*log* units) for 18 proteins. This shows that the errors in the rate prediction due to the differences in 3D-structures are negligible.

However, the errors in theoretical calculations may depend on the level of structural details incorporated in the model, for example different definitions of atomic contacts. Figure 4.7A shows the comparison between the observed and calculated mean mid-point rates for alternative representations of protein structure used in the model: Cα-Cα (<0.6 nm), Cβ-Cβ (<0.8nm), contacts between side chain heavy atoms (<0.6nm), center of masses for side-chain heavy atoms (Cα onwards)(0.6nm), all heavy atoms, combination of all the above definitions of atomic contacts and backbone hydrogen bonds. The interesting point from Figure 4.7A and B is that the variation in predicted rates seems to increase with the decease in folding rates. In the present dataset most proteins folding slower than 1 second belong to either β or αβ class for which model predictions are more sensitive to the details of structural description. This is obvious as any small changes in the mean values $k_{\Delta H}$ and $k_{\Delta C_p}$ with different representations of structure produces relatively larger changes in the free energy barriers of proteins with larger barriers, thus manifesting as greater variation in predicted rates of slow folding proteins. In case of α-helical proteins smaller exponents are required to produce lower barriers. Any further small variations in these exponents produce insignificant changes in the curvature of the enthalpy functionals of α-helical proteins and hence in their barrier heights. Using Cα-Cα contacts results in the least mean discrepancy between theoretical and observed rates even when compared with the average across all atomic models

(compare Figure 4.5B and Figure 4.7A). Hence for all calculations in this analysis pertaining to rate prediction from structural details, Cα-Cα contacts are used. Interestingly other statistical models predicting folding rates from native structures have also found the coarse-grained Cα description to perform equally well as more detailed atomic representations of structure. But Cα description can reproduce only

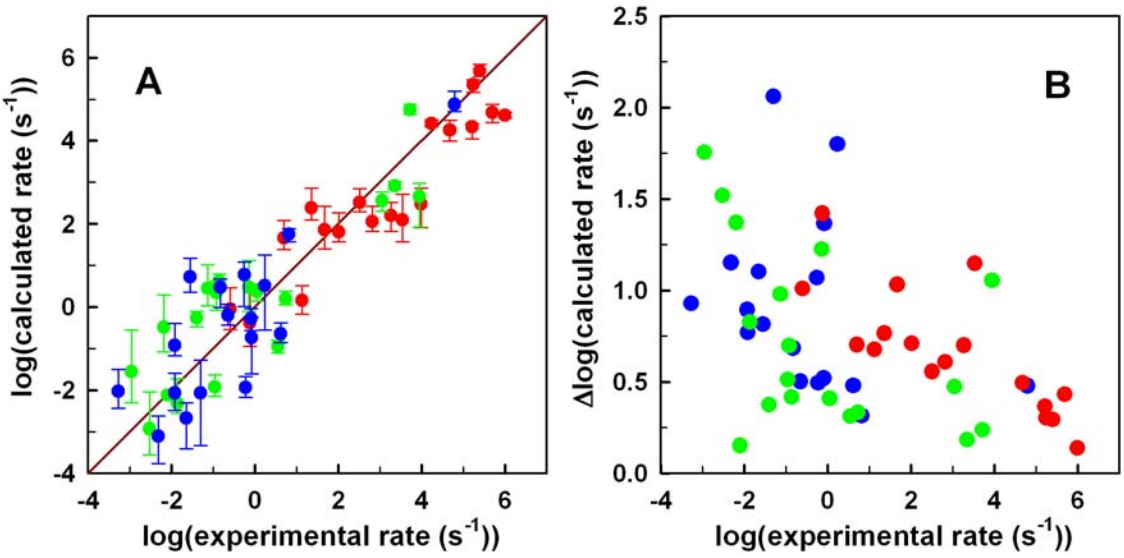


Figure 4.7 Comparison of calculated and observed rates at mid-point conditions for various atomic models

(A) Average mid-point rates predicted from using different representation of atomic contacts in the model are compared with experimental mid-point rates. The error bars represent the upper and lower limit of predicted rates. (B) Experimental midpoint rates are compared with the difference between the upper and lower limits of predicted rates (in A) at midpoint. The same color-coding as in previous Figures is followed.

the general behavior- the deviations in predicted rates are much larger at the level of individual proteins than the uncertainties in the observed rates. A possible reason for this limitation is the use of pair-wise contacts in prediction of rates. A description involving many-body interactions may be more suitable for representing protein structures. This becomes even more important with increasing complexity in protein topology as can be seen from Figure 4.7A. On the other hand, it is also possible that using more details may deteriorate the prediction due to the use of a noisy database. This effect is already seen from the increase in mean discrepancy between observed and predicted rates when sequence-dependent details are added in the model.

4.4.4 Prediction of folding rates at native conditions

Folding rates measured in the absence of denaturant are usually more prone to error than rates obtained at mid-transition. To calculate folding rates in the absence of denaturant using the simple model described here one more experimental parameter is required – protein stability at folding temperature i.e. unfolding free energy, the estimates of which are also associated with uncertainties. Hence to calibrate the heat capacity and enthalpy functionals the more reliable mid-point rates are used. Using stabilities reported in literature and the mean parameters obtained from the fitting of mid-point rates (using Cα-Cα contacts), folding rates in the absence of denaturant are calculated (Figure 4.8). The only adjustable parameter in this calculation is ΔH^0_{res} to reproduce equilibrium population that matches experimental stabilities. The calculated rate shows an 87% correlation with the observed rates.

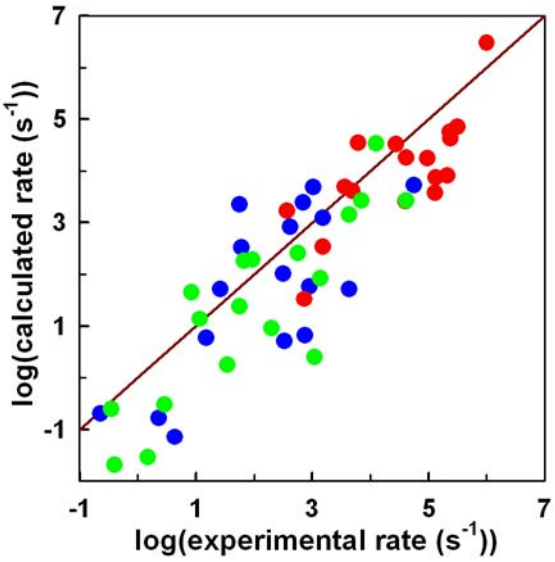


Figure 4.8 Comparison of calculated and observed rates in absence of denaturant

Rates in water calculated by the model (using Cα-Cα contacts) with the mean parameters obtained from the fitting of mid-point rates (Table 4.2) are compared with experimental folding rates measured in the absence of denaturant. The color-coding for α, β, and α-β classes is maintained from earlier Figures.

However for majority of the proteins the calculated rates are under-predicted with the mean values of $k_{\Delta H}$ and $k_{\Delta C_p}$ used for predicting mid-point rates. This is not surprising since folding barriers at mid-point denaturation are higher than folding barriers at more native-like conditions. Hence reproduction of folding rates in absence of denaturant would require lower values of $k_{\Delta H}$ and $k_{\Delta C_p}$ that would give rise to lower barriers. Keeping $k_{\Delta C_p}$ fixed to the same mean value obtained from prediction of mid-

point rates (i.e. assuming solvation effects to be similar in absence and presence of denaturant) the coefficient of the exponent of the enthalpy functional ($c_{\Delta H}$) is fitted for each individual protein to perfectly reproduce the observed experimental folding rates in absence of denaturant. The resulting $c_{\Delta H}$ ranges from –0.2 to 2.5 with mean 1.44, which is, as expected, lower than the mean value of 1.784 used from midpoint rate prediction.

Chapter 5: Estimation of conformational entropy from statistical analysis of protein structure database

5.1 Introduction

In spite of the importance of entropic factors in determining free energy barriers efforts towards quantitative understanding of protein folding entropy have been limited. Among thermodynamic parameters conformational entropy has remained one of the most difficult to evaluate. The change in entropy upon folding (ΔS) in aqueous environment is generally partitioned into two components: ΔS_{conf} associated with the loss in conformational freedom of the polypeptide chain and ΔS_{solv} is the de-solvation entropy arising from the burial of polar and non-polar groups (ΔS_{polar} and ΔS_{apolar})[155]. In earlier thermodynamic studies entropy and enthalpy changes of unfolding normalized with respect to number of residues were found to converge for a set of proteins at temperatures of 385 K and 373 K respectively[156]. Noticeably entropy of dissolution of liquid hydrocarbons and solid hydrophobic model compounds were also observed to converge around 385 K[41]. Due to this similarity it was hypothesized that at 385 K the only significant component i.e. hydrophobic (ΔS_{apolar}) contributions to ΔS were absent and, by difference, ΔS measured at 385 K corresponded only to conformational entropy ΔS_{conf}[41]. Although there have been controversies regarding the contributions from the polar groups, ΔS_{polar} have been, indeed, shown to be close to zero at ~335 K and have negligible contribution at 385 K. Using convergence temperatures of solvation entropies (ΔS_{polar} and ΔS_{apolar}) and parameterization of heat

capacity changes in terms of polar and apolar accessible surface areas given by Freire and co-workers[157] Akmal and Muñoz have calculated temperatures at which ΔS_{polar} and ΔS_{apolar} cancel out[62]. Hence at these temperatures (evaluated from heat capacity of activation obtained from fitting of temperature dependence of folding and unfolding rates) that were found to be only slightly higher than 385 K, ΔS was assumed to reflect only ΔS_{conf}. Estimate of average ΔS_{conf} per residue (~18 $J.mol^{-1}.K^{-1}.res^{-1}$)[62] from this analysis of kinetic data have been found to agree closely with those obtained by Robertson and Murphy using thermodynamic data (~17 $J.mol^{-1}.K^{-1}.res^{-1}$, see Figure 4.1)[41].

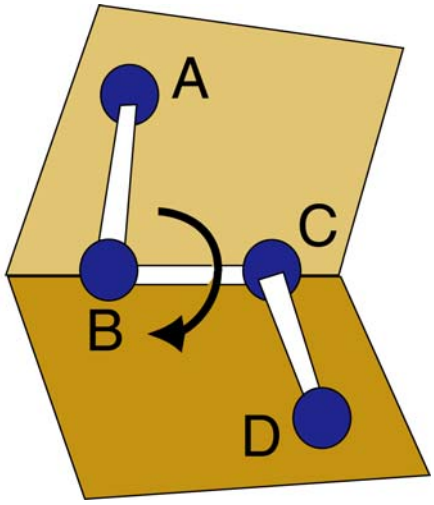
Theoretical estimates of ΔS_{conf} have been obtained from sampling of various conformational states for backbone and side chain ($\Delta S_{conf} = -R\sum_i p_i \ln p_i$) using Monte Carlo simulations and molecular mechanics force fields[158]. In a different approach side chain rotamer libraries and distribution of backbone dihedral angles derived from limited number of protein structures have been used[159]. These estimates have been successful in reproducing experimental helical and β-strand propensities of different amino acids[160].

Along similar lines, here, conformational entropies are evaluated in absolute terms from the statistical analysis of an expanded database of ~4000 protein structures. Conformational entropies estimated for each amino acid from this analysis directly form the sequence-dependent parameters of the model described in Chapter 4. In addition this analysis gives an opportunity to address questions such as: Are protein sequences and/or protein structures subjected to evolutionary selection to minimize or maximize conformational entropies?

5.2 Methods

The database of three-dimensional protein structures used in this analysis is a subset of the WHAT IF relational database[161]. The WHAT IF database has been built using an algorithm similar to the one developed by Hobohm and Sander[162] in which representative X-ray structures from PDB are selected based on their quality: R-factor, resolution and sequence homology. The database considered here is derived using a cutoff of 0.25 for the R-factor and contains 4013 X-ray structures with less than 30% sequence identity and less than 2.5A^{o} resolution.

Using Cartesian coordinates of relevant sets of four atoms, main chain (ϕ, ψ, ω) and side chain (χ_1,χ_2,χ_3,χ_4) dihedral angles are calculated. (See Figures A2 and A3 in appendix for the different side-chain angles applicable to each amino acid).



For example, to determine the dihedral angle between two bonds AB and CD about a common bond BC, the following vectors are calculated first from the x,y,z coordinates of the four atoms A,B,C,D:

$$\overrightarrow{AB} = B(x,y,z) - A(x,y,z) \quad \text{.....(5.1)}$$

$$\overrightarrow{BC} = C(x,y,z) - B(x,y,z) \quad \text{.....(5.2)}$$

$$\overrightarrow{CD} = D(x,y,z) - C(x,y,z) \quad \text{.....(5.3)}$$

Next, vectors defining the two planes are obtained from the cross product of respective bond vectors.

$$nABC = \overrightarrow{AB} \times \overrightarrow{BC} \quad \text{.....(5.4)}$$

$$nBCD = \overrightarrow{BC} \times \overrightarrow{CD} \quad \text{.....(5.5)}$$

The dihedral angle θ in degrees is then given by

$$\theta = \frac{180}{\pi}\left(\arccos\left(\frac{nABC \bullet nBCD}{|nABC||nBCD|}\right)\right. \quad \text{.....(5.6)}$$

if $nABC \bullet \overrightarrow{CD} < 0$, θ is negative.

The ϕ-ψ dihedral angle space of 899,172 amino acids is represented as the Ramachandran plot and divided into intervals of 9^{o}C in both directions (i.e. for ϕ and ψ) resulting in a 40×40 matrix. Each of the 1600 discrete regions can be addressed by the indices of the rows and columns of the matrix. Row and column indices run from 1 to 40 corresponding to values of –180 to +171 for ψ and ϕ respectively. For example, interval (1,1) corresponds to that region of the ϕ-ψ space including all values from –180 to –171 for both ϕ and ψ. The logarithm of number of hits for each region obtained from the database is shown in Figure 5.1.

The 40×40 matrix is next partitioned into 20 clusters using the K-means algorithm[163,164] implemented in MATLAB 6.5. This algorithm randomly assigns centroids/centers, one each for the specified number of clusters and calculates the Euclidean distance between each data point and every cluster centroid. An iterative algorithm aims at minimizing the sum of Euclidean squared distances within each

cluster by re-assigning/moving the data points between clusters at each step. Convergence is reached when no further changes can be made. K-means clustering forms mutually exclusive and compact partitions, however, the solution is not always optimum. Hence the algorithm is run several times and the result that approximately matches the natural clusters observed in ϕ-ψ distribution is chosen.

Similarly, one-dimensional matrices are built, one each for side chain dihedral angles, and each partitioned into 3 clusters. The overall distribution of each side chain dihedral angle and main chain ω for all amino acids is shown in Figure 5.2.

The probability of populating each cluster can be expressed as

$$p_i = \frac{\sum_j N_j}{N_t} \quad \text{.....(5.7)}$$

where N_j is the number of hits in region j of cluster i and N_t is the total number of hits over all clusters. The probability distribution obtained from a large database can be assumed to follow Boltzmann distribution. Each dihedral angle is assumed to attain either one of the two thermodynamic states: the native state corresponding to any one cluster and the nonnative state in which the dihedral angle can sample all other clusters in the Ramachandran plot except for the native cluster in question. Therefore the cost of fixing a dihedral angle in a particular cluster i is given by the difference between the entropy to be in all other clusters except for cluster i (nonnative state) and the entropy of cluster i (native state) (Equation 5.8).

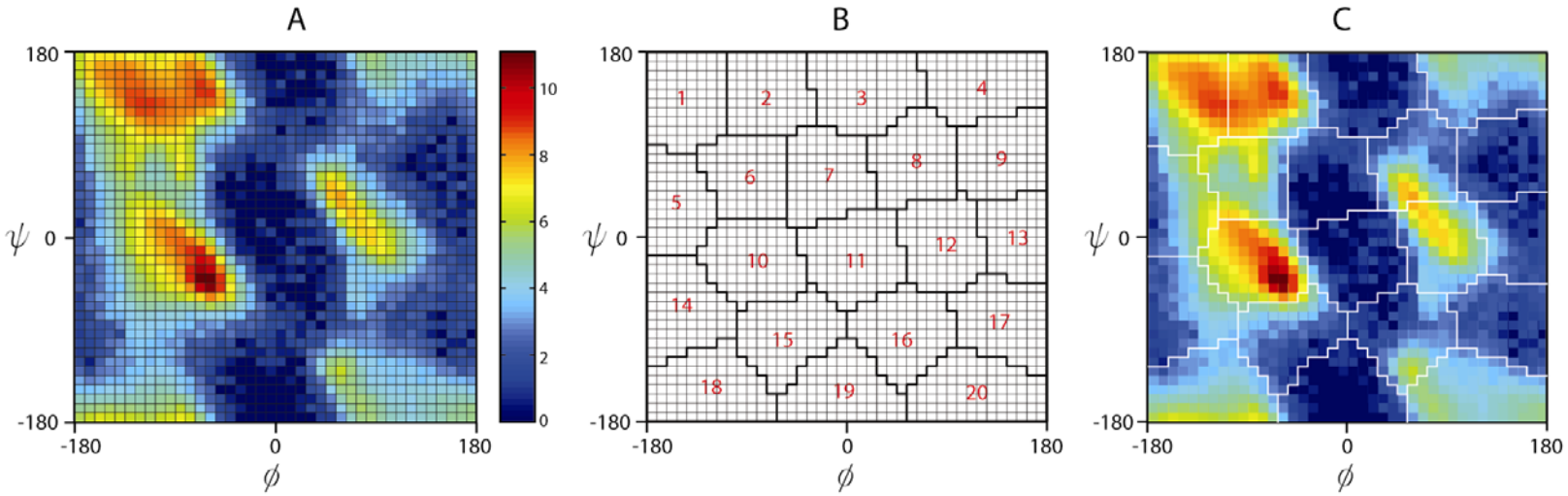


Figure 5.1 Distribution and clustering of ϕ-ψ dihedral angles

(A) The ϕ-ψ space is represented as a 40 × 40 matrix with each square corresponding to a region of 9^{o} × 9^{o}. The color bar indicates the value of logarithm of number of hits in each region. The highly populated areas correspond to α-helical and the β-strand conformations. (B)The ϕ-ψ space of (A) is shown to be divided into 20 clusters. Cluster indices are marked in red. (C) Super-imposition of (B) on (A) shows β-strand region divided into clusters 1 and 2 while α-helical region falls in cluster 10.Left-handed helical conformations are mainly included in clusters 8 and 12.

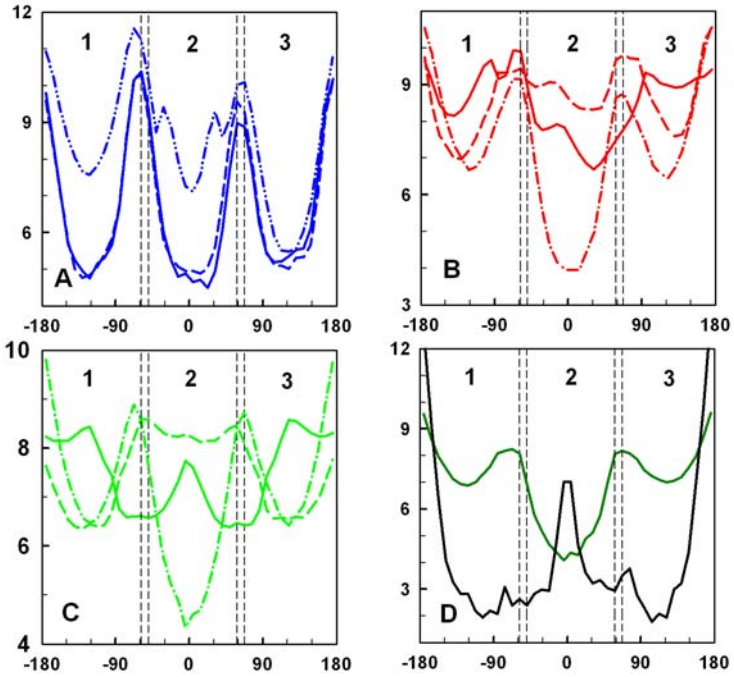


Figure 5.2 Distribution and clustering of χ and ω dihedral angles.

The logarithm of number of hits is plotted against dihedral angle values from –180 to +180. Dashed-dot lines show the distribution of χ angles (blue: χ_1; red: χ_2; green: χ_3). In certain amino acids there is branching at C_β, C_γ or C_δ atoms and thus each χ angle can have two alternative values χ_{x1} (dashed line) or χ_{x2} (solid line) where x=1,2 or 3 for χ_1, χ_2 or χ_3 respectively. For definition of each side chain dihedral angle see Figure A3. (D) shows the distribution of χ_4 (dark green) and ω (black) angles. Black dashed lines (around -60^o and +60^o) demarcate the regions corresponding to the three clusters.

$$\Delta S_i = S_{nonnative} - S_{native}$$

$$\Delta S_i = \left(-R\sum_i p_i \ln p_i - \left(-Rp_i \ln p_i\right)\right) - \left(-Rp_i \ln p_i\right)$$

$$\Delta S_i = -R\left(\sum_i p_i \ln p_i - 2p_i \ln p_i \right) \quad \text{.....(5.8)}$$

Distributions of ϕ-ψ angles, ω and side chain χ angles are generated from the protein database for each individual amino acid (see Figures A2 and A3 in appendix). The entropic costs of fixing the dihedral angles of a particular amino acid in any of the 20 different clusters (for ϕ and ψ) or 3 clusters (for χ's and ω) are evaluated from its probability distribution calculated from the total number of hits in each cluster and the overall total number of hits for that amino acid in the database. These are listed in Table 5.1 and 5.2.

5.3 Estimation of conformational entropies[d] of test Proteins

From 3-dimensional structures main chain and side chain dihedral angles are calculated for each protein. Depending on the values of the dihedral angles they are assigned to the clusters defined above. For each amino acid from a protein sequence the total conformational entropy is directly obtained from the sum of the backbone conformational entropies, ΔS_{bb}^{conf} and all of its relevant side chain conformational entropies, ΔS_{sc}^{conf} taken from Tables 5.1 and 5.2 for all residues. And therefore the total cost in conformational entropy for each protein (ΔS_{PDB}^{conf}) is simply obtained from the summation of conformational entropies of all its residues.

[d] In the text 'conformational entropy' actually refer to cost of conformational entropy (ΔS) unless otherwise specified.

Table 5.1 Backbone conformational entropies (ΔS_{bb}^{conf}) of different amino acids

CLUSTER NUMBER	GLY	ALA	VAL	LEU	ILE	MET	PRO	TRP	TYR	GLU
1	14.521	5.641	5.159	5.586	5.270	7.558	8.789	8.538	7.141	5.975
2	14.182	5.308	5.852	5.069	5.840	7.986	3.438	8.830	7.976	5.512
3	17.381	10.163	11.125	10.416	11.125	12.447	8.773	13.547	12.824	10.252
4	14.533	10.111	11.128	10.425	11.134	12.463	8.799	13.567	12.824	10.248
5	17.080	9.563	10.617	10.017	10.688	11.916	8.834	13.133	11.913	9.758
6	16.677	9.075	10.237	9.138	10.258	11.344	7.376	12.382	11.165	9.190
7	17.406	10.174	11.138	10.430	11.139	12.483	8.771	13.591	12.849	10.271
8	15.990	9.539	10.929	10.036	10.991	12.008	8.755	13.271	12.207	9.387
9	17.233	10.187	11.144	10.437	11.144	12.495	8.815	13.606	12.859	10.274
10	12.062	5.323	5.159	5.134	5.252	7.202	2.898	8.077	7.024	5.534
11	17.397	10.159	11.125	10.415	11.120	12.447	8.752	13.547	12.829	10.234
12	11.526	9.734	11.092	10.044	11.105	12.078	8.716	13.246	12.219	9.726
13	16.961	10.208	11.172	10.458	11.180	12.574	8.855	13.703	12.909	10.302
14	17.226	10.059	10.779	10.280	10.847	12.416	8.803	13.468	12.577	10.047
15	17.072	9.929	11.032	10.213	11.035	12.351	8.381	13.458	12.674	9.938
16	15.858	10.065	11.106	10.376	11.108	12.463	8.721	13.542	12.793	10.187
17	17.110	10.178	11.122	10.427	11.137	12.487	8.811	13.596	12.849	10.263
18	14.404	9.809	10.915	10.227	11.008	12.271	8.360	13.309	12.512	10.031
19	17.369	10.176	11.139	10.433	11.142	12.487	8.785	13.596	12.854	10.274
20	13.886	10.089	10.995	10.304	11.017	12.431	8.672	13.533	12.732	10.201
CLUSTER NUMBER	ASP	GLN	ASN	CYS	SER	THR	LYS	ARG	PHE	HIS
1	9.766	6.972	11.184	9.221	7.137	6.431	6.720	6.815	6.900	9.063
2	8.166	6.963	10.543	9.619	7.149	7.000	6.498	6.942	7.625	9.520
3	13.454	11.599	15.369	14.479	12.474	12.179	11.435	11.761	12.560	14.201
4	13.425	11.611	15.353	14.455	12.406	12.156	11.425	11.753	12.560	14.187
5	12.171	10.984	13.540	13.610	11.610	11.123	10.810	10.886	11.680	12.792
6	10.676	10.330	12.352	13.265	11.447	11.268	10.215	10.491	10.782	12.236
7	13.446	11.623	15.382	14.518	12.507	12.191	11.451	11.784	12.580	14.232
8	11.800	10.550	12.437	13.935	11.787	12.030	10.317	10.797	12.022	12.997
9	13.485	11.638	15.404	14.537	12.510	12.202	11.458	11.793	12.589	14.243
10	7.816	6.504	9.517	8.896	6.713	6.311	6.206	6.434	6.763	8.527
11	13.462	11.606	15.374	14.479	12.479	12.174	11.422	11.764	12.555	14.197
12	12.375	10.968	13.239	13.999	11.944	12.008	10.633	11.075	12.014	13.273
13	13.525	11.686	15.446	14.634	12.547	12.237	11.493	11.832	12.629	14.312
14	13.322	11.469	15.179	14.373	12.193	11.799	11.191	11.584	12.349	13.927
15	13.245	11.441	15.226	14.418	12.306	12.010	11.106	11.503	12.395	14.069
16	13.246	11.529	15.201	14.437	12.326	12.146	11.313	11.666	12.514	14.163
17	13.478	11.633	15.395	14.527	12.519	12.195	11.455	11.790	12.582	14.228
18	11.976	11.345	14.062	13.833	11.317	10.875	11.173	11.458	12.287	13.836
19	13.480	11.633	15.401	14.527	12.507	12.198	11.453	11.792	12.585	14.235
20	13.403	11.577	15.327	14.470	12.375	12.090	11.382	11.693	12.495	14.174

Table 5.2 Side chain conformational entropies (ΔS_{sc}^{conf}) of different amino acids.

ANGLE	CLUSTER NUMBER	GLY	ALA	VAL	LEU	ILE	MET	PRO	TRP	TYR	GLU
χ_1	1				2.340		2.447	-0.01	2.696	2.610	2.593
	2				2.479		2.407	0.015	2.820	3.051	2.450
	3				0.870		1.167	0.011	1.937	1.688	1.601
χ_{11}	1			1.092		2.329					
	2			4.258		0.799					
	3			1.859		2.766					
χ_{12}	1			2.530		1.127					
	2			2.587		3.353					
	3			3.514		2.382					
χ_2	1						1.751				1.080
	2						4.328				4.747
	3						1.317				1.150
χ_{21}	1				1.174	1.416			2.478	1.885	
	2				2.971	2.404			3.229	3.359	
	3				2.460	2.531			2.774	2.590	
χ_{22}	1				2.562				1.662	1.483	
	2				1.917				4.455	4.633	
	3				1.808				0.743	0.440	
χ_3	1						2.205				
	2						3.899				
	3						1.896				
χ_{31}	1										2.250
	2										2.700
	3										2.141
χ_{32}	1										2.242
	2										3.725
	3										2.360
χ_4	1										
	2										
	3										
ω	1	-0.41	-0.22	-0.40	-0.28	-0.45	-0.26	1.211	-0.15	-0.09	-0.22
	2	5.562	5.681	5.625	5.684	5.596	5.693	4.636	5.703	5.671	5.698
	3	0.523	0.300	0.442	0.322	0.490	0.299	1.274	0.229	0.247	0.275

Blank cells correspond to those side chain dihedrals not applicable to a particular type of amino acid. For example, both Gly and Ala have values only for peptide bond dihedrals but not for any side chain dihedrals. (see Figure A3 for the side chain dihedral angles applicable to each amino acid). Negative values for the cost in entropy in Tables 5.1 and 5.2, for ω dihedrals, reflect the rigidity of peptide bond, (for example cluster 1 corresponding to trans configuration is highly favorable).

Table 5.2 Side chain conformational entropies (ΔS_{sc}^{conf}) of different amino acids

(continued)

ANGLE	CLUSTER NUMBER	ASP	GLN	ASN	CYS	SER	THR	LYS	ARG	PHE	HIS
χ_1	1	2.598	2.537	2.545	2.729	2.834		2.519	2.542	2.590	2.681
	2	2.140	2.226	1.960	2.650	3.261		2.483	2.622	2.977	2.523
	3	1.772	1.509	1.706	2.083	2.855		1.355	1.401	1.588	1.906
χ_{11}	1						2.760				
	2						2.689				
	3						3.357				
χ_{12}	1						2.992				
	2						3.020				
	3						3.097				
χ_2	1		1.265					0.963	0.862		
	2		4.601					4.955	5.060		
	3		1.302					0.784	0.644		
χ_{21}	1	2.310		2.740						2.007	2.311
	2	2.594		2.822						3.248	3.743
	3	1.661		2.542						2.654	2.270
χ_{22}	1	1.643		2.141						1.587	2.384
	2	3.473		3.266						4.527	3.542
	3	2.487		2.682						0.541	2.575
χ_3	1							1.339	2.455		
	2							4.593	3.661		
	3							1.242	2.193		
χ_{31}	1		2.779								
	2		2.846								
	3		2.687								
χ_{32}	1		2.646								
	2		3.413								
	3		2.700								
χ_4	1							1.367	0.498		
	2							4.515	5.333		
	3							1.397	0.424		
ω	1	-0.29	-0.32	-0.14	-0.13	-0.08	-0.25	-0.28	-0.29	-0.21	-0.09
	2	5.639	5.657	5.692	5.704	5.694	5.659	5.667	5.655	5.683	5.716
	3	0.388	0.375	0.240	0.216	0.201	0.345	0.345	0.370	0.291	0.165

The mean conformational entropy per residue ($\Delta S^{conf}_{res,PDB} = \Delta S^{conf}_{PDB}/N$) for each protein forms the parameter $\Delta S^{n=0}_{res}$ of the simple model described in Chapter 4. $\Delta S^{conf}_{res,PDB(U-F)}$ i.e. the maximum difference in the entropy functional (Equation 4.3) reflects the total change in conformational entropy between the unfolded and native states and can be estimated from $\Delta S^{conf}_{res,PDB}$ using the following analytical relation:

$$\Delta S^{conf}_{res,PDB(U-F)} = R \ln\left(\exp\left(\frac{\Delta S^{conf}_{res,PDB}}{R} \right) + 1 \right) \quad \text{.....(5.9)}$$

(derivation shown in appendix)

$\Delta S^{conf}_{PDB(U-F)}$ obtained from ($\Delta S^{conf}_{res,PDB(U-F)}.N$) can then be directly compared with ΔS_{U-F} measured experimentally at 385 K.

5.4 Results and Discussion

5.4.1 Comparison of conformational entropies obtained from experiments and theory

Total conformational entropies $\Delta S^{conf}_{PDB(U-F)}$ have been calculated for the protein dataset used by Robertson and Murphy[41]. Table 5.3 shows the comparison between these theoretically calculated conformational entropies with those measured experimentally at 385 K ($\Delta S_{U-F}(385K)$). For this set of proteins $\Delta S^{conf}_{PDB(U-F)}$ exhibits a correlation of 96.3 % with $\Delta S_{U-F}(385K)$ (*p-value* < 1.0753e-030). R^2 of 0.9273 is

obtained, which indicates that 92.73% of the variability in $\Delta S_{U-F}(385K)$ is explained by the variability in theoretical estimates $\Delta S^{conf}_{PDB(U-F)}$.

Table 5.3 Comparison of conformational entropies from theory and experiment

Protein	PDB	N	$\Delta S_{U-F}(385K)$ $(Jmol^{-1}K^{-1})$	$\Delta S^{conf}_{PDB(U-F)}$ $(Jmol^{-1}K^{-1})$
1. α-chymotrypsin	5CHA	237	4420	4128
2. α-chymotrypsinogen	2CGA	245	3860	4352
3. α-lactalbumin	1HML	123	1910	2221
4. α-lactalbumin	1ALC	122	2400	2210
5. Acyl carrier protein	1ACP	77	1050	1294
6. Arabinose binding protein	1ABE	305	4480	5264
7. Arc repressor	1ARR	106	2000	1820
8. B1 domain of protein G	1PGB	56	886	1022
9. B2 domain of protein G	1PGX	56	932	1223
10. Barnase	1BNI	108	2450	1954
11. Barnase	1BNJ	109	2790	1957
12. Barstar	1BTA	89	1570	1552
13. Bovine Pancreatic Trypsin Inhibitor	5PTI	58	882	1003
14. Carbonic anhydrase B	2CAB	256	4530	4629
15. Chymotrypsin Inhibitor 2	1COA	64	1070	1121
16. Cytochrome b5	1CYO	88	1660	1599
17. Cytochrome c (horse)	1HRC	104	1910	1883
18. Cytochrome c (yeast iso-1)	1YCC	108	2000	1910
19. Cytochrome c (yeast iso-2)	1YEA	112	1700	1898
20. GCN4	2ZTA	62	1100	1058
21. Histidine containing protein	2HPR	87	1230	1480
22. Interleukin 1-β	6I1B	153	2410	2770
23. Lac repressor headpiece	1LCD	51	518	849
24. Lysozyme (human)	1LZ1	130	2250	2334
25. Lysozme (hen)	1LYS	129	2530	2315
26. Lysozyme (equine)	2EQL	129	2190	2293
27. Lysozyme (T4)	2LZM	164	3300	2900
28. Met repressor	1CMB	208	3030	3682
29. Myoglobin (horse)	1YMB	153	2280	2604
30. Myoglobin (whale)	4MBN	153	3470	2676
31. Myoglobin (whale)	1MBO	153	2380	2652

Table 5.3 Comparison of total conformational entropies from theory and experiment (continued)

Protein	PDB	N	$\Delta S_{U-F}(385K)$ $(Jmol^{-1}K^{-1})$	$\Delta S^{conf}_{PDB(U-F)}$ $(Jmol^{-1}K^{-1})$
32. 3rd domain of silver pheasant Ovomucoid	2OVO	56	891	977
33. Papain	9PAP	212	3570	3765
34. Parvalbumin	5CPV	108	1706	1866
35. Pepsin	5PEP	326	5910	5816
36. Pepsinogen	3PSG	365	6410	6409
37. Plasminogen K4 domain	1PMK	78	1670	1363
38. RNase T1	9RNT	104	2080	1885
39. RNase T1	8RNT	104	2210	1875
40. RNase A	3RN3	124	2090	2237
41. ROP	1RPR	126	2840	2184
42. Sac7d	1SAP	66	837	1116
43. α-Spectrin	1SHG	57	994	1038
44. Staphylococcus nuclease	1STN	136	2540	2380
45. Stefin A	1CYV	98	1720	1753
46. Stefin B	1STF	95	2080	1664
47. Subtilisin inhibitor	3SIC	107	2440	1780
48. Subtilisin BPN	2ST1	275	4120	4729
49. Tendamistat	3AIT	74	985	1283
50. Thioredoxin	2TRX	108	1600	1921
51. Trp repressor	2WRP	105	1590	1809
52. Trp repressor	3WRP	101	1590	1736
53. Ubiquitin	1UBQ	76	1040	1371

Experimental mean $\Delta S_{U-F}(385K)/N$ for 53 proteins is ~17.4 ± 3 ($\pm$ standard deviation) $Jmol^{-1}K^{-1}$ res^{-1} and is close to the theoretical mean $\Delta S^{conf}_{PDB(U-F)}/N$ of 17.7 ± 0.73 ($\pm$ standard deviation) $Jmol^{-1}K^{-1}$ res^{-1}. However, the spread in experimental mean $\Delta S_{U-F}(385K)$ per residue seems to be almost 4 (3/0.73) times higher than that observed in theoretical $\Delta S^{conf}_{PDB(U-F)}$ per residue. The values of conformational entropy per residue obtained by Robertson and Murphy are calculated from $\Delta S_{U-F}(385K)$, which are not directly measured but propagated to 385 K using ΔH_m and ΔC_p. Due to this, the errors in the estimation and/or calculation of ΔH_m and

ΔC_p are also propagated into $\Delta S_{U-F}(385K)$. Akmal and Muñoz estimated the entropy of activation for folding and unfolding from kinetic data available at different temperatures[62]. The mean conformational entropies per residue estimated by them for 6 proteins also show a larger spread (18 ± 4 $Jmol^{-1}K^{-1}$ res^{-1}). Using the simple model described in Chapter 4 the values for the entropic parameter $\Delta S_{res}^{n=0}$ are obtained for different proteins that perfectly reproduce their experimentally measured mid-point rates. The mean $\Delta S_{PDB(U-F)}^{conf}$ per residue derived from $\Delta S_{res}^{n=0}$ for the same set of proteins as used by other groups (Table 5.4) is ~19.7 ± 3.7 $Jmol^{-1}K^{-1}$ res^{-1} (for all 54 proteins in Table 4.1 used in rate prediction the mean is ~17.4 ± 2.7 $Jmol^{-1}K^{-1}res^{-1}$). This clearly shows that estimating conformational entropy per residue from empirical thermodynamic or kinetic data involves a larger variability and thus leads to an overestimation of the real variability in proteins. Estimates from distribution of dihedral angles show smaller variation around the mean and compare more closely with the (error-free) linear scaling of conformational entropy with protein size (showing a mean discrepancy of ~0.5 as compared to ~2.3 between experimental values and those obtained from linear scaling).

Hence, evaluating mean conformational entropy per residue for each protein from its structural information can be a reliable approach in sequence-dependent parameterization of the 1-D simple model described in Chapter 4. Toward this end, conformational entropies per residue are calculated for the protein dataset used in prediction of mid-point rates from the PDB files listed in Table 4.1.

Table 5.4 Comparison of estimates of conformational entropy per residue obtained from thermodynamic and kinetic data and from protein structure statistics

Protein	$\Delta S_{U-F}(385K)/N$ ($Jmol^{-1}K^{-1}res^{-1}$) (Robertson & Murphy)	ΔS_{res}^{conf} ($Jmol^{-1}K^{-1}res^{-1}$) (Akmal & Muñoz)	$\Delta S_{PDB(U-F)}^{conf}/N$ ($Jmol^{-1}K^{-1}res^{-1}$) (Calculated from structures)	ΔS_{res}^{conf} ($Jmol^{-1}K^{-1}res^{-1}$) (fitting protein folding rates with simple model)
1. CI2	16.719	24.267	17.519	23.852
2. α-Spectrin SH3	17.439	-	18.201	22.571
3. Tendamistat	13.311	-	17.334	-
4. CspB-Bs	-	15.899	17.672	13.678
5. N-PTL9	-	19.246	17.425	19.316
6. Protein L	-	15.062	17.592	21.332
7. FKBP12	-	20.502	17.763	17.399
8. GCN4	17.742	13.389	16.792	-

5.4.2 Comparison of conformational entropies of natural proteins and random heteropolymers.

From the number of occurrences of each amino acid in the database of ~4000 proteins, composition of natural sequences is obtained. Most frequently occurring amino acids are non-polar amino acids with the exception of Ile, Met, Cys and aromatic residues. Next in line are amino acids with polar groups followed by aromatic residues. Random polymers, on the other hand, do not show specific preferences and hence it can be assumed to have equivalent probability for all amino acids (i.e. 1/20).

Table 5.5 Mean conformational entropy per residue for proteins used in the prediction of folding rates

Protein	ΔS_{PDB}^{conf}	Protein	ΔS_{PDB}^{conf}
1. BBL	15.749		
2. BBL(H166W)	15.495	29. CspB-Bs	16.615
3. E3BD	14.975	30. CspB-Bc	17.206
4. E3BD(F166W)	15.923	31. CspB-Tm	16.949
5. POB	14.848	32. CspA	16.79
6. EngHD	16.849	33. Fibronectin	16.677
7. hTRF1	16.163	34. Tenascin	16.779
8. hRAP1	16.431	35. TI27	16.766
9. c-Myb	17.91	36. Twitchin	16.073
10. FSD	17.261	37. Tendamistat	16.191
11. Trp Cage	14.657	38. GPW	16.035
12. α-3D	15.419	39. mAcP	16.283
13. BdpA	16.231	40. ctAcP	17.34
14.Villin-HP35 (N27H)	15.894	41. CI2	16.442
15. $\lambda_{6\text{-}85}$	15.943	42. C-PTL9	16.254
16. ACBP	14.782	43. N-PTL9	16.334
17. Im9	16.679	44. Protein G	17.269
18. Im7	15.890	45. Protein L	16.524
19. Pin WW	16.682	46. Ubiquitin	17.027
20. YAP65	16.787	47. ADAh2	16.792
21. WW Prototype	17.070	48. U1A	16.749
22. FBP28 (W30A)	16.458	49. S6	16.826
23. α-Spectrin SH3	17.214	50. FKBP12	16.718
24. Fyn SH3	16.904	51. Hpr	16.217
25. Src SH3	17.096	52. Villin14T	16.852
26. PI3K SH3	16.982	53. RafRBD	16.935
27. ABP1 SH3	17.369	54. Prb (K5I/K39V)	16.202
28. Sso7d (Y34W)	16.900	55. BBA5	15.070

Table 5.6 Sequence composition of natural proteins and average conformational entropy for individual amino acids

Amino Acid	% occurence	Average conformational entropy per residue	Amino Acid	% occurence	Average conformational entropy per residue
Gly	7.6619	13.333	Asp	5.9159	16.16
Ala	8.3256	5.7448	Gln	3.676	16.648
Val	7.2295	10.166	Asn	4.3832	18.766
Leu	8.9085	12.284	Cys	1.4619	12.661
Ile	5.6513	12.385	Ser	5.9024	10.468
Met	1.961	14.332	Thr	5.5243	12.954
Pro	4.6587	5.0227	Lys	5.8989	13.223
Trp	1.4325	15.924	Arg	4.9261	13.432
Tyr	3.56	14.084	Phe	4.0286	13.898
Glu	6.577	14.693	His	2.3166	17.402

The manner in which each amino acid samples the conformational space is different in natural proteins resulting in a range of weighted entropic cost averaged over all clusters specified in this work (i.e. 20 for ϕ-ψ and 3 for side-chain and ω dihedrals). The general trend is longer side chains or side chains with bulky groups are more restricted and hence have a larger conformational entropic cost. Counter-intuitively Gly has a large value by virtue of its maximum amount of sampling of conformational space. Since several regions are populated, the probability of being in any one region is lower and thus the cost of entropy higher. To investigate whether protein sequences and protein structures are naturally selected such that the average cost in conformational entropy is minimized or maximized the following two approaches are considered:

(A) Assignment of random sequence to natural structures

The dihedral angles calculated from PDB structures of proteins listed in Table 5.5 fall in different regions of conformational space. For each protein the corresponding cluster number for every residue is noted. A random sequence is generated for each template having the same number of residues as the protein in question. This is accomplished by assuming equal probability for each amino acid. Each residue of the random sequence is assigned the same cluster as that of the residues in the original sequence. The cost in entropy for each residue is then calculated using Equation 5.8 and the probability distribution of each cluster for each amino acid given in Tables 5.1 and 5.2. This procedure is repeated 1000 times for each template yielding a value of 13.5 ± 0.38 $Jmol^{-1}K^{-1}res^{-1}$ for the average cost in conformational entropy per residue. In contrast, for the same set of proteins in Table 5.5 the mean entropic cost per residue is 16.5 ± 0.7. Hence it appears that natural sequences are selected to have higher per-residue entropy costs. Notwithstanding the energetic effects, this result is consistent with the fact of natural proteins having entropic barriers.

Interestingly when random sequences are generated for natural structural templates following the same distribution of amino acids as natural sequences an estimate for conformational entropy cost (13.1 ± 0.38 $Jmol^{-1}K^{-1}res^{-1}$) results that is similar to the one obtained using completely random sequences. Although this case simulates natural selection and overall sequence composition for 55,000 sequences (1000 sequences each for 55 proteins) is the same as the natural % occurrence for each amino acid, the individual sequence composition for each protein is of course not preserved. Hence it appears that the sequence specificity of natural proteins tends

to increase the average entropic cost per residue than that expected from a Boltzmann distribution.

(B) Assignment of random structures to natural sequences

In order to mimic the sequence composition of natural proteins, 1000 heteropolymer sequences are generated with 60 residues each according to the probability distribution of individual amino acids (Table 5.6). One out of 20 clusters for main chain dihedrals and one out of 3 clusters for side chain and ω dihedrals are picked at random for each residue such that each cluster has equal probability to get populated. Hence each cluster (of 20) in ϕ-ψ space has probability (p_i=1/20) and the entropic cost of fixing a residue in any one cluster as calculated from Equation 5.8 is 22.4 $Jmol^{-1}K^{-1}res^{-1}$. Similarly, for each side chain dihedral angle and ω, p_i=1/3 and the entropic cost is ~3 $Jmol^{-1}K^{-1}res^{-1}$ resulting in a total of 37.4 $Jmol^{-1}K^{-1}res^{-1}$ for each residue irrespective of the sequence identity. This number is more than twice the estimate expected for natural proteins, which is expected, since in case of natural proteins all the regions of the conformational space are not sampled with equal probability. Some regions are more highly populated than others while some regions are not populated at all.

When clusters are chosen at random but according to their probabilities obtained from the overall distribution of dihedral angles from natural proteins and conformational entropies are evaluated using sequence- and cluster-dependent values from Tables 5.1 and 5.2, a mean entropic cost per residue of 12.7 ± 0.5 $Jmol^{-1}K^{-1}res^{-1}$ is obtained for 1000 sequences. This is equivalent to identifying the amino acid sequence of polypeptide chain and using sequence-dependent average values from

Table 5.6. Applying this procedure to the amino acid sequences of proteins in Table 5.5 yields a very similar average cost per residue of 12.74 ± 0.45 $Jmol^{-1}K^{-1}res^{-1}$. Again this estimate is significantly lower than the value of 16.5 $Jmol^{-1}K^{-1}res^{-1}$ for the same dataset. These results show that even when sequence composition and sampling of conformational space of 1000 polypeptide chains overall mimics that found in natural proteins, random assignment of structure tends towards lower entropic cost per residue. This work also warrants against the use of average entropy costs for amino acids while calculating total conformational entropy, as has been done in previous theoretical studies[155]. Conformational entropy for each protein can be properly evaluated not just by identifying its sequence but also considering its unique structural information (i.e. the probabilities of its each amino acid for populating different regions of conformational space). It can be argued that per-residue entropic cost of 16.5 $Jmol^{-1}K^{-1}res^{-1}$ is only for a small protein dataset. But similar values are obtained even for a different set of 53 proteins used by Robertson and Murphy. Also, using 1000 sequences with 30 and 100 residues do not change the mean 12.7 $Jmol^{-1}K^{-1}res^{-1}$ significantly, only the standard deviation seem to decrease with increase in chain length. Hence, individual protein structure and its sequence appear to have evolved to have higher average conformational entropy cost per residue than that expected from a Boltzmann distribution obtained from a protein database.

Chapter 6: Analysis of protein folding experiments with 1-D free energy surface model

6.1 Introduction

The 1-D free energy surface simple model described in Chapter 4 provides the foundation for studying the general properties of protein folding. As demonstrated in section 4.4.2 the DSC profiles and Chevron plots simulated by the model exhibit the essential characteristics observed in thermal and chemical denaturation experiments. This suggests that the model can be directly used to analyze experimental data. Unlike the traditional chemical models, the 1-D free energy surface model does not presume the presence of a free energy barrier or the number of macrostates. More importantly the model provides an opportunity to obtain barrier heights from experimental data that can be compared to those extracted by other independent studies. Here, the original equilibrium thermal denaturation data from DSC experiments and kinetic data from chemical denaturation studies are accumulated for a group of proteins that include representatives from different structural classes as well as folding regimes. The DSC profiles and Chevron plots are subjected to direct fitting by the model. Recently the model is also applied to analyze data obtained from laser-induced T-jump studies on temperature dependence of relaxation rates of proteins folding in the microsecond timescale. Such analysis helps in the estimation of both free energy barrier heights as well as pre-exponential from folding rates. Additionally, diffusion on 1-D free energy surfaces to obtain relaxation rates provides information regarding the diffusion coefficients of different proteins.

6.2 Methods

The published data on DSC experiments and the dependence of relaxation rates on the concentration of chemical denaturant and temperature are digitized using DigitizeIt 1.5.8. Fitting of data to the model is performed using the 'lsqcurvefit' function (non-linear least square minimization) in Matlab 6.5.

6.2.2 Fitting DSC profiles

Determination of unfolded baselines

To fit experimental profiles the unfolded baseline is calculated according to the expression given by Privalov and Makhatadze[165]:

$$C_{p,U}(T) = \sum_{i}^{N} a_{res,i}\left(T-T_r\right)^3 + b_{res,i}\left(T-T_r\right)^2 + c_{res,i}\left(T-T_r\right) + d_{res,i} +$$
$$\sum_{i}^{N-1} a_{pb,i}\left(T-T_r\right)^3 + b_{pb,i}\left(T-T_r\right)^2 + c_{pb,i}\left(T-T_r\right) + d_{pb,i} +$$
$$a_{ter}\left(T-T_r\right)^3 + b_{ter}\left(T-T_r\right)^2 + c_{ter}\left(T-T_r\right) + d_{ter} \quad \text{..... (6.1)}$$

where T_r=273.15 K, N is the total number of amino acids and coefficients with subscripts *res*, *pb* and *ter* refer to the values specific to individual amino acids, peptide bond units and amino and carboxy terminals respectively. The assumption here is that in the unfolded conformation the heat capacity contributions from the individual components of the protein are additive. The heat capacity functional is assumed to have a linear temperature dependence of the form

$$\Delta C_p(n,T) = N\left(\Delta C_{p,res} + m_{\Delta C_{p,res}}\left(T-T_m\right)\right)\left[\left(\exp(k_{\Delta C_p}n)-1\right)\Big/\left(1-\exp(k_{\Delta C_p})\right)\right] \quad \text{..... (6.2)}$$

where the exponential part is similar to Equation 4.6 except that the unfolded state is taken to be the reference state here. The experimental DSC profile is then fitted to the following equation

$$\left\langle \Delta C_p(T) \right\rangle = \Delta C_p^{excess}(T) + \sum_{i(n=0)}^{i(n=1)} \Delta C_{p_i}\left(n,T\right) p_i(T,n) + C_{p,U}(T) \quad \text{..... (6.3)}$$

$\Delta C_p^{excess}(T)$ is obtained from Equation 4.15.

The fitting procedure involves 6 parameters: $k_{\Delta H}$, $k_{\Delta C_p}$, T_m, $\Delta C_{p,\,res}$, $m_{\Delta C_{p,res}}$ and one parameter that allows the shifting of the unfolded baseline in the up- or down direction.

Determination of folded baselines

In an alternative procedure the native state is assumed to be the reference state with its baseline having a linear temperature dependence of the form:

$$C_{p,N}(n,T) = N\left(C_{p,\,\mathrm{res}} + m_{C_{p,res}}\left(T - T_r\right)\right) \quad \text{..... (6.4)}$$

The above is used in combination with the temperature dependence of the heat capacity functional

$$\Delta C_p(n,T) = N\left(\Delta C_{p,\,res} + m_{\Delta C_{p,res}}\left(T - T_r\right)\right)\left[1 + \left(\exp(k_{\Delta C_p} n) - 1\right) \Big/ \left(1 - \exp(k_{\Delta C_p})\right)\right] \quad \text{..... (6.5)}$$

where T_r is the temperature corresponding to the first of the data point.

The experimental DSC profile is then fitted to the following equation

$$\left\langle \Delta C_p(T) \right\rangle = \Delta C_p^{excess}(T) + \sum_{i(n=0)}^{i(n=1)} \left(\Delta C_{p_i}\left(n,T\right) + C_{p,N}(n,T)\right) p_i(T,n) \quad \text{..... (6.6)}$$

This fitting procedure involves 7 parameters: $k_{\Delta H}, k_{\Delta C_p}$, T_m, $\Delta C_{p,res}, m_{\Delta C_{p,res}}$, $C_{p,res}, m_{C_{p,res}}$. In another trial only temperature-dependent native baseline is used keeping heat capacity functional constant with temperature.

6.2.3 Fitting Chevron plots

The procedure for generating Chevron plots is the same as outlined in section 4.2.3 and 4.2.5. Here, the 5 fitting parameters are $k_{\Delta H}$, E_d and d_m in Equation 4.10, C and j in Equation 4.11. E_d corresponds to m_{eq} from experiments and will serve as a common parameter if global fitting of equilibrium and kinetic unfolding data is performed. In the present analysis, however, as only the Chevron plots are fitted E_d is an independent parameter. d_m denotes the denaturant concentration at mid-transition. Relaxation kinetics after a chemical jump is calculated from diffusion on 1-D free energy surfaces obtained from Equation 4.10 in the same manner as in Chapter 3. The diffusion coefficient along the reaction coordinate is given by $D(n) = D/(\Delta n)^2$ and is assumed to be constant. The relaxation rates as a function of chemical denaturant concentration are then evaluated from the time corresponding to the decay of the total signal A to A/e. The phenomenological pre-exponential is obtained from the calculated rates and barrier heights as follows: $k = k_o \exp\left(-\Delta G^{\ddagger\text{-}U} \big/ RT\right)$.

6.3 Results and Discussion

6.3.1 Analysis of DSC thermograms and Chevron plots

The fitting of DSC thermograms is very sensitive to the choice of initial parameters and hence it is necessary to impose restrictions on the parameter space and ensure global minimization. Among thermodynamic parameters the cost per residue in conformational entropy is fixed to a mean value of 16.5 $Jmol^{-1}K^{-1}res^{-1}$. The variable parameter space is reduced by performing global fitting of DSC thermograms of all 15 proteins that include previously classified two state, fast folding and downhill folding proteins. In this procedure a common $k_{\Delta C_p}$ and $\Delta C_{p,res}$ is fitted for all proteins while allowing other parameters to float. Out of these parameters $k_{\Delta H}$ and T_m determine the thermodynamic properties specific to each protein (i.e. T_m determines the midpoint condition where the population on either side of the barrier equals 50% and $k_{\Delta C_p}$ controls the curvature of the enthalpy functional and thus the height of the free energy barrier at T_m) and the rest define the temperature dependence of the baselines. This exercise resulted in a mean value of 4.3 for $k_{\Delta C_p}$ and ~50 $J.mol^{-1}.K^{-1}$ for $\Delta C_{p,res}$. Now, using these values as the starting parameters and not allowing them to change by more than 5-10% as far as possible the DSC thermograms for each protein is fitted individually. In contrast to DSC profiles all the parameters used for fitting Chevron plots are highly coupled and specific for each protein thereby preventing global analysis. The ideal way to analyze chemical denaturation data is to do a global fit of equilibrium and kinetic data. This requires modeling the signal decay of the different probes chosen for studying each protein, which is not trivial in

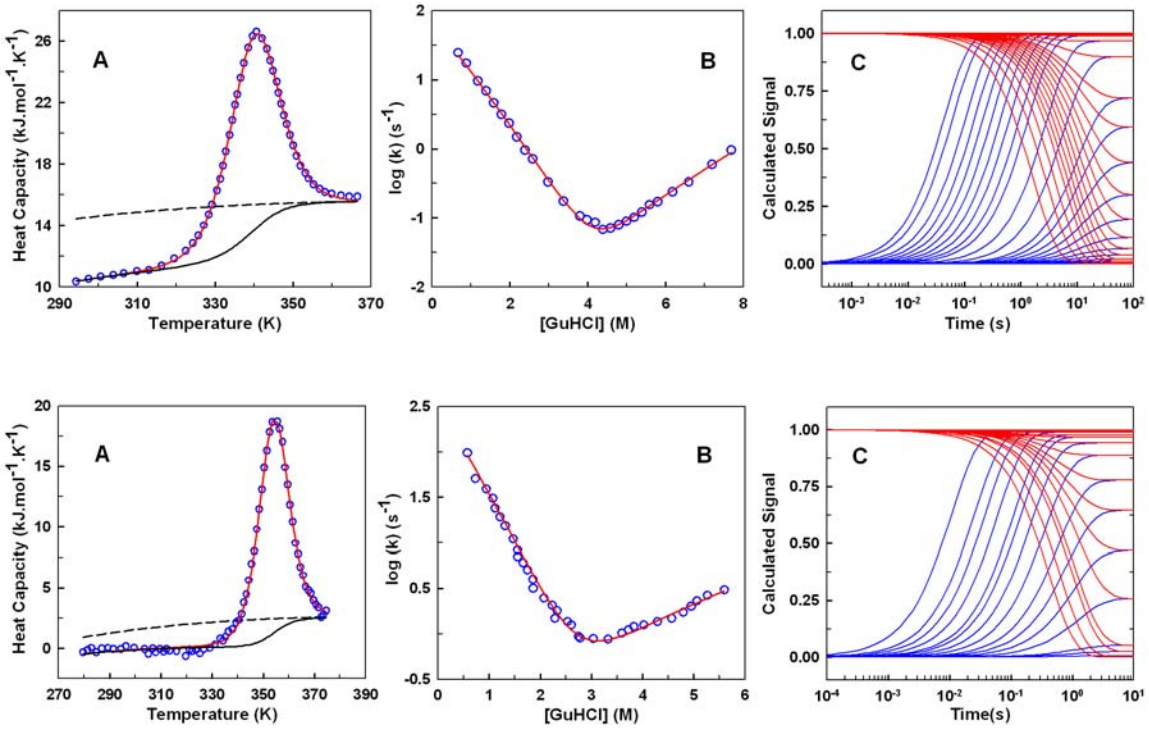


Figure 6.1 Fits of DSC thermograms and Chevron plots to 1-D free energy surface model

Upper Panel: Fyn SH3; Lower panel: Protein G.

(A) DSC data (blue circles) with fit (red line). The chemical baseline is shown in black whereas the dotted line is the unfolded baseline; (B) Experimental Chevron plot (blue circles) with fit (red line); (C) Relaxation traces obtained from the model after simulating various chemical-jumps starting from highly destabilizing (blue) and stabilizing (red) conditions. The relaxation rates obtained from the decays are plotted in (B).

a comparative analysis between several proteins. Besides, it is better to fit relaxation time courses (that are generally not reported) than the individual rate constants forming the Chevron plot. Here, Chevron plots are fitted individually by adjusting 5 parameters for each protein. Figures 6.1 and 6.2 show the fitted DSC profiles and Chevron plots for Fyn SH3, Protein G, CspB and POB (a homolog of E3BD). It is clear from these Figures that the model can reproduce both the sets of data very well. The sensitivity of fitting of DSC profiles to the baselines is investigated by several trials involving combinations of linear temperature dependence/independence of the heat capacity changes with temperature-dependent/independent unfolding baselines as well as native baselines. The performance of each combination is judged from the overall goodness of fit (sum of least squares) and the (non-) occurrence of baseline crossing in the transition region. The results from using either Equations 6.1-6.3 or 6.4-6.6 are comparable with the advantage of using one less parameter in the former case. Using a temperature independent heat capacity functional with a temperature-dependent native baseline does not fit the post-transition region as well as the rest of the profile because the higher slope of the pre-transition region results in slight overestimation of the slope of the post-transition tail.

For fitting Chevron plots, estimates of m_{eq} and midpoint denaturant concentration obtained from experiments are used as starting parameters for E_d and d_m respectively. Although the parameter space is reduced by grid analysis for C and j (where different values of C and j are fixed while allowing others to vary) and good fits are obtained, the fitting procedure is very cumbersome. Different combinations of the parameters give rise to similar local minima that are close in space and probably

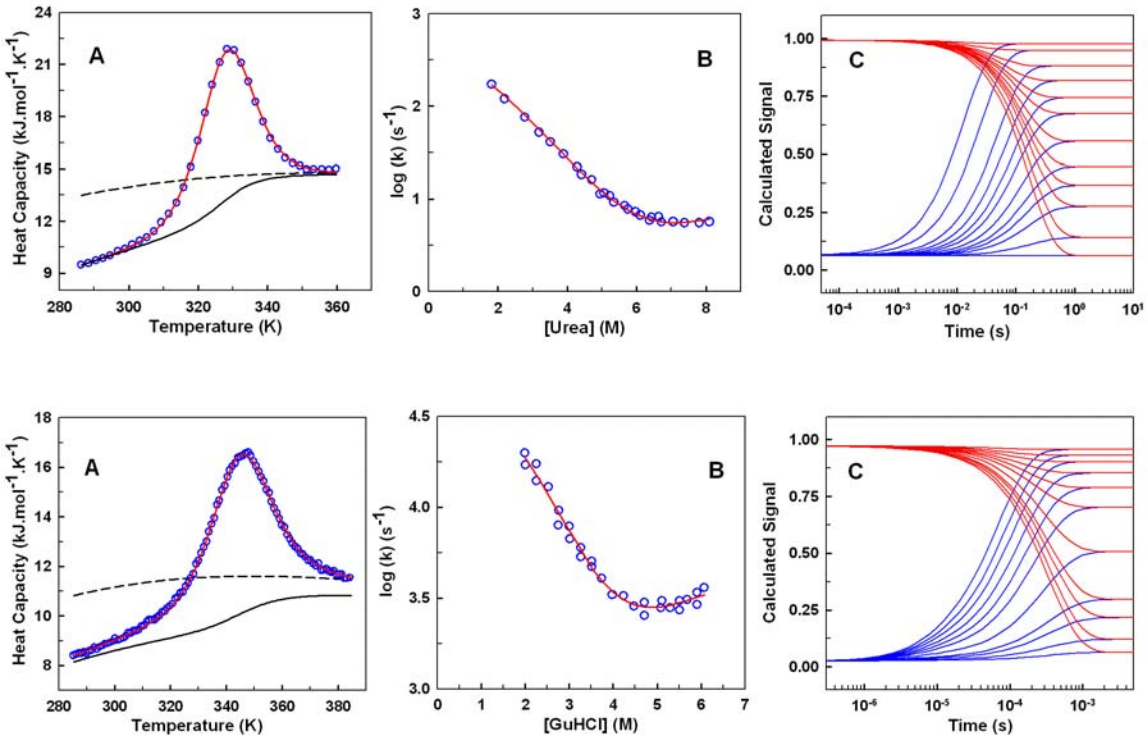


Figure 6.2 Fits of DSC thermograms and Chevron plots to 1-D free energy surface model

Upper Panel: CspB; Lower panel: POB (E3-binding domain of dihydro-lipoamide succinyl transferase).

(A) DSC data (blue circles) with fit (red line). The chemical baseline is shown in black whereas the dotted line is the unfolded baseline; (B) Experimental Chevron plot (blue circles) with fit (red line); (C) Relaxation traces obtained from the model after simulating various chemical-jumps starting from highly destabilizing (blue) and stabilizing (red) conditions. The relaxation rates obtained from the decays are plotted in (B).

in the broad global minima region. Hence more sophisticated optimization procedures such as genetic algorithms are required that explore the parameter space in a more efficient manner.

Table 6.1 lists the free energy barrier heights at mid-point denaturation for proteins for which DSC and Chevron plot fits are shown. The physical or chemical details about how urea or guanidinium salts denature the proteins are not known. Hence it is not possible to model the denaturing effects of each agent. Despite this limitation and using a general procedure to model the effects of chemical denaturation it is interesting to note that consistently for all proteins the barriers at their chemical midpoints are higher than that obtained at T_m. The 1-D free energy surface analysis provides the opportunity to compare barrier heights obtained from two different sets of experiments- one from thermal denaturation under equilibrium conditions and the other from kinetic chemical denaturation. It appears that chemical destabilization more strongly affects the free energy surface at mid-point conditions than thermal denaturation. The barrier heights obtained from fitting DSC thermograms are also very similar to earlier estimates obtained by Naganathan and Muñoz from analysis of DSC data with a variable barrier model[63]. The barrier height of Fyn SH3, the slowest folder in the group, is ~14RT at its chemical mid-point, consistent with its classification as a two-state folder. For POB, the barrier of ~3RT is again in agreement with lower barriers suggested by analytical theory for fast folders. The phenomenological pre-exponential (i.e. the dynamic term in the rate expression) of (1/7)-(1/10) μs^{-1} at temperatures close to 298 K obtained from diffusion on 1-D free energy surfaces of the above proteins is also compatible with earlier estimates. In

addition the relaxation traces shown in Figures 6.1C and 6.2C exhibit single exponential decay as expected for two-state proteins.

Table 6.1 Barrier heights obtained from the analysis of DSC thermogram and Chevron plots

Protein	N	k_m (s^{-1})	Barrier heights ($kJ.mol^{-1}$) from fitting DSC thermograms	Barrier heights ($kJ.mol^{-1}$) from fitting Chevron plots
Fyn SH3	67	0.08	~20	~34
Protein G	56	1.1	~16	~29
CspB	67	7.1	~15	~23
POB	51	3711	~2.5	~8

6.3.2 Analysis of temperature dependence of relaxation rates of fast folding proteins

For small organic molecules the pre-exponential can be obtained from the Arrhenius plots (logarithm of rates vs. inverse of temperature). However, for proteins this procedure does not work because both the dynamic and the energetic components depend on temperature. For proteins with low barriers any small change due to temperature effects produces relatively larger changes in the relaxation rate. Hence analyzing the temperature dependence of relaxation rate with free energy surface approach helps in discerning the contributions from barrier heights and the dynamic term. Using the 1-D free energy surface model the kinetic data from laser induced T-jump studies on proteins folding in the microsecond timescale have been recently analyzed by Athi N Naganathan[166]. The relaxation rates as a function of temperature are calculated by solving diffusion equation on free energy surfaces obtained from

Equation 4.9. From the analysis of free energy surfaces it is found that the fast folding proteins have marginal barriers at T_m and negligible barriers under native conditions. Also the effective diffusion coefficient, $D(n,T) = \left[k_0 \exp\left(-E_{a,res} N / RT\right)\right] /(\Delta n)^2$, is found to be strongly dependent on temperature with an activation energy, $E_{a,res}$, of ~1 $kJ.mol^{-1}$ per residue for all proteins. The folding speed limits for the proteins investigated shows a large variation ranging from ~4 to ~80 μs at room temperature. This analysis is thus able to explain the basis for the differences in rates of fast folding proteins (i.e. arising from either the differences in the barrier heights or the dynamic term). The success in interpreting experimental results on fast folding proteins demonstrates the empirical validation of the 1-D free energy surface models. Furthermore, folding barrier heights obtained from the simulations of chemical denaturation experiments shows a clear relation with experimental m-values, which can be used to classify proteins with different folding regimes[166].

Chapter 7: Summary and concluding remarks

Nucleation-elongation theory has proved extremely successful in explaining the thermodynamic properties of α-helix formation. However, there are very few examples of its applicability in interpreting kinetic experiments. Recently increased temporal resolution and use of more protein-like sequences have revealed rich kinetic behavior with multiphasic relaxation. The apparent relaxation times have been found to depend on the magnitude of thermal perturbation and on the specific regions of the peptide. These results have been interpreted with a diffusive search description that is incompatible with the nucleation elongation theory. In order to explain these results, here, a detailed kinetic model has been developed based on helix-coil theory that employs double sequence approximation and takes into account the sequence specific details from the AGADIR force field. Theoretical investigation carried out with this model clearly demonstrates that the observed complex kinetics are consequences of the inherent characteristics of helix-coil transition. This work resolves the controversy related to the mechanism of α-helix formation by proving that nucleation-elongation is still a valid description for α-helix formation[167]. The success of the kinetic model opens the possibility of testing experimentally the specific predictions made by the model, which in turn will help in the refinement of model parameters. The physical basis of the complex kinetics observed in experiments is explained with a simple 1-D projection of the free energy surface.

Furthermore it is demonstrated that diffusion on such 1-D free energy surface can reproduce all the kinetic results as predicted by the detailed model. This provides empirical validation for 1-D free energy surfaces on a quantitative level and for the number of helical peptide bonds as the reaction coordinate[168]. Length dependence of relaxation times upon T-jump has also been investigated with the diffusive model. There are a number of recent experiments that have examined the dependence of relaxation rates on peptide length, stability and sequence[169-171]. However, they still remain to be analyzed on a quantitative basis. In this respect the diffusive model that is computationally much less intensive than the detailed kinetic model (for example the order of the rate matrix is drastically reduced from 6196 to 20 for a 21-residue peptide) can prove extremely useful.

Computer simulations on folding of proteins and peptides provide a vast amount of structural information at atomistic level. This information when consolidated in low-dimensional projections of the free energy landscape provides an opportunity to elucidate the underlying principles and determine the thermodynamic properties of folding (i.e. relevant conformational ensembles and free energy barriers). Here, in a much simpler approach, a mean field model is formulated that accounts for the average energy and entropy as functions of an order parameter and generates 1-D free energy profiles. These profiles are able to predict a range of folding behaviors – from two-state-like with large barriers to completely downhill. The model is consistent with the experimental observations of linear scaling of thermodynamic parameters and convergence of entropy at 385 K. This model explains the origin of the relationship of protein size and native topology with folding

rates. By directly relating the only model parameter that determines the free energy barrier to protein size and 3-D structure, the model is able to predict protein folding rates within a factor of 6. This precision is sufficient to study general folding properties and can be improved by replacing the average values of thermodynamic parameters by sequence-dependent ones. The simplicity of the model facilitates the incorporation of sequence-specific energetic details. This allows determining the contributions of size, structure and energetics and opens the avenue for testing empirical force fields for prediction of protein kinetics. An algorithm is developed that automatically extracts Cartesian coordinates from PDB files according to a selected atomic description (i.e. Cα, Cβ or heavy atoms); calculates the contact maps, uses the information from these contact maps to generate 1-D free energy profiles and calculates folding rates by solving the diffusion equation on these free energy profiles. This opens the exciting possibility of scanning protein structure databases to identify fast folding proteins.

The model also describes the effects of thermal and chemical denaturation in a quantitative manner making it a suitable tool for direct analysis of folding experiments. A recent application of the model in the analysis of temperature dependence of rates of fast folding proteins has been successful and revealed the separate contributions of energetic (barrier height) and dynamic (diffusion coefficient) terms to folding rates[166]. Using the model to analyze chemical denaturation experiments of fast folding proteins and their mutants has demonstrated that these proteins show systematic deviations from two-state behaviors (i.e. ratio of

kinetic and equilibrium m-values is lower than 1 and a decrease in the m-values with the increase in the folding rates.

These 1-D free energy surface models once combined with improved minimization procedures can be implemented in a web-based application and made accessible to the protein folding community for the analysis of new equilibrium and kinetic experiments.

Appendices

Table A1. Experimental protein folding rates at thermal or chemical mid-point and in absence of denaturant

Protein	k_m (s^{-1})	T_m (K)	d_m (M)	T (K)	k_f(s^{-1})
BBL	2.00E+06	325		298	6.25E+04
BBL (H166W)	1.97E+04		3.25	298	1.30E+05
E3BD	3.50E+04	325.4		298	1.80E+04
E3BD (F166W)	6.53E+02		4.86	298	2.75E+04
POB	3.71E+03		4.00	298	2.10E+05
EngHD	6.92E+03		2.96	298	3.99E+04
EngHD	9.60E+04	325.14		298	3.75E+04
hTRF1	1.00E+01		2.96	298	3.70E+02
hRAP1	9.38E+01		3.78	298	3.60E+03
c-Myb	4.59E+01		5.10	298	6.20E+03
FSD	3.30E+05	313		298	4.17E+04
Trp Cage	1.00E+06	316		295.7	2.40E+05
α3D	5.00E+05	346.2		328	3.16E+05
BdpA	2.08E+02		3.52	298	9.68E+04
Villin-HP35 (N27H)	3.50E+05	342		300	2.33E+05
$\lambda_{6\text{-}85}$	1.33E+03		2.74	310	4.90E+03
ACBP	1.50E+00		1.84	298	1.05E+03
Im9	5.08E-01		4.76	298	1.53E+03
Im7	2.67E+01		2.63	298	7.35E+02
Pin WW	1.05E+04	334		314.3	1.25E+04
YAP65	2.25E+03		2.94	298	4.30E+03
WW Prototype	4.53E+03		3.14	298	7.00E+03
FBP28 (W30A)	1.79E+04		4.72	298	4.10E+04
α-Spectrin SH3	2.77E-01		4.38	298	8.41E+00
Fyn SH3	8.03E-02		4.15	293	9.43E+01
Src SH3	1.45E+00		2.58	295	5.67E+01
PI3K SH3	1.60E-02		1.55	293	3.53E-01
ABP1 SH3	2.40E-01		1.78	298	1.17E+01

Table A1. Experimental protein folding rates at thermal or chemical mid-point and in absence of denaturant (continued)					
Protein	**k_m (s^{-1})**	**T_m (K)**	***d_m* (M)**	**T (K)**	**k_f(s^{-1})**
Sso7d (Y34W)	3.46E+00		3.68	293	1.04E+03
CspB-Bs	7.10E+00		5.96	288	1.09E+03
CspB-Bc	2.27E+00		2.57	298	1.37E+03
CspB-Tm	1.48E-01		3.36	298	5.65E+02
CspA	1.08E+01		5.01	298	1.99E+02
Fibronectin	2.21E-01		0.40	298	4.00E-01
Tenascin	2.75E-02		3.63	293	2.90E+00
TI27	2.20E-03		3.42	298	3.50E+01
Twitchin	5.97E-03		4.47	293	1.50E+00
Tendamistat	1.29E-02		6.58	298	6.66E+01
GPW	1.25E+05	337		315	5.56E+04
mAcP	1.07E-03		3.77	301	2.30E-01
ctAcP	2.37E-02		2.58	301	2.31E+00
CI2	5.63E-02		4.18	298	5.63E+01
C-PTL9	2.41E-02		6.16	298	2.63E+01
N-PTL9	1.31E+01		6.29	298	6.99E+02
Protein G	1.11E+00		2.62	295	4.14E+02
Protein L	2.95E-01		2.35	295	6.06E+01
Ubiquitin	4.49E-01		3.86	298	1.53E+03
ADAh2	8.32E+00		4.40	298	7.57E+02
U1A	2.95E+00		3.44	298	3.16E+02
S6	4.49E-02		7.94	298	3.32E+02
FKBP12	9.62E-03		3.63	298	4.30E+00
Hpr	1.00E-01		2.01	293	1.49E+01
Villin14T	1.62E+00		3.89	310	8.98E+02
RafRBD	1.68E+00		6.24	298	4.27E+03
Prb (K5I/K39V)				347	1.00E+06
BBA5				298	1.33E+05

Table A2. Summary of structural information of proteins used in the analysis

Protein	Fold	Superfamily	Family	Architecture	Topology	Homology
1. BBL	PSBD of 2-oxo acid dehydrogenase complex	PSBD of 2-oxo acid dehydrogenase complex	E3BD of dihydro-lipoamide succinyl transferase	Irregular	Dihydrolipoamide Transferase	Dihydrolipoamide Transferase
2. BBL (H166W)	PSBD of 2-oxo acid dehydrogenase complex	PSBD of 2-oxo acid dehydrogenase complex	E3BD of dihydro-lipoamide succinyl transferase	Irregular	Dihydrolipoamide Transferase	Dihydrolipoamide Transferase
3. E3BD	PSBD of 2-oxo acid dehydrogenase complex	PSBDof 2-oxo acid dehydrogenase complex	E3BD of dihydro-lipoamide acetyl transferase	Irregular	Dihydrolipoamide Transferase	Dihydrolipoamide Transferase
4. E3BD (F166W)	PSBD of 2-oxo acid dehydrogenase complex	PSBD of 2-oxo acid dehydrogenase complex	E3BDof dihydro-lipoamide acetyl transferase	Irregular	Dihydrolipoamide Transferase	Dihydrolipoamide Transferase
5. POB	PSBD of 2-oxo acid dehydrogenase complex	PSBD of 2-oxo acid dehydrogenase complex	E3BD of dihydro-lipoamide succinyl transferase	Irregular	Dihydrolipoamide Transferase	Dihydrolipoamide Transferase
6. EngHD	DNA-RNA-binding 3-helical bundle	Homeodomain-like	Homeodomain	Orthogonal bundle	Arc Repressor Mutant, subunit A	Homeodomain-like
7. hTRF1	DNA-RNA-binding 3-helical bundle	Homeodomain-like	DNA-binding domain of human telomeric protein, hTRF1	Orthogonal bundle	Arc Repressor Mutant, subunit A	Homeodomain-like
8. hRAP1	DNA-RNA-binding 3-helical bundle	Homeodomain-like	Myb/SANT domain	Orthogonal bundle	Arc Repressor Mutant, subunit A	Homeodomain-like
9. c-Myb	DNA-RNA-binding 3-helical bundle	Homeodomain-like	Myb/SANT domain	Orthogonal bundle	Arc Repressor Mutant, subunit A	Homeodomain-like
10. FSD	Zinc finger based ββα motif	Zinc finger based ββα motif	Zinc finger based ββα motif	-	-	-
11. Trp Cage	Trp-cage mini protein	Trp-cage mini protein	Trp-cage mini protein	-	-	-

Table A2. Summary of structural information of proteins used in the analysis (continued)

Protein	Fold	Superfamily	Family	Architecture	Topology	Homology
12. α-3D	Three helix bundle	Three helix bundle	Three helix bundle	Up-down bundle	Methane Monooxygenase Hydroxylase; Chain G, domain 1	Three helix bundle
13. BdpA	Immunoglobulin/ albumin binding domain-like	Bacterial Immunoglobulin/ albumin binding domains	Immunoglobulin binding protein A modules	Up-down bundle	Single α-helices involved in coiled-coils or other helix-helix interfaces	Complex(Skeletal muscle/Muscle protein)
14. Villin-HP35 (N27H)	VHP Villin headpiece domain	VHP Villin headpiece domain	VHP Villin headpiece domain	-	-	-
15. $\lambda_{6\text{-}85}$	Lambda repressor-like DNA-binding domains	Lambda repressor-like DNA-binding domains	Phage repressors	Orthogonal bundle	434 Repressor (Amino terminal domain)	Lyase
16. ACBP	ACBP- like	ACBP	ACBP	Up-down bundle	ACBP	Structural protein
17. Im9	Acyl-carrier protein-like	Colicin E immunity proteins	Colicin E immunity proteins	Orthogonal bundle	Non-ribosomal Peptide Synthetase Peptidyl Carrier Protein; Chain A	Immune System
18. Im7	Acyl-carrier protein-like	Colicin E immunity proteins	Colicin E immunity proteins	Orthogonal bundle	Non-ribosomal Peptide Synthetase Peptidyl Carrier Protein; Chain A	Immune System
19. Pin WW	WW domain-like	WW domain	WW domain	Single Sheet	Ubiquitin Ligase Nedd4; Chain: W;	Complex (Isomerase/ Dipeptide)
20. YAP65	WW domain-like	WW domain	WW domain	Single Sheet	Ubiquitin Ligase Nedd4; Chain: W;	Ligase
21. WW Prototype	WW domain-like	WW domain	WW domain	Single Sheet	-	-

Table A2. Summary of structural information of proteins used in the analysis (continued)

Protein	Fold	Superfamily	Family	Architecture	Topology	Homology
22. FBP28 (W30A)	WW domain-like	WW domain	WW domain	Single Sheet	-	-
23. α-Spectrin SH3	SH3-like barrel	SH3-domain	SH3-domain	Roll	SH3-type barrels	SH3 domains
24. Fyn SH3	SH3-like barrel	SH3-domain	SH3-domain	Roll	SH3-type barrels	SH3 domains
25. Src SH3	SH3-like barrel	SH3-domain	SH3-domain	Roll	SH3-type barrels	SH3 domains
26. PI3K SH3	SH3-like barrel	SH3-domain	SH3-domain	Roll	SH3-type barrels	SH3 domains
27. ABP1 SH3	SH3-like barrel	SH3-domain	SH3-domain	Roll	SH3-type barrels	SH3 domains
28. Sso7d (Y34W)	SH3-like barrel	Chromodomain- like	Histone-like proteins from archaea	Barrel	OB fold Dihydrolipoamide Acetyltransferase,E2P	Peptide Binding Protein
29. CspB-Bs	OB-fold	Nucleic acid- binding proteins	Cold shock DNA-binding domain-like	Barrel	OB fold Dihydrolipoamide Acetyltransferase,E2P	Nucleic acid-binding proteins
30. CspB-Bc	OB-fold	Nucleic acid- binding proteins	Cold shock DNA-binding domain-like	Barrel	OB fold Dihydrolipoamide Acetyltransferase,E2P	Nucleic acid-binding proteins
31. CspB-Tm	OB-fold	Nucleic acid- binding proteins	Cold shock DNA-binding domain-like	Barrel	OB fold Dihydrolipoamide Acetyltransferase,E2P	Nucleic acid-binding proteins
32. CspA	OB-fold	Nucleic acid- binding proteins	Cold shock DNA-binding domain-like	Barrel	OB fold Dihydrolipoamide Acetyltransferase,E2P	Nucleic acid-binding proteins
33. Fibronectin	Immunoglobulin-like β-sandwich	Fibronectin type III	Fibronectin type III	Sandwich	Immunoglobulin-like	Fibronectin type III
34. Tenascin	Immunoglobulin-like β-sandwich	Fibronectin type III	Fibronectin type III	Sandwich	Immunoglobulin-like	Fibronectin type III
35. TI27	Immunoglobulin-like β-sandwich	Immunoglobulin	I set domains	Sandwich	Immunoglobulin-like	Immunoglobulins

Table A2. Summary of structural information of proteins used in the analysis (continued)

Protein	Fold	Superfamily	Family	Architecture	Topology	Homology
36. Twitchin	Immunoglobulin-like β-sandwich	Immunoglobulin	I set domains	Sandwich	Immunoglobulin-like	Immunoglobulins
37. Tendamistat	α-amylase inhibitor tendamistat	α-amylase inhibitor tendamistat	α-amylase inhibitor tendamistat	Sandwich	Immunoglobulin-like	α-amylase inhibitor
38. GPW	Head to tail joining protein W	Head to tail joining protein W	Head to tail joining protein W	Head to tail joining protein W	-	-
39. mAcP	Ferredoxin-like	Acyl phosphatase-like	Acyl phosphatase-like	2 layer sandwich	α-β plaits	Metal transport
40. ctAcP	Ferredoxin-like	Acyl phosphatase-like	Acyl phosphatase-like	2 layer sandwich	α-β plaits	Metal transport
41. CI2	CI2 family of serine protease inhibitors	CI2 family of serine protease inhibitors	CI2 family of serine protease inhibitors	2 layer sandwich	Trypsin Inhibitor V; Chain A	Trypsin Inhibitor V; subunit A
42. C-PTL9	Ribosomal protein L9 C-domain	Ribosomal protein L9 C-domain	Ribosomal protein L9 C-domain	3-layer (αβα) sandwich	Ribosomal protein L9 domain1	Ribosomal protein L9 domain1
43. N-PTL9	L9 N-domain-like	Ribosomal protein L9 N-domain	Ribosomal protein L9 N-domain	Roll	Ribosomal protein L9 domain2	Ribosomal protein L9 domain2
44. Protein G	β-Grasp (Ubiquitin-like)	Immunoglobulin-binding domains	Immunoglobulin-binding domains	Roll	Ubiquitin-like (UB Roll)	Immunoglobulin-binding proteins
45. Protein L	β-Grasp (Ubiquitin-like)	Immunoglobulin-binding domains	Immunoglobulin-binding domains	Roll	Ubiquitin-like (UB Roll)	Immunoglobulin-binding proteins
46. Ubiquitin	β-Grasp (Ubiquitin-like)	Ubiquitin-like	Ubiquitin-related	Roll	Ubiquitin-like (UB Roll)	Chromosomal Protein
47. ADAh2	Ferredoxin-like	Protease-propeptides/inhibitors	Pancreatic procarboxy peptidase activation domain	3-layer (αβα) sandwich	Aminopeptidase	Zinc peptidase

Table A2. Summary of structural information of proteins used in the analysis (continued)

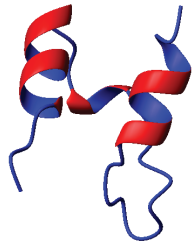
Protein	Fold	Superfamily	Family	Architecture	Topology	Homology
48. U1A	Ferredoxin-like	RNA-binding domain	Canonical RNA-binding domain	2 layer sandwich	α-β plaits	RNA binding protein
49. S6	Ferredoxin-like	Ribosomal protein S6	Ribosomal protein S6	2 layer sandwich	α-β plaits	Ribosomal protein
50. FKBP12	FKBP-like	FKBP-like	FKBP immunophilin/ proline isomerase	Roll	Chitinase A; domain 3	Isomerase
51. Hpr	Hpr-like	Hpr-like	Hpr-like	2 layer sandwich	Histidine containing protein; Chain A	Phosphotransferase
52. Villin14T	Gelsolin-like	Actin depolymerizing proteins	Gelsolin-like	3-layer (αβα) sandwich	Severin	Severin
53. RafRBD	β-Grasp (Ubiquitin-like)	Ubiquitin-like	Ras-binding domain	Roll	Ubiquitin-like (UB Roll)	Chromosomal Protein
54. Prb (K5I/K39V)	Immunoglobulin/ albumin binding domain-like	Bacterial Immunoglobulin /albumin binding domains	GA module, albumin binding domain	Orthogonal bundle	Helicase, Ruva protein; domain 3	Albumin binding domain
55. BBA5	Zinc finger based ββα motif	Zinc finger based ββα motif	Zinc finger based ββα motif	-	-	-

Abbreviations: BD: Binding domain; PSBD: Peripheral subunit-binding domain

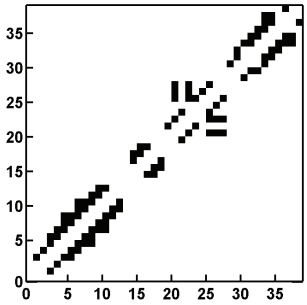
Figure A1 Three-dimensional structures and contact maps of proteins used in the analysis

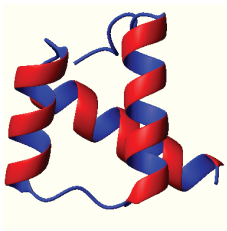
All Cα-Cα contacts within 0.6 nm are shown in the left panel. The structures plotted with MOLMOL are shown in the right panel. The panels are labeled according to the PDB file names listed in Table 4.1.

(See Next 6 pages)

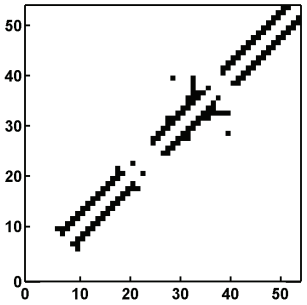


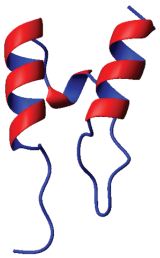
2CYU



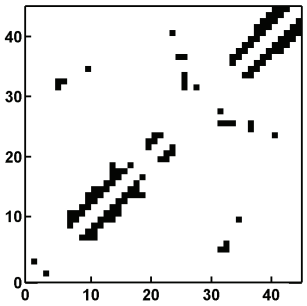


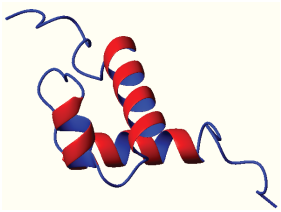
1ENH



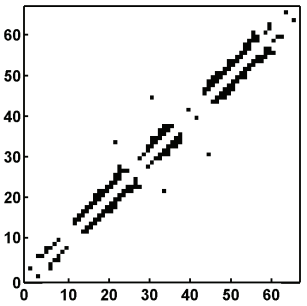


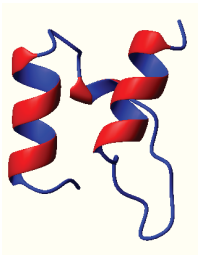
2BTH



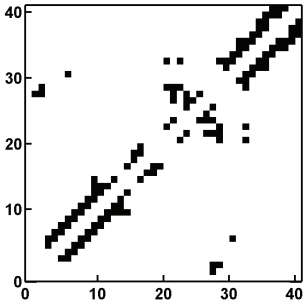


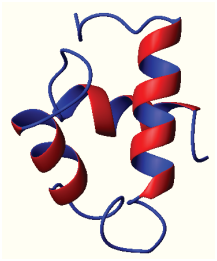
1ITY



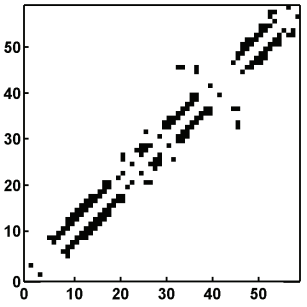


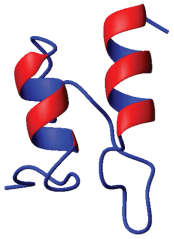
1EBD



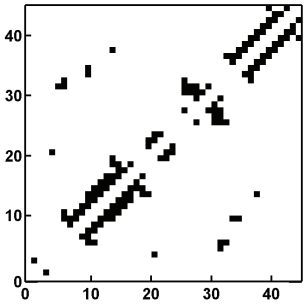


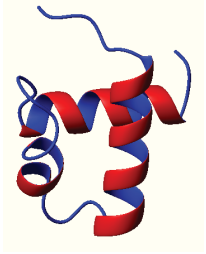
1FEX



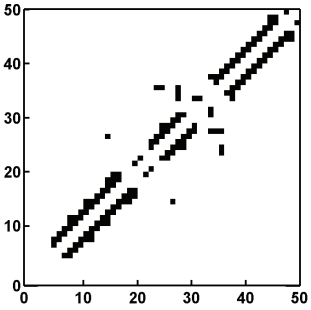


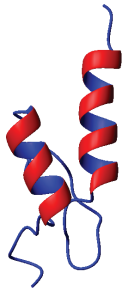
1W4E



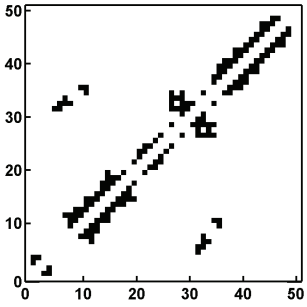


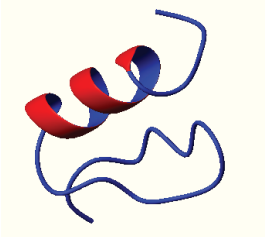
1GUU



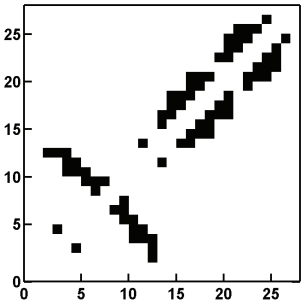


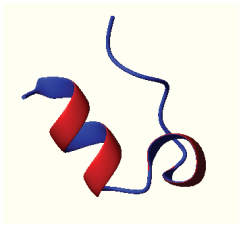
1W4J



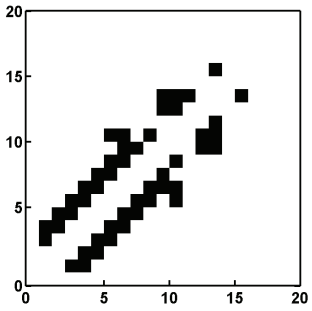


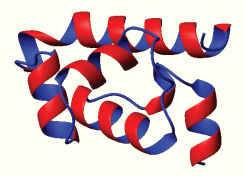
1FME



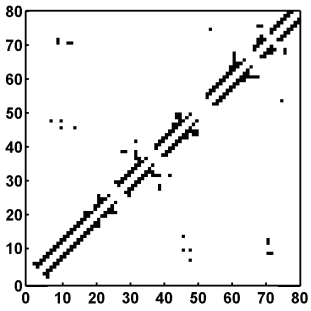


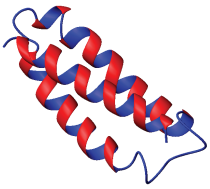
1L2Y



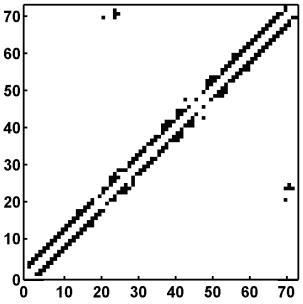


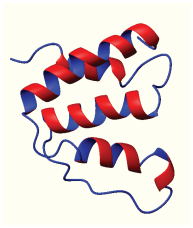
1LMB



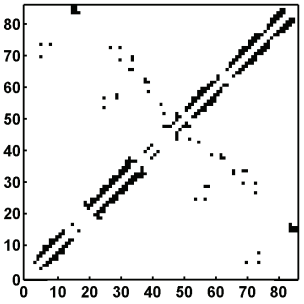


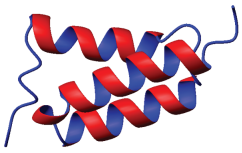
2A3D



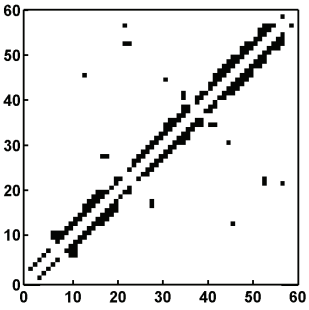


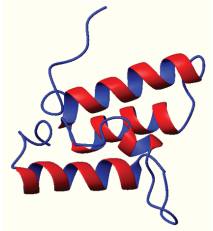
2ABD



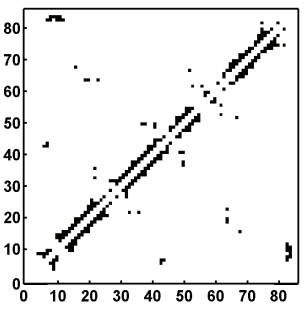


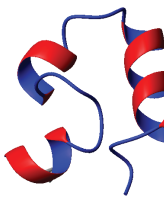
1SS1



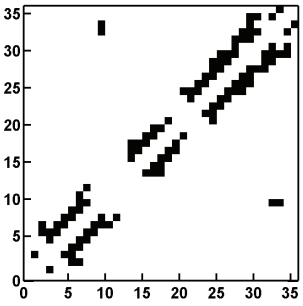


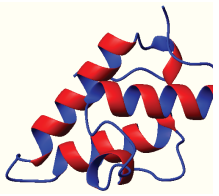
1IMQ



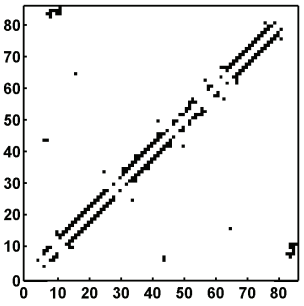


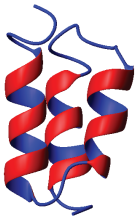
1VII



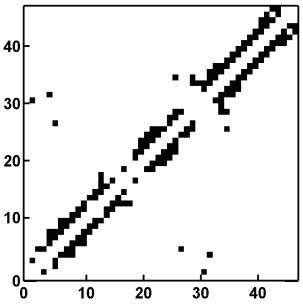


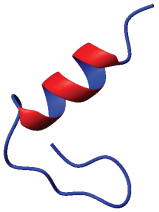
1AYI



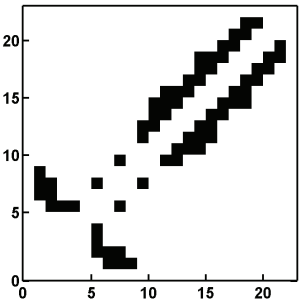


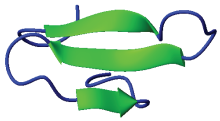
1PRB



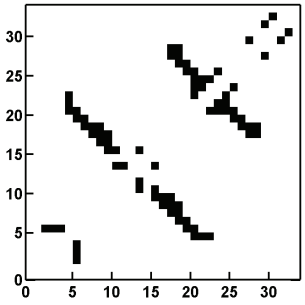


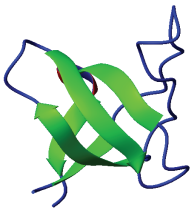
1T8J



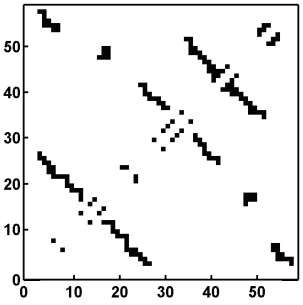


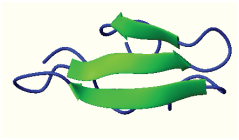
1PIN



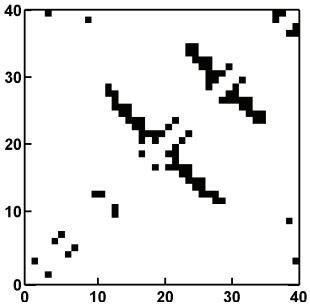


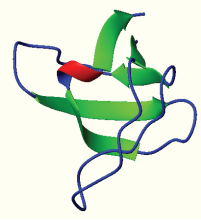
1SHF



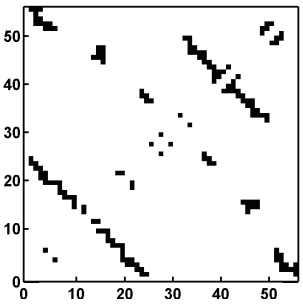


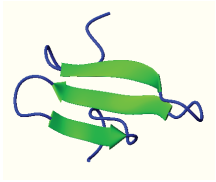
1K9Q



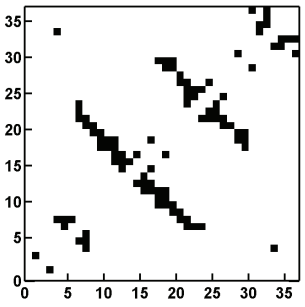


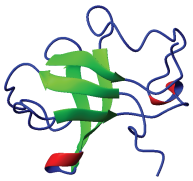
1FMK



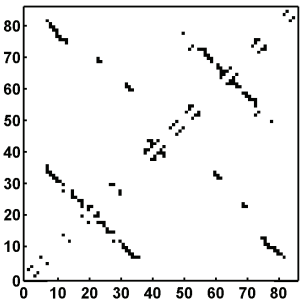


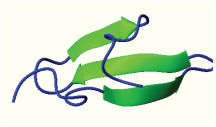
1E0M



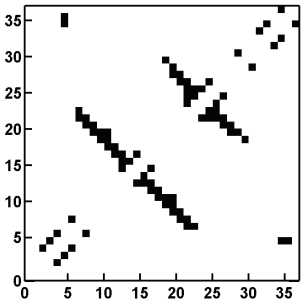


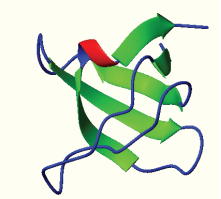
1PNJ



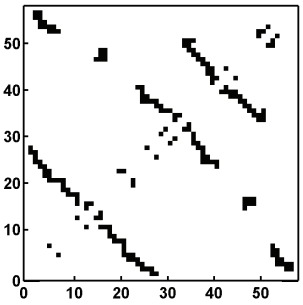


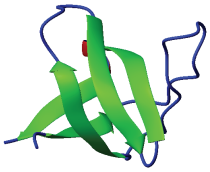
1E0L



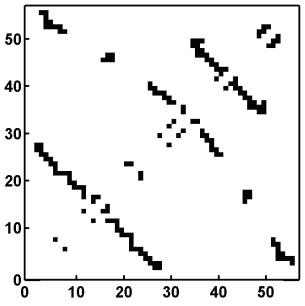


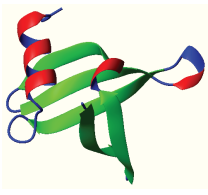
1JO8



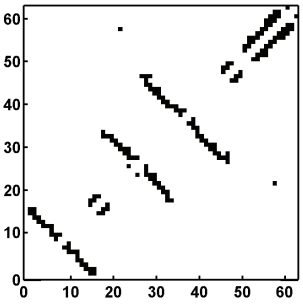


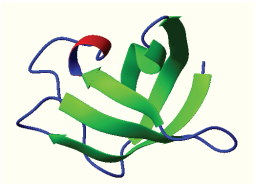
1SHG



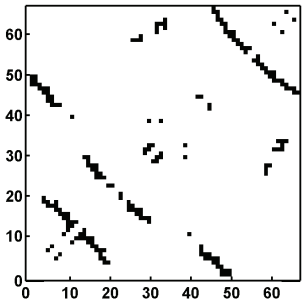


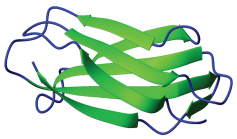
1BF4



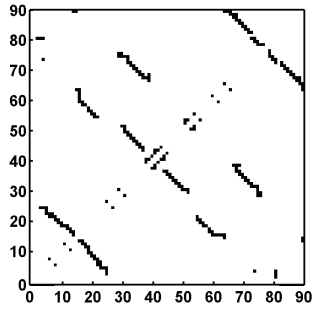


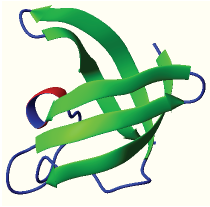
1CSP



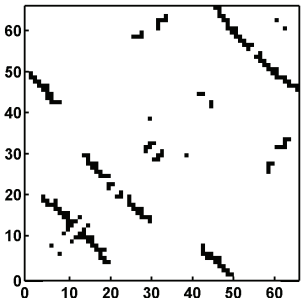


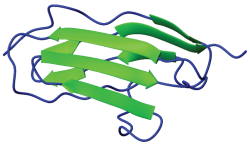
1TEN



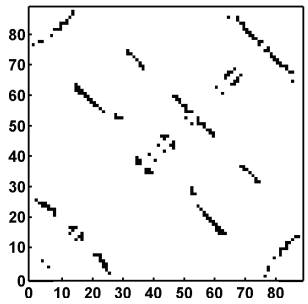


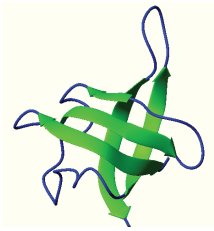
1C9O



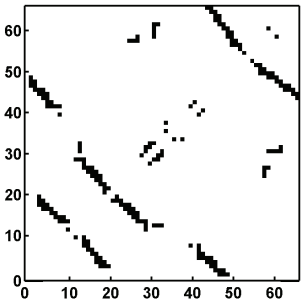


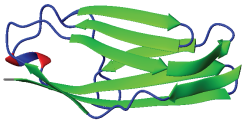
1TIT



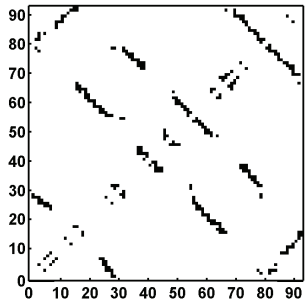


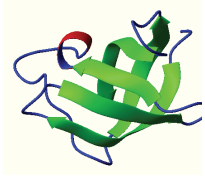
1G6P



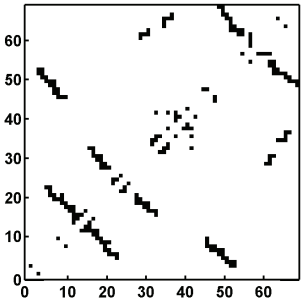


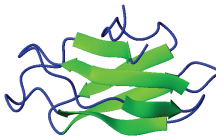
1WIT



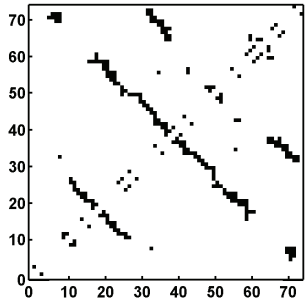


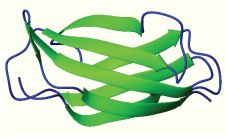
1MJC



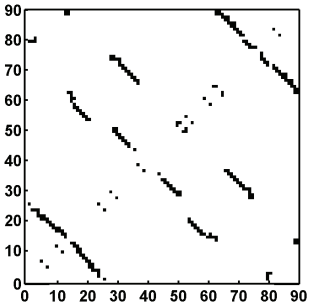


2AIT

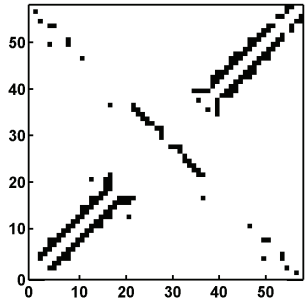




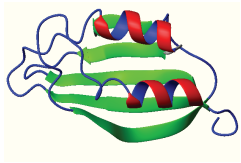
1FNF

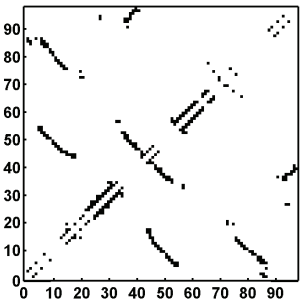


1HYW

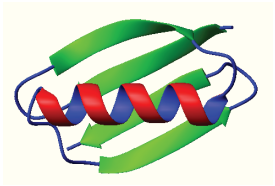


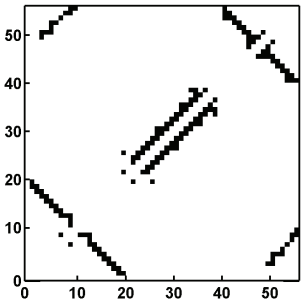
1APS



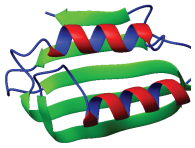


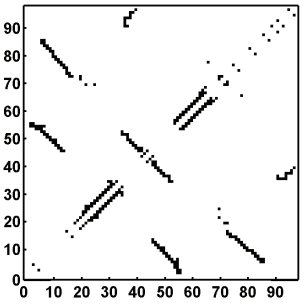
1PGB



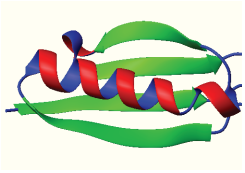


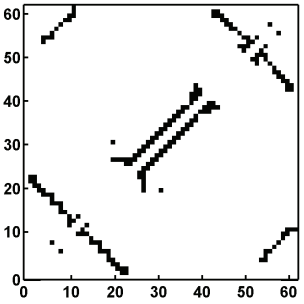
2ACY



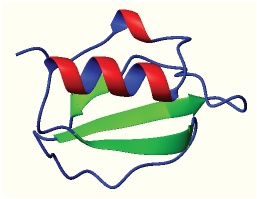


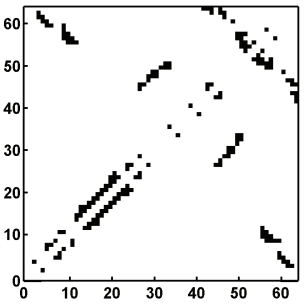
1HZ6



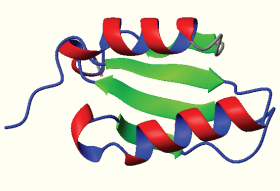


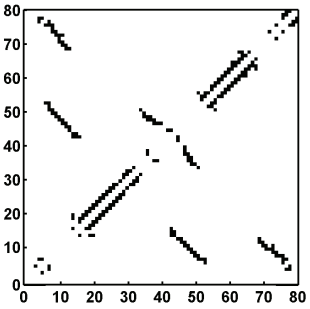
1COA



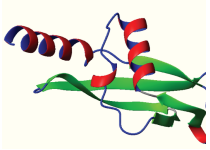


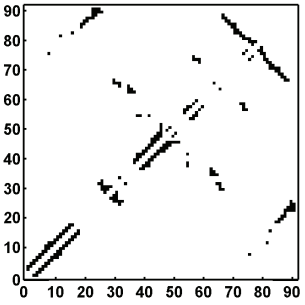
1AYE



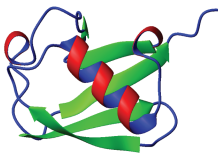


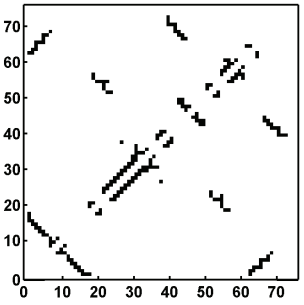
1DIVc



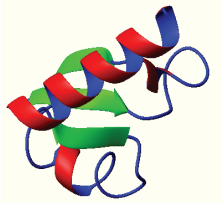


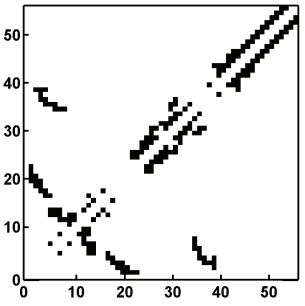
1UBQ



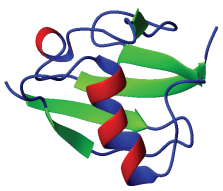


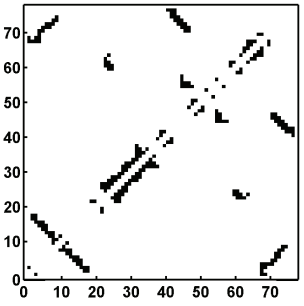
1DIVn

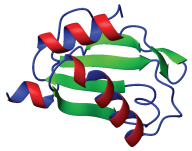




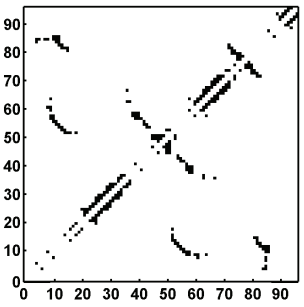
1RFA

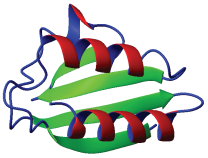




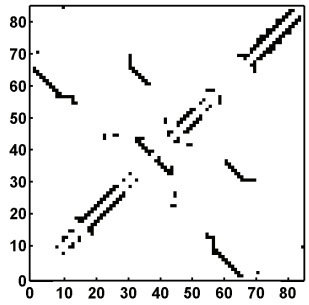


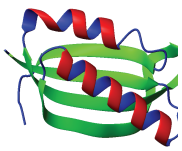
1URN



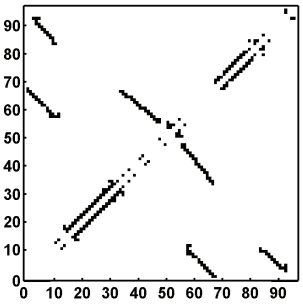


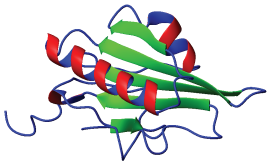
1POH



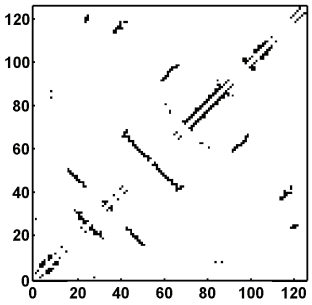


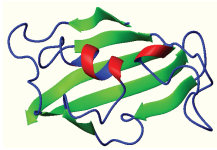
1RIS





2VIK





1FKB

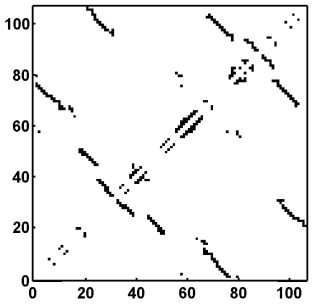


Figure A2 Distribution of ϕ-ψ dihedral angles for each individual amino acid

The ϕ-ψ space is represented as a 40 X 40 matrix with each square corresponding to a region of 9^o X 9^o. The color bar indicates the value of logarithm of number of hits in each region (Notice the different scale for each amino acid).

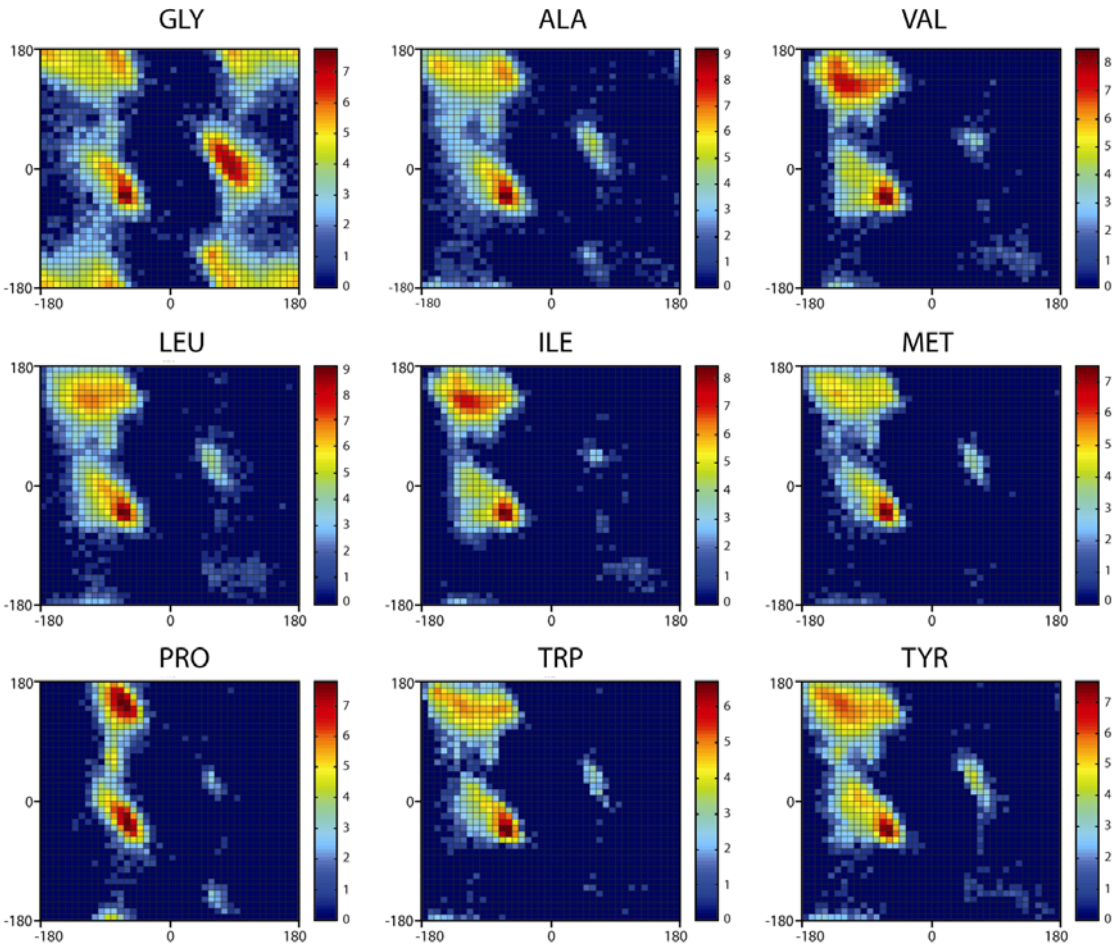


Figure A2 Distribution of ϕ-ψ dihedral angles for each individual amino acid (continued)

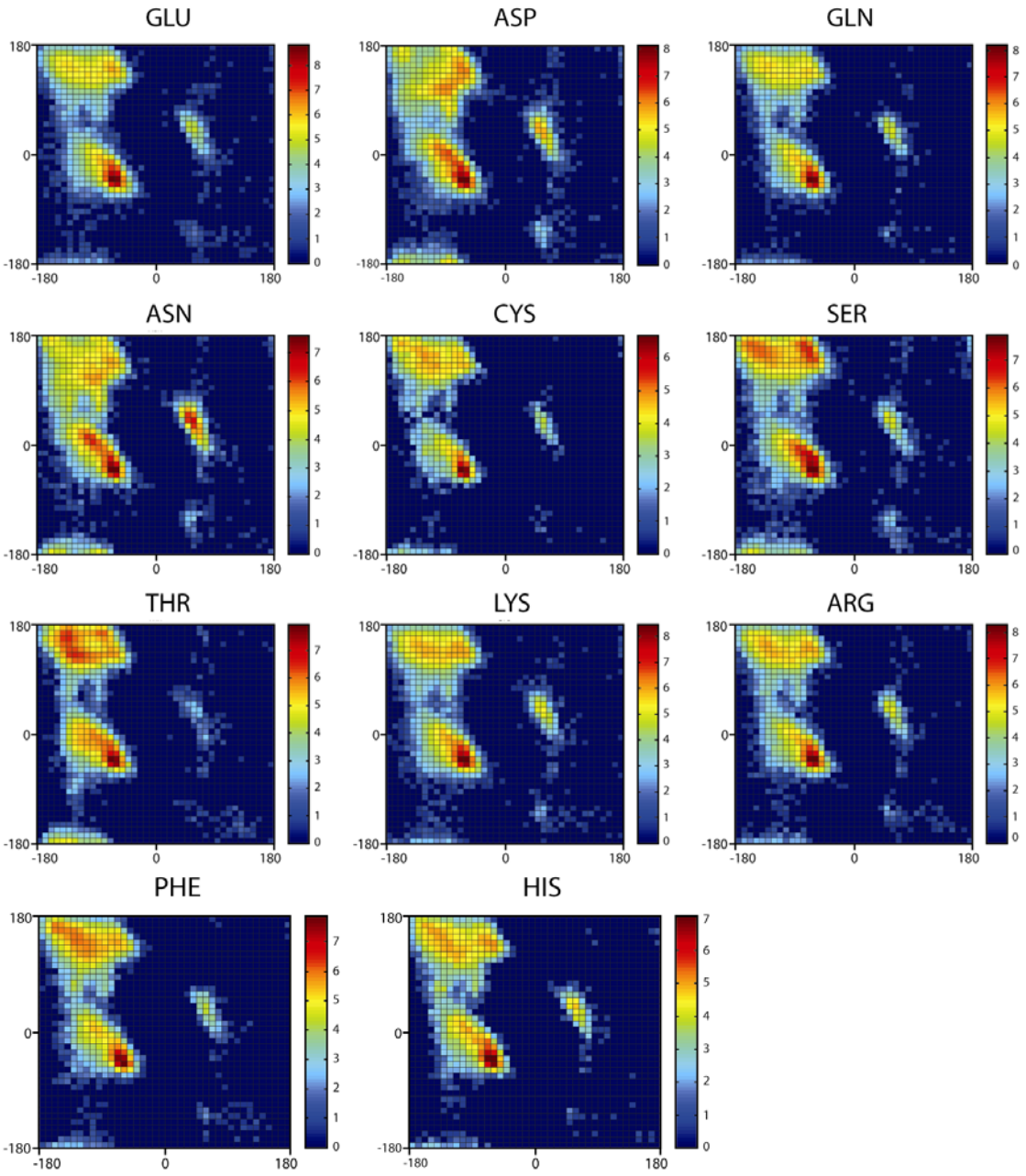


Figure A3 Distribution of side chain and peptide bond dihedral angles for each amino acid

The distribution of χ_1 is shown as blue, χ_2 as red, χ_3 as green and χ_4 as dark green lines. The dotted lines correspond to χ_{12}, χ_{22} and χ_{32} with the same color as their respective counterparts. ω dihedrals around the peptide bond are shown in black.

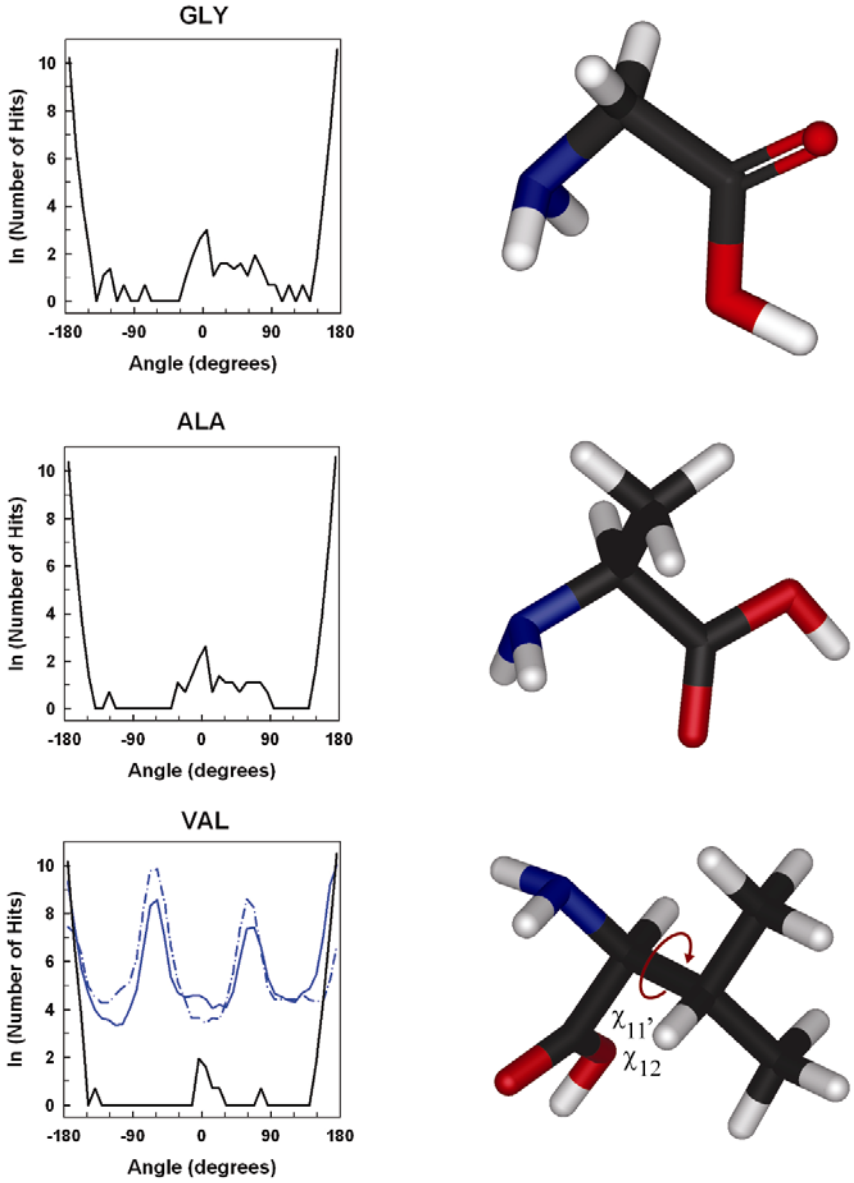


Figure A3 Distribution of side chain and peptide bond dihedral angles for each amino acid (continued)

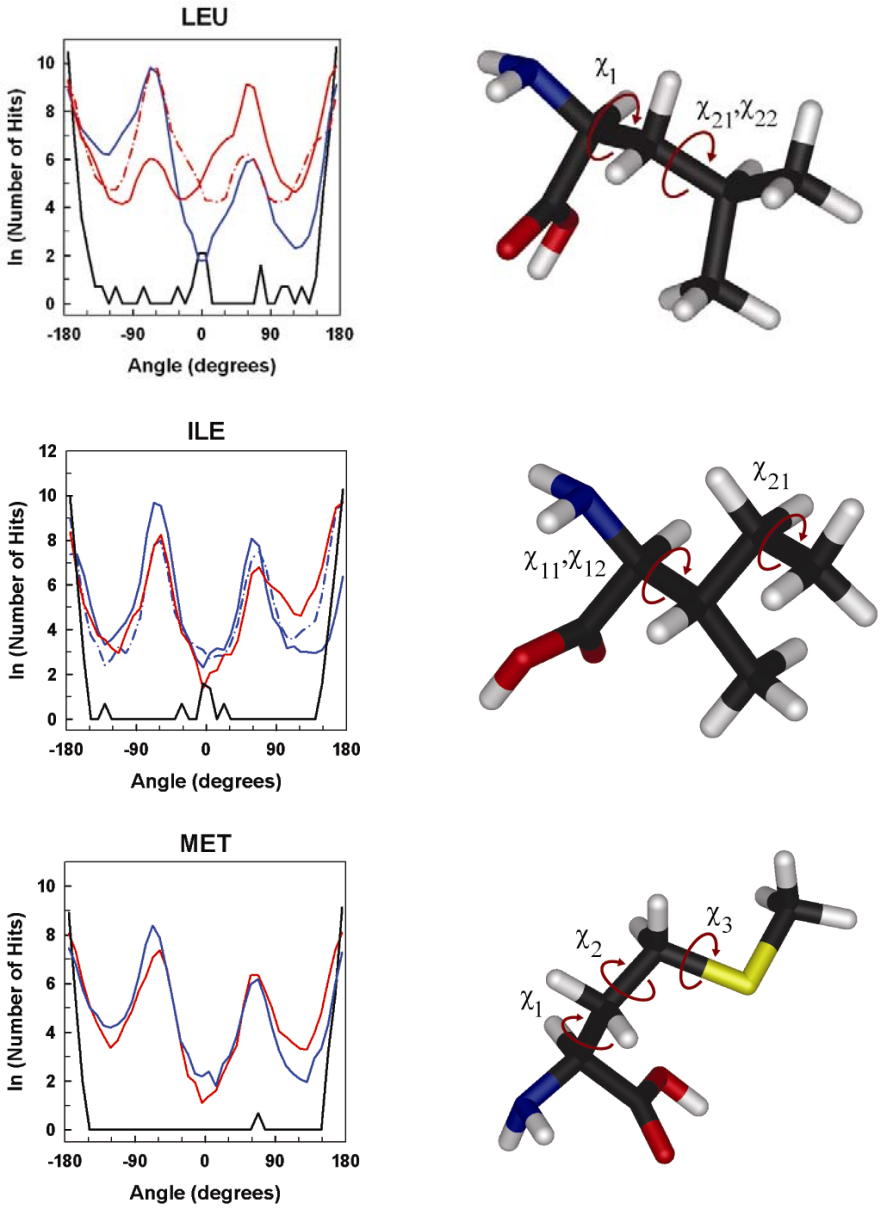
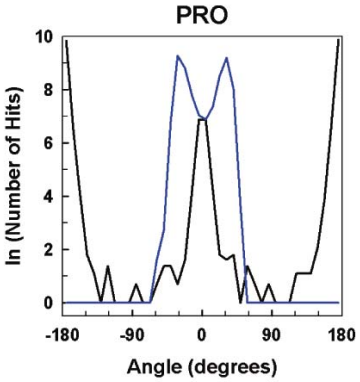
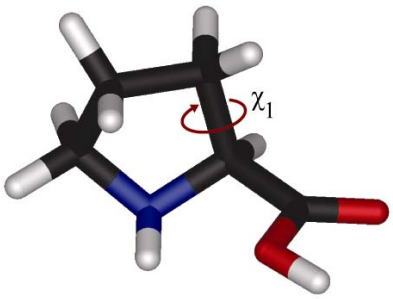
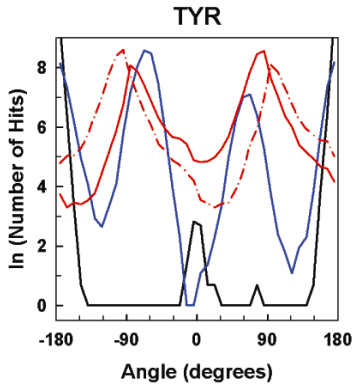
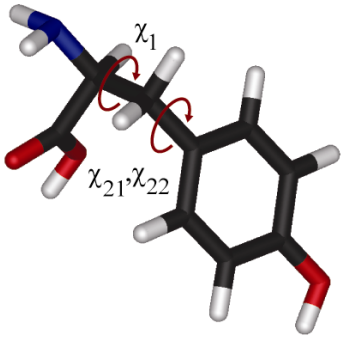


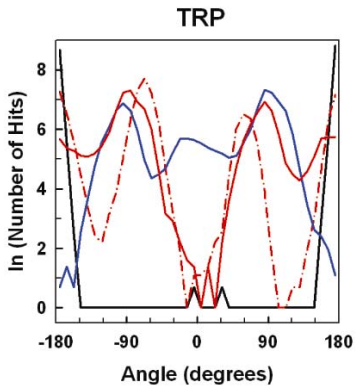
Figure A3 Distribution of side chain and peptide bond dihedral angles for each amino acid (continued)











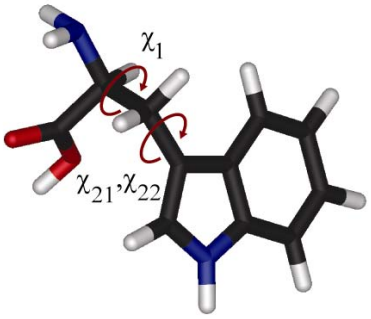
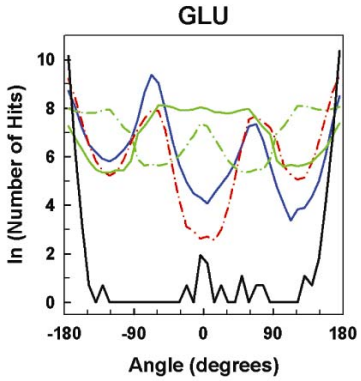
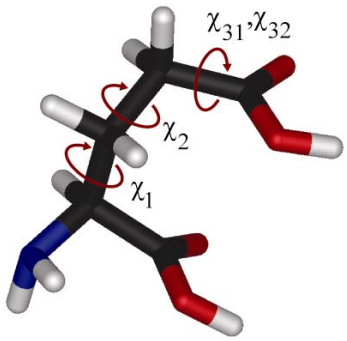
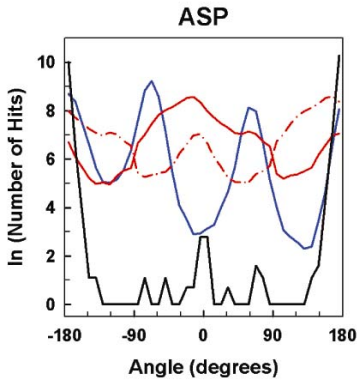
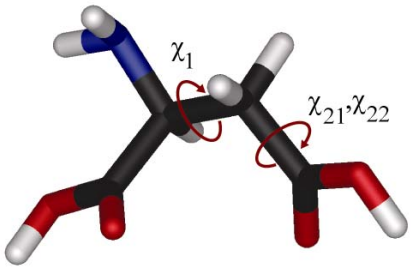


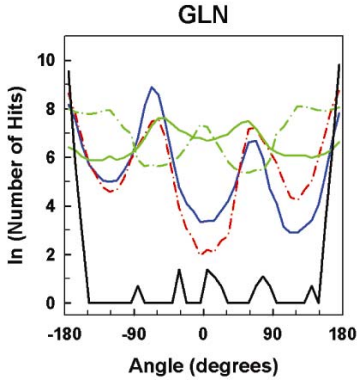
Figure A3 Distribution of side chain and peptide bond dihedral angles for each amino acid (continued)











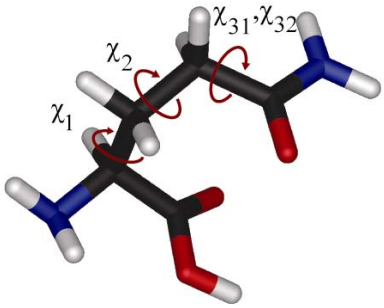
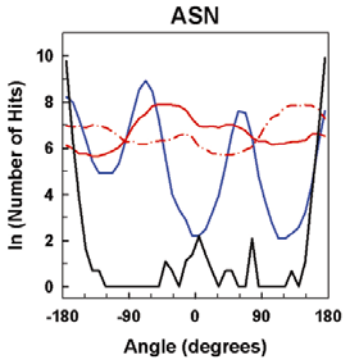
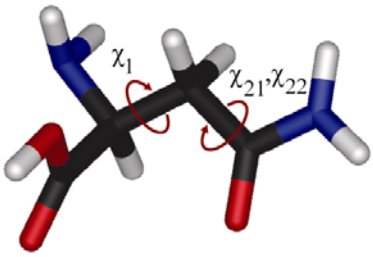
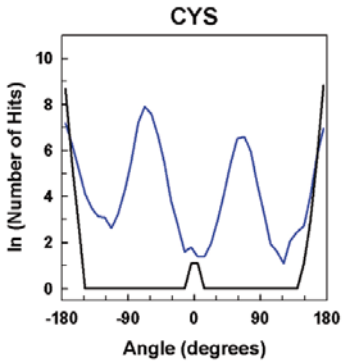
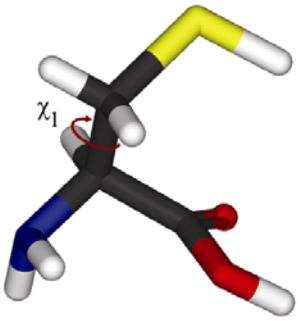


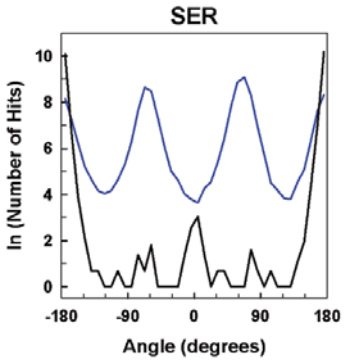
Figure A3 Distribution of side chain and peptide bond dihedral angles for each amino acid (continued)











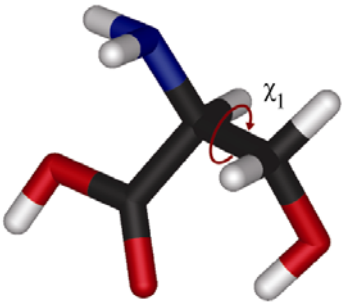
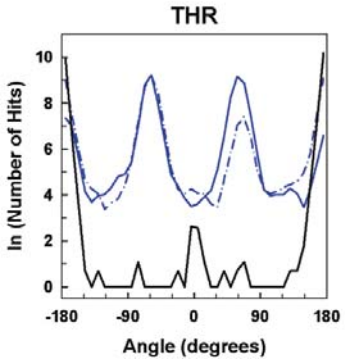
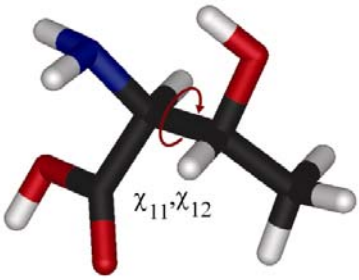
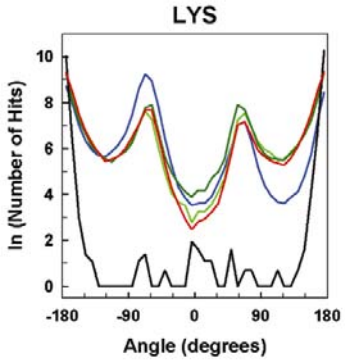
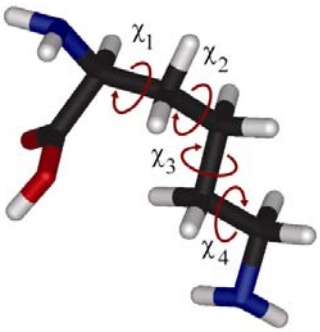


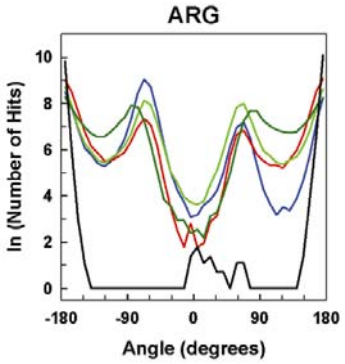
Figure A3 Distribution of side chain and peptide bond dihedral angles for each amino acid (continued)











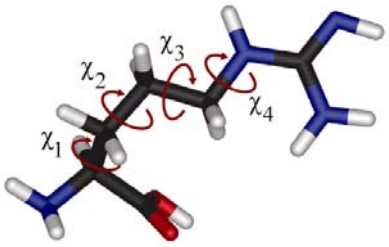
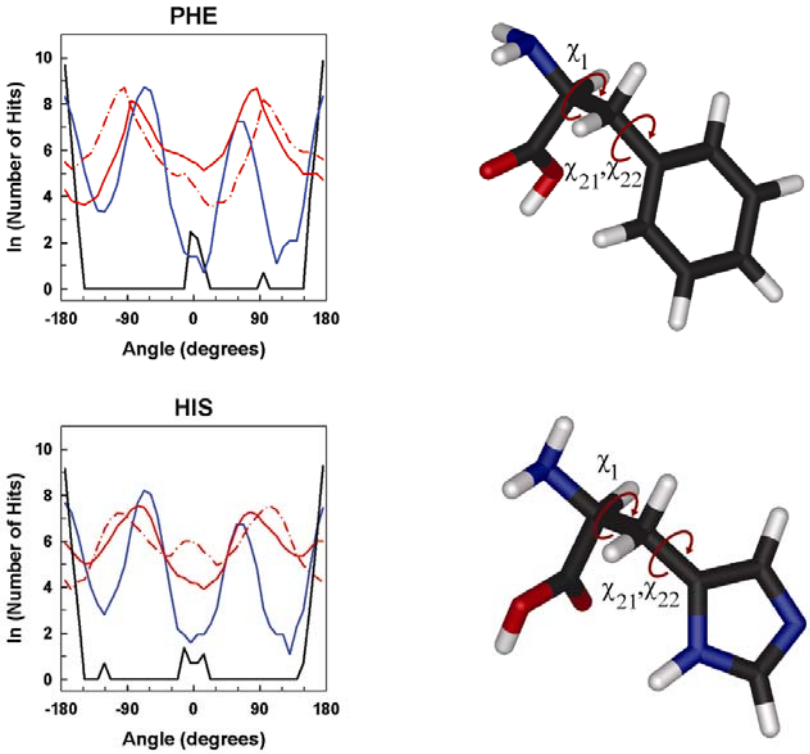
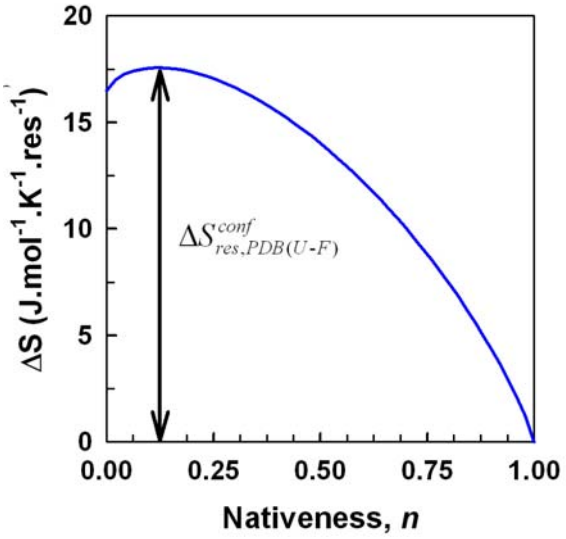


Figure A3 Distribution of side chain and peptide bond dihedral angles for each amino acid (continued)



Derivation of $\Delta S^{conf}_{res,PDB(U-F)}$ from $\Delta S^{conf}_{res,PDB}$



The blue curve in the above Figure is the entropy functional (Equation 4.3) given by

$$f = -R\left[n\ln(n) + (1-n)\ln(1-n)\right] + n\Delta S^{n=1}_{res} + (1-n)\Delta S^{n=0}_{res}$$

Since ΔS at n=1 is the reference state, $\Delta S^{n=1}_{res}$=0 and the functional becomes

$$\Delta S^{conf}_{res}(n) = -R\left[n\ln(n) + (1-n)\ln(1-n)\right] + (1-n)\Delta S^{n=0}_{res}$$

Here, $\Delta S^{n=0}_{res}$ corresponds to $\Delta S^{conf}_{res,PDB}$ and the maximum in the functional $\Delta S^{conf}_{res,PDB(U-F)}$ can be obtained from the first order derivative of the functional as follows

$$\frac{\partial f}{\partial n} = -R\left[\frac{\partial}{\partial n}\left(n\ln n\right) + \frac{\partial}{\partial n}\left((1-n)\ln(1-n)\right)\right] + \Delta S^{n=0}_{res}\frac{\partial}{\partial n}(1-n)$$

$$\frac{\partial f}{\partial n} = -R\left[\frac{\partial n}{\partial n}\left(\ln n\right) + n\frac{\partial}{\partial n}\left(\ln n\right) + \left(\frac{\partial}{\partial n}(1-n)\right)\ln(1-n) + (1-n)\frac{\partial}{\partial n}\left(\ln(1-n)\right)\right] + \Delta S^{n=0}_{res}\frac{\partial}{\partial n}(1-n)$$

$$\frac{\partial f}{\partial n} = -R\left[\ln n + 1 - \ln(1-n) - 1\right] - \Delta S^{n=0}_{res}$$

$$\frac{\partial f}{\partial n} = -R\left[\ln\left(\frac{n}{1-n}\right)\right] - \Delta S^{n=0}_{res}$$

At maximum point, $\frac{\partial f}{\partial n} = 0$, i.e. $-R\left[\ln(\frac{n}{1-n})\right] - \Delta S_{res}^{n=0} = 0$

$$\frac{n}{1-n} = \exp(-\frac{\Delta S_{res}^{n=0}}{R}) \quad \text{and} \quad n_{\max} = \frac{\exp(-\frac{\Delta S_{res}^{n=0}}{R})}{1+\exp(-\frac{\Delta S_{res}^{n=0}}{R})}$$

Let $s = \exp(-\frac{\Delta S_{res}^{n=0}}{R})$; $\Delta S_{res,PDB(U-F)}^{conf}$ is then given by

$$f(n_{\max}) = -R\left[\frac{s}{1+s}\ln\frac{s}{1+s} + \frac{1}{1+s}\ln\frac{1}{1+s}\right] + \Delta S_{res}^{n=0}\frac{1}{1+s}$$

$$f(n_{\max}) = -R\left[\frac{s}{1+s}\ln\frac{s}{1+s} + \frac{1}{1+s}\ln\frac{1}{1+s}\right] + \Delta S_{res}^{n=0}\frac{1}{1+s}$$

$$f(n_{\max}) = -R\left[\frac{s}{1+s}\bullet(-\frac{\Delta S_{res}^{n=0}}{r}) + \frac{s}{1+s}\ln\frac{1}{1+s} + \frac{1}{1+s}\ln\frac{1}{1+s}\right] + \Delta S_{res}^{n=0}\frac{1}{1+s}$$

$$f(n_{\max}) = -R\left[\frac{s}{1+s}\bullet(-\frac{\Delta S_{res}^{n=0}}{r}) + \ln\frac{1}{1+s}\right] + \Delta S_{res}^{n=0}\frac{1}{1+s)}$$

$$f(n_{\max}) = \Delta S_{res}^{n=0}\frac{s}{1+s} + R\ln(1+s) + \Delta S_{res}^{n=0}\frac{1}{1+s}$$

$f(n_{\max}) = R\ln(1+s) + \Delta S_{res}^{n=0}$ and therefore,

$$\Delta S_{res(U-F)}^{conf} = R\ln\left(1+\exp(-\frac{\Delta S_{res}^{n=0}}{R})\right) + \Delta S_{res}^{n=0} \quad \text{i.e.}$$

$$\Delta S_{res,PDB(U-F)}^{conf} = R\ln\left(1+\exp(-\frac{\Delta S_{res,PDB}^{conf}}{R})\right) + \Delta S_{res,PDB}^{conf}$$

$$\Delta S_{res,PDB(U-F)}^{conf} = R\ln\left(1+\frac{1}{\exp(\frac{\Delta S_{res,PDB}^{conf}}{R})}\right) + \Delta S_{res,PDB}^{conf}$$

$$\Delta S_{res,PDB(U-F)}^{conf} = R\ln\left(1+\frac{1}{\exp(\frac{\Delta S_{res,PDB}^{conf}}{R})}\right) + \Delta S_{res,PDB}^{conf}$$

$$\Delta S^{conf}_{res,PDB(U-F)} = R\ln\left(1+\frac{1}{\exp(\frac{\Delta S^{conf}_{res,PDB}}{R})}\right)+\Delta S^{conf}_{res,PDB}$$

$$\Delta S^{conf}_{res,PDB(U-F)} = R\ln\left(\frac{\exp(\frac{\Delta S^{conf}_{res,PDB}}{R})+1}{\exp(\frac{\Delta S^{conf}_{res,PDB}}{R})}\right)+\Delta S^{conf}_{res,PDB}$$

$$\Delta S^{conf}_{res,PDB(U-F)} = R\left(\ln\left(\exp(\frac{\Delta S^{conf}_{res,PDB}}{R})+1\right)-\ln\left(\exp(\frac{\Delta S^{conf}_{res,PDB}}{R})\right)\right)+\Delta S^{conf}_{res,PDB}$$

$$\Delta S^{conf}_{res,PDB(U-F)} = R\left(\ln\left(\exp(\frac{\Delta S^{conf}_{res,PDB}}{R})+1\right)-\frac{\Delta S^{conf}_{res,PDB}}{R}\right)+\Delta S^{conf}_{res,PDB}$$

$$\Delta S^{conf}_{res,PDB(U-F)} = R\left(\ln\left(\exp(\frac{\Delta S^{conf}_{res,PDB}}{R})+1\right)\right)$$ (i.e. Equation 5.9)

References

1. Anfinsen, C. B. (1973). Principles that govern folding of protein chains. *Science* **181**(4096), 223-230.
2. Levinthal, C. (1969). How to fold graciously. *Mossbaun Spectroscopy in Biol. Sys. Proc., University of Illinois Bulletin* **67**(41), 22-24.
3. Bryngelson, J. D., Onuchic, J. N., Socci, N. D.*, et al.* (1995). Funnels, pathways, and the energy landscape of protein-folding - a synthesis. *Proteins: Struct., Funct., Genet.* **21**(3), 167-195.
4. Pande, V. S., Grosberg, A. Y., Tanaka, T.*, et al.* (1998). Pathways for protein folding: Is a new view needed? *Curr. Opin. Struct. Biol.* **8**(1), 68-79.
5. Wolynes, P. G. (2001). Landscapes, funnels, glasses, and folding. *Proc. Amer. Phil. Soc.* **145**(4), 555-563.
6. Oliveberg, M. & Wolynes, P. G. (2005). The experimental survey of protein-folding energy landscapes. *Quart. Rev. Biophys.* **38**(3), 245-288.
7. Muñoz, V. (2002). Thermodynamics and kinetics of downhill protein folding investigated with a simple statistical mechanical model. *Int. J. Quantum Chem.* **90**(4-5), 1522-1528.
8. Wolynes, P. G. (1997). Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc. Natl. Acad. Sci. USA* **94**(12), 6170-6175.
9. Socci, N. D., Onuchic, J. N. & Wolynes, P. G. (1996). Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**(15), 5860-5868.
10. Lapidus, L. J., Steinbach, P. J., Eaton, W. A.*, et al.* (2002). Effects of chain stiffness on the dynamics of loop formation in polypeptides. Appendix: Testing a 1-dimensional diffusion model for peptide dynamics. *J. Phys. Chem. B* **106**(44), 11628-11640.
11. Muñoz, V. & Eaton, W. A. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA* **96**(20), 11311-11316.
12. Azuma, T., Hamaguch.K & Migita, S. (1972). Denaturation of Bence-Jones proteins by guanidine hydrochloride. *J. Biochem.* **72**(6), 1457-1467.
13. Griko, Y. V., Freire, E., Privalov, G.*, et al.* (1995). The unfolding thermodynamics of C-type lysozymes - A calorimetric study of the heat denaturation of equine lysozyme. *J. Mol. Biol.* **252**(4), 447-459.
14. Krantz, B. A., Mayne, L., Rumbley, J.*, et al.* (2002). Fast and slow intermediate accumulation and the initial barrier mechanism in protein folding. *J. Mol. Biol.* **324**(2), 359-371.
15. Jackson, S. E. (1998). How do small single-domain proteins fold? *Folding Des.* **3**(4), R81-R91.

16. Matouschek, A., Kellis, J. T., Serrano, L.*, et al.* (1989). Mapping the transition-state and pathway of protein folding by protein engineering. *Nature* **340**(6229), 122-126.
17. Gorski, S. A., Capaldi, A. P., Kleanthous, C.*, et al.* (2001). Acidic conditions stabilise intermediates populated during the folding of Im7 and Im9. *J. Mol. Biol.* **312**(4), 849-863.
18. Went, H. M., Benitez-Cardoza, C.G. & Jackson, S.E. (2004). Is an intermediate state populated on the folding pathway of Ubiquitin? *Febs Letters* **567**(2-3), 333-338.
19. Bieri, O., Wildegger, G., Bachmann, A.*, et al.* (1999). A salt-induced kinetic intermediate is on a new parallel pathway of lysozyme folding. *Biochemistry* **38**(38), 12460-12470.
20. Kiefhaber, T. (1995). Kinetic traps in lysozyme folding. *Proc. Natl. Acad. Sci. USA* **92**(20), 9029-9033.
21. Silow, M. & Oliveberg, M. (1997). Transient aggregates in protein folding are easily mistaken for folding intermediates. *Proc. Natl. Acad. Sci. USA* **94**(12), 6084-6086.
22. Krantz, B. A. & Sosnick, T. R. (2000). Distinguishing between two-state and three-state models for ubiquitin folding. *Biochemistry* **39**(38), 11696-11701.
23. Rios, M. A. D. & Plaxco, K. W. (2005). Apparent debye-huckel electrostatic effects in the folding of a simple, single domain protein. *Biochemistry* **44**(4), 1243-1250.
24. Park, S. H., Shastry, M. C. R. & Roder, H. (1999). Folding dynamics of the B1 domain of protein C explored by ultrarapid mixing. *Nature Struc. Biol.* **6**(10), 943-947.
25. Otzen, D. E., Kristensen, O., Proctor, M.*, et al.* (1999). Structural changes in the transition state of protein folding: Alternative interpretations of curved chevron plots. *Biochemistry* **38**(20), 6499-6511.
26. Kaya, H. & Chan, H. S. (2003). Origins of chevron rollovers in non-two-state protein folding kinetics. *Phys. Rev. Lett.* **90**(25).
27. Gruebele, M. (2005). Downhill protein folding: Evolution meets physics. *Comp. Rend. Biol.* **328**(8), 701-712.
28. Garcia-Mira, M. M., Sadqi, M., Fischer, N.*, et al.* (2002). Experimental identification of downhill protein folding. *Science* **298**(5601), 2191-2195.
29. Naganathan, A. N., Perez-Jimenez, R., Sanchez-Ruiz, J. M.*, et al.* (2005). Robustness of downhill folding: Guidelines for the analysis of equilibrium folding experiments on small proteins. *Biochemistry* **44**(20), 7435-7449.
30. Oliva, F. Y. & Muñoz, V. (2004). A simple thermodynamic test to discriminate between two'state and downhill folding. *J. Am. Chem. Soc.* **126**, 8596-8597.
31. Gomez, J., Hilser, V. J., Xie, D.*, et al.* (1995). The heat-capacity of proteins. *Proteins: Struct., Funct., Genet.* **22**(4), 404-412.
32. Ma, H. R. & Gruebele, M. (2005). Kinetics are probe-dependent during downhill folding of an engineered $\lambda_{(6\text{-}85)}$ protein. *Proc. Natl. Acad. Sci. USA* **102**(7), 2283-2287.

33. Sadqi, M., Fushman, D. & Muñoz, V. (2006). Atom-by-atom analysis of global downhill protein folding. *Nature* **442**(7100), 317-321.
34. Hagen, S. J. (2003). Exponential decay kinetics in "downhill" protein folding. *Proteins: Struct., Funct., Genet.* **50**(1), 1-4.
35. Nymeyer, H., Garcia, A. E. & Onuchic, J. N. (1998). Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci. USA* **95**(11), 5921-5928.
36. Onuchic, J. N., LutheySchulten, Z. & Wolynes, P. G. (1997). Theory of protein folding: The energy landscape perspective. *Ann. Rev. Phys. Chem.* **48**, 545-600.
37. Onuchic, J. N., Wolynes, P. G., Lutheyschulten, Z*., et al.* (1995). Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl. Acad. Sci. USA* **92**(8), 3626-3630.
38. Yang, W. Y. & Gruebele, M. (2004). Folding λ-repressor at its speed limit. *Biophys. J.* **87**, 596-608.
39. Kaya, H. & Chan, H. S. (2002). Towards a consistent modeling of protein thermodynamic and kinetic co-operativity: How applicable is the transition state picture to folding and unfolding? *J. Mol. Biol.* **315**(4), 899-909.
40. Knott, M. & Chan, H. S. (2006). Criteria for downhill protein folding: Calorimetry, chevron plot, kinetic relaxation, and single-molecule radius of gyration in chain models with subdued degrees of cooperativity. *Proteins: Struct., Func., Bioinf.* **65**(2), 373-391.
41. Robertson, A. D. & Murphy, K. P. (1997). Protein structure and the energetics of protein stability. *Chem. Rev.* **97**(5), 1251-1267.
42. Finkelstein, A. V. & Badretdinov, A. Y. (1997). Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Folding Des.* **2**(2), 115-121.
43. Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. (1996). Chain length scaling of protein folding time. *Phys. Rev. Lett.* **77**(27), 5433-5436.
44. Thirumalai, D. (1995). From minimal models to real proteins - time scales for protein-folding kinetics. *J. Phys. I* **5**(11), 1457-1467.
45. Li, M. S., Klimov, D. K. & Thirumalai, D. (2004). Thermal denaturation and folding rates of single domain proteins: Size matters. *Polymer* **45**(2), 573-579.
46. Galzitskaya, O. V., Garbuzynskiy, S. O., Ivankov, D. N*., et al.* (2003). Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins: Struct., Funct., Genet.* **51**(2), 162-166.
47. Naganathan, A. N. & Muñoz, V. (2005). Scaling of folding times with protein size. *J. Am. Chem. Soc.* **127**, 480-481.
48. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**(4), 985-994.
49. Gromiha, M. M. & Selvaraj, S. (2001). Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction. *J. Mol. Biol.* **310**(1), 27-32.

50. Zhou, H. Y. & Zhou, Y. Q. (2002). Folding rate prediction using total contact distance. *Biophys. J.* **82**(1), 458-463.
51. Ivankov, D. N. & Finkelstein, A. V. (2004). Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl. Acad. Sci. USA* **101**(24), 8942-8944.
52. Zwanzig, R. (1995). A simple model of protein folding kinetics. *Proc. Natl. Acad. Sci. USA* **92**, 9801-9804.
53. Galzitskaya, O. V. & Finkelstein, A. V. (1999). A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. USA* **96**(20), 11299-11304.
54. Finkelstein, A. V. & Badretdinov, A. Y. (1998). Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold - Influence of chain knotting on the rate of folding (vol 2, pg 115, 1997). *Folding & Design* **3**(1), 67-68.
55. Alm, E. & Baker, D. (1999). Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA* **96**(20), 11305-11310.
56. Alm, E., Morozov, A. V., Kortemme, T.*, et al.* (2002). Simple physical models connect theory and experiment in protein folding kinetics. *J. Mol. Biol.* **322**(2), 463-476.
57. Ivankov, D. N. & Finkelstein, A. V. (2001). Theoretical study of a landscape of protein folding-unfolding pathways. Folding rates at mid-transition. *Biochemistry* **40**(33), 9957-9961.
58. Henry, E. R. & Eaton, W. A. (2004). Combinatorial modeling of protein folding kinetics: Free energy profiles and rates. *Chem. Phys.* **307**(2-3), 163-185.
59. Karplus, M. (2000). Aspects of protein reaction dynamics: Deviations from simple behavior. *J. Phys. Chem. B* **104**(1), 11-27.
60. Kramer, H. A. (1940). *Physica* **7**, 284-307.
61. Hanggi, P., Talkner, P. & Borkovec, M. (1990). Reaction-rate theory - 50 years after Kramers. *Rev. Mod. Phys.* **62**(2), 251-341.
62. Akmal, A. & Muñoz, V. (2004). The nature of the free energy barriers to two-state folding. *Proteins: Struct., Funct., Bioinf.* **57**(1), 142-152.
63. Naganathan, A. N., Sanchez-Ruiz, J. M. & Muñoz, V. (2005). Direct measurement of barrier heights in protein folding. *J. Am. Chem. Soc.* **127**(51), 17970-17971.
64. Muñoz, V. & Sanchez-Ruiz, J. M. (2004). Exploring protein-folding ensembles: A variable-barrier model for the analysis of equilibrium unfolding experiments. *Proc. Natl. Acad. Sci. USA* **101**, 17646-17651.
65. Hofrichter, J., Thompson, P. A., Muñoz, V.*, et al.* (1998). Folding dynamics of an α-helix and a β-hairpin. *Biophys. J.* **74**(2), A3-A3.
66. Huang, C. Y., Klemke, J. W., Getahun, Z.*, et al.* (2001). Temperature-dependent helix-coil transition of an alanine based peptide. *J. Am. Chem. Soc.* **123**(38), 9235-9238.

67. Williams, S., Causgrove, T. P., Gilmanshin, R.*, et al.* (1996). Fast events in protein folding: Helix melting and formation in a small peptide. *Biochemistry* **35**(3), 691-697.
68. Lednev, I. K., Karnoup, A. S., Sparrow, M. C.*, et al.* (1999). Nanosecond UV resonance Raman examination of initial steps in α-helix secondary structure evolution. *J. Am. Chem. Soc.* **121**(16), 4076-4077.
69. Muñoz, V., Ghirlando, R., Blanco, F. J.*, et al.* (2006). Folding and aggregation kinetics of a β-hairpin. *Biochemistry* **45**(23), 7023-7035.
70. Thompson, P. A., Muñoz, V., Jas, G. S.*, et al.* (2000). The helix-coil kinetics of a heteropeptide. *J. Phys. Chem. B* **104**(2), 378-389.
71. Muñoz, V., Thompson, P. A., Henry, E. R.*, et al.* (1998). Folding dynamics of a β-hairpin studied by laser temperature jump and kinetic modelling. *Biophys. J.* **74**(2), A175-A175.
72. Sadqi, M., Lapidus, L. J. & Muñoz, V. (2003). How fast is protein hydrophobic collapse? *Proc. Natl. Acad. Sci. USA* **100**(21), 12117-12122.
73. Hagen, S. J. & Eaton, W. A. (2000). Two-state expansion and collapse of a polypeptide. *J. Mol. Biol.* **301**(4), 1019-1027.
74. Lapidus, L. J., Eaton, W. A. & Hofrichter, J. (2000). Measuring the rate of intramolecular contact formation in polypeptides. *Proc. Natl. Acad. Sci. USA* **97**(13), 7220-7225.
75. Hagen, S. J., Hofrichter, J., Szabo, A.*, et al.* (1996). Diffusion-limited contact formation in unfolded cytochrome c: Estimating the maximum rate of protein folding. *Proc. Natl. Acad. Sci. USA* **93**(21), 11615-11617.
76. Yang, W. Y. & Gruebele, M. (2003). Folding at the speed limit. *Nature* **423**(6936), 193-197.
77. Kubelka, J., Hofrichter, J. & Eaton, W. A. (2004). The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.* **14**(1), 76-88.
78. Zimm, B. H. & Bragg, J. K. (1959). Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* **31**(2), 526-535.
79. Lifson, S. & Roig, A. (1961). On the theory of helix-coil transition in polypeptides. *J. Chem. Phys.* **34**(6), 1963-1974.
80. Scholtz, J. M. & Baldwin, R. L. (1992). The mechanism of α-helix formation by peptides. *Ann. Rev. Biophys. Biomol. Struct.* **21**, 95-118.
81. Muñoz, V. & Serrano, L. (1994). Elucidating the folding problem of helical peptides using empirical parameters. *Nature Struct. Biol.* **1**(6), 399-409.
82. Chakrabartty, A. & Baldwin, R. L. (1995). Stability of α-helices. *Adv. Prot. Chem., Vol 46* **46**, 141-176.
83. Scheraga, H. A. (1978). Use of random copolymers to determine helix-coil stability-constants of naturally occurring amino-acids. *Pure Appl. Chem.* **50**(4), 315-324.
84. Doig, A. J. (2006). The α-helix as the simplest protein model: Helix-coil theory, stability and design (Muñoz, V., ed.).
85. Muñoz, V. & Serrano, L. (1995). Elucidating the folding problem of helical peptides using empirical parameters .2. Helix macrodipole effects and rational modification of the helical content of natural peptides. *J. Mol. Biol.* **245**(3), 275-296.

86. Muñoz, V. & Serrano, L. (1995). Elucidating the folding problem of helical peptides using empirical parameters .3. Temperature and pH-dependence. *J. Mol. Biol.* **245**(3), 297-308.
87. Muñoz, V. & Serrano, L. (1997). Development of the multiple sequence approximation within the AGADIR model of α-helix formation: Comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers* **41**(5), 495-509.
88. Schwarz, G. (1965). On kinetics of helix-coil transition of polypeptides in solution. *J. Mol. Biol.* **11**(1), 64-&.
89. Gruenewald, B., Nicola, C. U., Lustig, A.*, et al.* (1979). Kinetics of the helix-coil transition of a polypeptide with nonionic side groups, derived from ultrasonic relaxation measurements. *Biophys. Chem.* **9**(2), 137-147.
90. Zana, R. (1975). Rate-determining step for helix propagation in helix-coil transition of polypeptides in solution. *Biopolymers* **14**(11), 2425-2428.
91. Eaton, W. A., Muñoz, V., Hagen, S. J.*, et al.* (2000). Fast kinetics and mechanisms in protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 327-359.
92. Thompson, P. A., Eaton, A. & Hofrichter, J. (1997). Laser temperature-jump studies of the helix-coil transition of an alanine-based peptide. *Biophys. J.* **72**(2), WP377-WP377.
93. Huang, C. Y., Getahun, Z., Zhu, Y. J.*, et al.* (2002). Helix formation via conformation diffusion search. *Proc. Natl. Acad. Sci. USA* **99**(5), 2788-2793.
94. Clarke, D. T., Doig, A. J., Stapley, B. J.*, et al.* (1999). The α-helix folds on the millisecond time scale. *Proc. Natl. Acad. Sci. USA* **96**(13), 7232-7237.
95. Thompson, P. A., Eaton, W. A. & Hofrichter, J. (1997). Laser temperature-jump study of the helix-coil kinetics of an alanine peptide interpreted with a 'kinetic zipper' model. *Biochemistry* **36**(30), 9200-9210.
96. Hummer, G., Garcia, A. E. & Garde, S. (2000). Conformational diffusion and helix formation kinetics. *Phys. Rev. Lett.* **85**(12), 2637-2640.
97. Hummer, G., Garcia, A. E. & Garde, S. (2001). Helix nucleation kinetics from molecular simulations in explicit solvent. *Proteins: Struct. Func. Genet.* **42**(1), 77-84.
98. Huang, C. Y., Getahun, Z., Wang, T.*, et al.* (2001). Time-resolved infrared study of the helix-coil transition using ^{13}C-labeled helical peptides. *J. Am. Chem. Soc.* **123**(48), 12111-12112.
99. Sorin, E. J. & Pande, V. S. (2005). Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.* **88**(4), 2472-2493.
100. Schneider, J. P. & DeGrado, W. F. (1998). The design of efficient α-helical C-capping auxiliaries. *J. Am. Chem. Soc.* **120**(12), 2764-2767.
101. Cheam, T. C. & Krimm, S. (1985). Vibrational analysis of peptides, polypeptides, and proteins .26. Infrared intensities of amide modes in N-methylacetamide and poly(glycine-i) from Ab-initio calculations of dipole-moment derivatives of N-methylacetamide. *J. Chem. Phys.* **82**(4), 1631-1641.
102. Krimm, S. & Bandekar, J. (1986). Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins. *Adv. Prot. Chem.* **38**, 181-364.

103. Muñoz, V., Henry, E. R., Hofrichter, J.*, et al.* (1998). A statistical mechanical model for β-hairpin kinetics. *Proc. Natl. Acad. Sci. USA* **95**(11), 5872-5879.
104. Strehlow, K. G., Robertson, A. D. & Baldwin, R. L. (1991). Proline for alanine substitutions in the C-peptide helix of Ribonuclease-A. *Biochemistry* **30**(23), 5810-5814.
105. Lednev, I. K., Karnoup, A. S., Sparrow, M. C.*, et al.* (1999). α-helix peptide folding and unfolding activation barriers: A nanosecond UV resonance Raman study. *J. Am. Chem. Soc.* **121**(35), 8074-8086.
106. Naganathan, A. N., Doshi, U., Fung, A.*, et al.* (2006). Dynamics, energetics, and structure in protein folding. *Biochemistry* **45**(28), 8466-8475.
107. Ivankov, D. N., Garbuzynskiy, S. O., Alm, E.*, et al.* (2003). Contact order revisited: Influence of protein size on the folding rate. *Prot. Sci.* **12**(9), 2057-2062.
108. Kortemme, T., Morozov, A. V. & Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* **326**(4), 1239-1259.
109. Berrera, M., Molinari, H. & Fogolari, F. (2003). Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics* **4:8**.
110. Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol.Biol.* **256**(3), 623-644.
111. Ferguson, N., Sharpe, T. D., Schartau, P. J.*, et al.* (2005). Ultra-fast barrier-limited folding in the peripheral subunit-binding domain family. *J. Mol. Biol.* **353**(2), 427-446.
112. Spector, S. & Raleigh, D. P. (1999). Submillisecond folding of the peripheral subunit-binding domain. *J. Mol. Biol.* **293**(4), 763-768.
113. Mayor, U., Johnson, C. M., Daggett, V.*, et al.* (2000). Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc. Natl. Acad. Sci. USA* **97**(25), 13518-13522.
114. Gianni, S., Guydosh, N. R., Khan, F.*, et al.* (2003). Unifying features in protein-folding mechanisms. *Proc. Natl. Acad. Sci. USA* **100**(23), 13286-13291.
115. Qiu, L. L., Pabit, S. A., Roitberg, A. E.*, et al.* (2002). Smaller and faster: The 20-residue Trp-cage protein folds in 4 μs. *J. Am. Chem. Soc.* **124**(44), 12952-12953.
116. Zhu, Y., Alonso, D. O. V., Maki, K.*, et al.* (2003). Ultrafast folding of α3D: A *de novo* designed three-helix bundle protein. *Proc. Natl. Acad. Sci. USA* **100**(26), 15486-15491.
117. Sato, S., Religa, T. L. & Fersht, A. R. (2006). Phi-analysis of the folding of the B domain of protein A using multiple optical probes. *J. Mol. Biol.* **360**(4), 850-864.
118. Kubelka, J., Eaton, W. A. & Hofrichter, J. (2003). Experimental tests of villin subdomain folding simulations. *J. Mol. Biol.* **329**(4), 625-630.

119. Burton, R. E., Huang, G. S., Daugherty, M. A., *et al.* (1996). Microsecond protein folding through a compact transition state. *J. Mol. Biol.* **263**(2), 311-322.
120. Maxwell, K. L., Wildes, D., Zarrine-Afsar, A., *et al.* (2005). Protein folding: Defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Prot. Sci.* **14**(3), 602-616.
121. Ferguson, N., Capaldi, A. P., James, R., *et al.* (1999). Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9. *J. Mol. Biol.* **286**(5), 1597-1608.
122. Jager, M., Nguyen, H., Crane, J. C., *et al.* (2001). The folding mechanism of a β-sheet: The WW domain. *J. Mol. Biol.* **311**(2), 373-393.
123. Ferguson, N., Johnson, C. M., Macias, M., *et al.* (2001). Ultrafast folding of WW domains without structured aromatic clusters in the denatured state. *Proc. Natl. Acad. Sci. USA* **98**(23), 13002-13007.
124. Viguera, A. R., Martinez, J. C., Filimonov, V. V., *et al.* (1994). Thermodynamic and kinetic-analysis of the SH3 domain of Spectrin shows a 2-state folding transition. *Biochemistry* **33**(8), 2142-2150.
125. Plaxco, K. W., Guijarro, J. I., Morton, C. J., *et al.* (1998). The folding kinetics and thermodynamics of the Fyn-SH3 domain. *Biochemistry* **37**(8), 2529-2537.
126. Grantcharova, V. P. & Baker, D. (1997). Folding dynamics of the src SH3 domain. *Biochemistry* **36**(50), 15685-15692.
127. Guijarro, J. I., Morton, C. J., Plaxco, K. W., *et al.* (1998). Folding kinetics of the SH3 domain of PI3 kinase by real-time NMR combined with optical spectroscopy. *J. Mol. Biol.* **276**(3), 657-667.
128. Guerois, R. & Serrano, L. (2000). The SH3-fold family: Experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304**(5), 967-982.
129. Garcia-Mira, M. M., Boehringer, D. & Schmid, F. X. (2004). The folding transition state of the cold shock protein is strongly polarized. *J. Mol. Biol.* **339**(3), 555-569.
130. Perl, D., Welker, C., Schindler, T., *et al.* (1998). Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nature Struct. Biol.* **5**(3), 229-235.
131. Reid, K. L., Rodriguez, H. M., Hillier, B. J., *et al.* (1998). Stability and folding properties of a model β-sheet protein, *Escherichia coli* CspA. *Prot. Sci.* **7**(2), 470-479.
132. Plaxco, K. W., Spitzfaden, C., Campbell, I. D., *et al.* (1997). A comparison of the folding kinetics and thermodynamics of two homologous Fibronectin type III modules. *J. Mol. Biol.* **270**(5), 763-770.
133. 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**(5), 771-778.
134. Clarke, J., Cota, E., Fowler, S. B., *et al.* (1999). Folding studies of immunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway. *Struct. Fold. & Des.* **7**(9), 1145-1153.

135. Schonbrunner, N., Koller, K. P. & Kiefhaber, T. (1997). Folding of the disulfide-bonded β-sheet protein tendamistat: Rapid two-state folding without hydrophobic collapse. *J. Mol. Biol.* **268**(2), 526-538.
136. van Nuland, N. A. J., Chiti, F., Taddei, N.*, et al.* (1998). Slow folding of muscle Acyl Phosphatase in the absence of intermediates. *J. Mol. Biol.* **283**(4), 883-891.
137. Taddei, N., Chiti, F., Paoli, P.*, et al.* (1999). Thermodynamics and kinetics of folding of common-type Acyl Phosphatase: Comparison to the highly homologous muscle isoenzyme. *Biochemistry* **38**(7), 2135-2142.
138. 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**(2), 260-288.
139. Jackson, S. E. & Fersht, A. R. (1991). Folding of Chymotrypsin Inhibitor-2 .1. Evidence for a 2-state transition. *Biochemistry* **30**(43), 10428-10435.
140. Sato, S. & Raleigh, D. P. (2002). pH-dependent stability and folding kinetics of a protein with an unusual α-β topology: The C-terminal domain of the ribosomal protein L9. *J. Mol. Biol.* **318**(2), 571-582.
141. Kuhlman, B., Luisi, D. L., Evans, P. A.*, et al.* (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**(5), 1661-1670.
142. McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of β-hairpin formation in protein g folding. *Nature Struct. Biol.* **7**(8), 669-673.
143. Kim, D. E., Fisher, C. & Baker, D. (2000). A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298**(5), 971-984.
144. Villegas, V., Martinez, J. C., Aviles, F. X.*, et al.* (1998). Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* **283**(5), 1027-1036.
145. Silow, M. & Oliveberg, M. (1997). High-energy channeling in protein folding. *Biochemistry* **36**(25), 7633-7637.
146. Main, E. R. G., Fulton, K. F. & Jackson, S. E. (1999). Folding pathway of FKBP12 and characterization of the transition state. *J. Mol. Biol.* **291**(2), 429-444.
147. Van Nuland, N. A. J., Meijberg, W., Warner, J.*, et al.* (1998). Slow cooperative folding of a small globular protein Hpr. *Biochemistry* **37**(2), 622-637.
148. Choe, S. E., Matsudaira, P. T., Osterhout, J.*, et al.* (1998). Folding kinetics of villin 14T, a protein domain with a central β-sheet and two hydrophobic cores. *Biochemistry* **37**(41), 14508-14518.
149. Wang, T., Zhu, Y. J. & Gai, F. (2004). Folding of a three-helix bundle at the folding speed limit. *J. Phys. Chem. B* **108**(12), 3694-3697.
150. Snow, C. D., Nguyen, N., Pande, V. S.*, et al.* (2002). Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **420**(6911), 102-106.
151. Greene, L. H., Lewis, T. E., Addou, S.*, et al.* (2007). The CATH domain structure database: New protocols and classification levels give a more

comprehensive resource for exploring evolution. *Nuc. Acid. Res.* **35**, D291-D297.
152. Murzin, A. G., Brenner, S. E., Hubbard, T.*, et al.* (1995). SCOP - a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**(4), 536-540.
153. De Los Rios, M. A., Muralidhara, B. K., Wildes, D.*, et al.* (2006). On the precision of experimentally determined protein folding rates and phi-values. *Prot. Sci.* **15**(3), 553-563.
154. Martinez, J. C. & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6**(11), 1010-1016.
155. D'Aquino, J. A., Gomez, J., Hilser, V. J.*, et al.* (1996). The magnitude of the backbone conformational entropy change in protein folding. *Proteins: Struct., Funct., Genet.* **25**(2), 143-156.
156. Fu, L. & Freire, E. (1992). On the origin of the enthalpy and entropy convergence temperatures in protein folding. *Proc. Natl. Acad. Sci. USA* **89**(19), 9335-9338.
157. Luque, I. & Freire, E. (1998). Structure-based prediction of binding affinities and molecular design of peptide ligands. *Methods Enzymol.* **295**, 100-127.
158. Creamer, T. P. & Rose, G. D. (1994). α-helix-forming propensities in peptides and proteins. *Proteins: Struct., Funct., Genet.* **19**(2), 85-97.
159. Pickett, S. D. & Sternberg, M. J. E. (1993). Empirical scale of side-chain conformational entropy in protein-folding. *J. Mol. Biol.* **231**(3), 825-839.
160. Muñoz, V. & Serrano, L. (1994). Intrinsic secondary structure propensities of the amino-acids, using statistical phi-psi matrices - comparison with experimental scales. *Proteins: Struct., Funct., Genet.* **20**(4), 301-311.
161. Hooft, R. W. W., Sander, C. & Vriend, G. (1996). Verification of protein structures: Side-chain planarity. *J. App. Cryst.* **29**, 714-716.
162. Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Prot. Sci.* **3**(3), 522-524.
163. Spath, H. (1980). *Cluster analysis algorithms for data reduction and classification of objects*, Halsted Press, New York.
164. Martinez, W. L. & Martinez, A. R. (2002). *Computational statistics handbook with matlab*, Chapman and Hall/CRC.
165. Makhatadze, G. I. & Privalov, P. L. (1990). Heat-capacity of proteins .1. Partial molar heat-capacity of individual amino-acid-residues in aqueous-solution - hydration effect. *J. Mol. Biol.* **213**(2), 375-384.
166. Naganathan, A. N., Doshi, U. & Muñoz, V. (2007). Protein folding kineitcs: Barrier effects in chemical and thermal denaturation experiments. *J. Am. Chem. Soc.* (Published online on April 10, 2007.).
167. Doshi, U. R. & Muñoz, V. (2004). The principles of α-helix formation: Explaining complex kinetics with nucleation-elongation theory. *J. Phys. Chem. B* **108**(24), 8497-8506.
168. Doshi, U. & Muñoz, V. (2004). Kinetics of α-helix formation as diffusion on a one-dimensional free energy surface. *Chem. Phys.* **307**(2-3), 129-136.

169. Wang, T., Du, D. G. & Gai, F. (2003). Helix-coil kinetics of two 14-residue peptides. *Chem. Phys. Lett.* **370**(5-6), 842-848.
170. Wang, T., Zhu, Y. J., Getahun, Z.*, et al.* (2004). Length dependent helix-coil transition kinetics of nine alanine-based peptides. *J. Phys. Chem. B* **108**(39), 15301-15310.
171. Gooding, E. A., Ramajo, A. P., Wang, J. W.*, et al.* (2005). The effects of individual amino acids on the fast folding dynamics of alpha-helical peptides. *Chem. Comm.*(48), 5985-5987.