ANALYSIS OF FOLDING KINETICS FOR SIMPLIFIED MODEL PROTEINS

by

Ş. Banu Özkan

BS. in Ch.E. Boğaziçi University, 1995

MS. in Ch.E., Boğaziçi University, 1997

Bogazici University Library

39001100869240

14

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor
of
Philosophy

Boğaziçi University

2001

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my thesis supervisor, Prof. Dr. İvet Bahar, for her invaluable supervision during this study and creating an excellent working environment. I feel that I am really lucky for having such a perfect advisor in all respects.

I would like to thank to Prof. Dr. Burak Erman for offering so much of his knowledge throughout of my study and his kind comments on the thesis. I am also grateful to Doç. Dr. Türkan Haliloğlu for her help and support in and outside of PRC. Her smiling face can make good and bad days better. Very special thanks to Prof. Ali Rana Atılgan, his enthusiasm about science and the stimulating discussions meant much to me. I would like to express my gratitude to Assit Prof. Dr. Candan T. Behar who devoted her valuable time to reading and commenting on the thesis.

It was a great pleasure for me to collobrate with Prof. Dr Ken A. Dill who made me see with other perspectives. Thanks to Assoc. Prof. Pemra Doruker for her help and patience.

I am so much indebted to all my friends for their help and support, Alpay Temiz (the problem solver!), Şafak Kırca (my thesis format advisor!), Basak Işın (who supports me with joyful laughters), Neşe Kurt and Zerrin Bağcı. I would like to also thank to my friends, Berna Sarıyar and Ebru Öner for their kind help and patience and support, Esra Küçükpınar whom I owe a lot, Arturo Cerda (my lovely English teacher!!) and Özlem Keskin who is always with me even over the ocean.

No words would be enough to thank to my family...

ABSTRACT

ANALYSIS OF FOLDING KINETICS FOR SIMPLIFIED MODEL PROTEINS

The conformational stochastics of simplified model chains that show an apparent two-state kinetics was explored. An apparent two-state kinetics refer to the occurrence of a single exponential time evolution dominating the folding process, even though the individual chains follow a broad ensemble of micropaths. A fundamental question addressed in the present analysis is to understand if the folding takes place through a continuum of paths with no distinct on-pathway forms, or if a preferred pathway involving subcooperative folding events can be discerned. To this aim, the complete sets of conformations for short model chains were generated as self-avoiding walks on a square lattice. Native-like contacts have been assigned attractive potentials, and transition rates have been assigned on the basis of native-like contacts and root-mean-square deviations between conformations. The time evolution of all conformational transitions has been analyzed starting from a uniform distribution of conformations, using a master equation formalism, which enabled us to capture the microscopic details of the folding kinetics.

It is found that the folding macropath can be described in terms of a particular sequence of events in which local interactions generally preceed the more nonlocal contacts, resembling a zippers process. The innermost contacts form first. A key conclusion from the present work is that: (i) The lack of intermediates that define two-state kinetics does not preclude folding through a specific sequence of events. (ii) Φ-value analysis, a measure of the stability and change in folding kinetics due to mutation performed on this exact lattice model reveals that non classical Φ-values can arise from parallel microscopic flow processes. Negative Φ values result when a mutation destabilizes a slow flow channel, causing an overflow into a faster flow channel. Φ-values greater than one occur when mutations redirect a fast flow into a slower channel.

ÖZET

BASİT PROTEİN MODELLERİNDEKİ KATLANMA KİNETİĞİ ANALİZİ

İki halli katlanma kinetiği gözlenen, basit model zincirlerinin stokastik konformasyonel özellikleri incelenmiştir. İki halli katlanma kinetiği, tek exponansiyel zaman evrimini işaret etmekle birlikte, moleküler düzeydeki kinetik proses herbir zincirin çok sayıda mikroskopik yoldan geçişi ile gerçekleşmektedir. Bu çalışmada şu temel soruya yanıt aranmaktadır: Protein katlanması geçiş yollarına bağlı olmaksızın, sürekli yollardan geçerek mi ya da kooperatif katlanma olaylarından dolayı belirli bir yolu seçerek mi gerçekleşir? Bu amaçla, küçük model zincirlerinin tüm konformasyonları, kare kafes üzerinde oluşturulmuştur. Doğal yapıda bulunan etkileşimlere negatif enerji değeri verilmiştir. Konformasyonlar arasındaki geçiş hızı, doğal halde bulunan etkileşimlere ve konformasyonlar arasındaki ortalama karekök sapma uzunluklarına göre belirlenmiştir. Tüm konformasyonlar arasındaki geçişlerin evrimi, herbir konformasyonun eşit olasılıkta olduğu bir ilk konumdan başlanarak, mikroskopik ayrıntıları yakalamamızı sağlayan bir temel denklem formulasyonu ile incelenmiştir.

Katlanma olaylarını denetleyen makroskopik yolların, lokal etkileşimlerin lokal olmayan etkileşimler tarafından izlendiği fermuar mekanizması ile tanımlanacağı sonucuna varılmıştır. Zincirin üç boyutlu yapısına göre en iç konumdaki etkileşimler önce oluşmaktadır. Bu çalışmadan elde edilen ana sonuçlar şunlardır: (i) İki halli kinetik ve ona karşı gelen huniyi andıran bazı enerji yüzeyleri, katlanmanın belirli olaylar dizisi şeklinde gelişmesini engellememektedir. (ii) Mutasyonların kararlılığa ve katlanma kinetiğine etkisini ölçen Φ değerlerinin negatif ve birden büyük olması, paralel katlanma prosesi nedeniyle ortaya çıkmaktadır. Yavaş katlanma kanallarını engelleyen mutasyonlar, hızlı kanallara akımı arttırmakta ve negatif Φ değeri oluşumuna neden olmaktadır. Kinetiği hızlı kanaldan yavaşına yönlendiren mutasyonlar ise birden büyük Φ değerlerine yol açmaktadır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

A	Transition
a_k	Equilibrium characteristic of the k^{th} mode
B	Matrix of Eigenvectors
$\mathbf{B^{-1}}$	Inverse of Matrix of Eigenvectors
C(t)	Conditional probability matrix
<CO>	Average contact order
d	Lattice spacing
$\Delta G^{\ddagger}$	Gibbs free energy of transition state
ΔG_D	Gibss free energy of denatured state
$\Delta\Delta G$	Gibbs free energy difference between the transition states of wild type and mutant
$\Delta^2 G_{int}$	Pairwise coupling Gibbs free energy
H	Heavyside step function
I	Intermediate
K_{mut}	Equilibrium constant of mutant
K_{wt}	Equilibrium constant of wild type
k_B	Boltzmann constant
k_f	Folding rate
k_{mut}	Folding rate of mutant
k_{wt}	Folding rate of wildtype
m	Number of native contact
N	Native state
P(t)	Instantaneous probability vector
q_i	Number of native contact of i^{th} conformation
R	Gas constant
R_{ij}	Distance between i^{ith} and j^{th} monoers
$<(\Delta r_{ij})^2>^{1/2}$	Rms deviation between conformation i and j
S	Substrate
T	Temperature
U	Unfolded

V	Potential associated with the vibrations of interaction sites
W_{mac}	Number of macroconformation
W_{mic}	Number of microconformation
α	Helix
β	Strand or Brønsted value
ε	Potential energy per native contact
κ	Tranmission coefficient
ν	Characteristic vibrational frequency
Λ	Diagonal matrix of eigenvalues of **B**
τ	Characteristic time
$\langle\tau_{wt}\rangle$	Average characteristic time of wild type contact
$\langle\tau_{mut}\rangle$	Average characteristic time of destabilized contact
δ	Delta dirac function
HX	Hydrogen Exchange
Ile	Isoluecine
MC	Monte Carlo
MD	Molecular dynamics
MG	Molten Globule
NMR	Nuclear Magnetic Resonance
rms	Root mean square
SVD	Singular Value Decompostion
TS	Transition State

1. INTRODUCTION

Understanding the process of protein folding is one of the major challenges of modern structural biology. The spontaneous folding of protein molecules with a huge number of degrees of freedom into a unique three-dimensional structure that carries out a biological function is the simplest case of biological self-organization. Thus, protein folding is a subject that attracts scientists from a wide range of disciplines. The protein folding problem is generally divided into two parts. In the first part, the aim is to predict the unique three-dimensional structure among the huge conformational space. The second part is to understand the relation between protein sequences and folding mechanism. In protein folding, mechanism is the distribution of microscopic pathways that connects countless structures of the denatured state with the unique structure of the native state.

For over a quarter of a century, ideas on protein folding mechanism have been dominated by two inter-related concepts: the Levinthal paradox, and necessity for folding intermediates. Levinthal argued that, since there is an astronomical number of conformations open to denatured state of a protein, an unbiased search through these would take forever. It was thus a short logical step to argue that there must be defined pathways to simplify the choices in folding.

Since Anfinsen's original demonstration (1973) of spontaneous protein refolding, major advances have occurred in experimental studies to determine the mechanism of folding (Fersht *et al.,* 1992; Baldwin, 1995; Creighton, 1995). A turning point was the idea of studying small, single proteins that began with the equilibrium and kinetic experiments on chymotrypsin inhibitor 2 (CI2). CI2 was found to fold and unfold as a simple two–state system with no kinetic intermediates. In subsequent work, the role of individual residues in the transition state (TS) for folding was investigated by mutational analysis, Φ-value analysis (Matouscheck *et al.,* 1989; Matouscheck *et al.,* 1992). This method has been also used by many theoreticians to understand the TS structure and the folding mechanism of simple protein models (Lazaridis and Karplus, 1997; Klimov and Thirumalai, 1998; Li *et al.*, 2000; Clementi *et al.*, 2000; Nymeyer *et al.,* 2000). Thus this method allows us to

understand the folding mechanisms relatively and unambiguously, stimulating interaction between experimentalists and theoreticians (Takada, 2000).

A second major advance in experimental studies has resulted from the introduction of a new generation of experiments with dramatically improved time resolution. Until a few years ago the kinetics of folding have been studied using stopped-flow techniques. Stopped-flow experiments have yielded an enormous amount of valuable information that provided the basis of the kinetics of folding. The fundamental limitation of this method was the poor resolution that the spectroscopic changes associated with folding already occurred within the dead time, i.e. the time required for the solutions to be mixed (Matthews, 1993; Ptitsyn, 1995). On the other hand, the new rapid mixing techniques allow us to overcome the time limit. They can be roughly classified into three categories: photochemical triggering, temperature or pressure jump and ultrarapid mixing methods. These new techniques have provided major insights into folding mechanisms. These include a much deeper understanding of the mechanism of secondary structure formation, the introduction of the notion of an upper limit on the rate of protein folding (a "speed limit"), discovery of unusual kinetics suggesting that very fast folding is continuously downhill in free energy (Gruebele, 1999; Munoz *et al.*, 2000).

There have been many theoretical approaches that shed light on the nature of folding pathways, their transition states, and the role of intermediates in folding. One of the most important progress in theoretical approaches was the energy landscape theory (Baldwin, 1995; Bryngelson *et al.*, 1995; Chan and Dill, 1997) that gave us the general framework of the folding mechanisms based on statistical physics.

So far the most important insights have come from simulations of simplified representation of proteins in lattice and off-lattice models (Shakhnovich, 1997; Chan and Dill, 1998; Pande *et al.*, 1998; Thirumalai and Klimov, 1999; Dinner *et al.*, 2000). Such models provide simple examples that can help to clarify basic principles of folding kinetics. On the other hand, all-atom molecular dynamic (MD) simulations can give the most detailed and realistic information. These simulations have been restricted by computer time to one or just a few trajectories of tens of nanoseconds. They are not capable of direct simulation of protein folding. However, the all-atom MD simulations of

unfolding trajectories of CI2 under extreme conditions (500 K and 26 atmospheres) conducted by Daggett *et al.* (1996) gave insights into the nature of the TS ensemble. Under these conditions, unfolding was accelerated by six orders of magnitude, from milliseconds to nanoseconds, and became accessible to study. They argued that the TS should correspond to a rapid change in the conformation of protein with time, and identified related conformations in four unfolding trajectories as putative TS. The remarkable success of the study was the consistency between the residues that Daggett *et al.* (1996) identified as important in the TS and those implicated by Fersht and coworkers using mutational analysis (Otzen *et al.*, 1994). Lazaridis and Karplus (1997) have also analyzed the unfolding MD trajectories of CI2 to clarify the TS. They observed that the TS region for folding and unfolding occurs early with only 25 per cent of the native contacts. Another interesting result was that the statistically preferred unfolding pathway emerged from the simulations. Zhou and Karplus (1999) used a discrete MD technique to study the folding of the small three-helix bundle fragment of Protein A. In this study, two different dominant trajectories (fast and slow tracks) were observed for the folding of helical proteins when the single energy parameter (the difference between the strength of native and non-native contacts) was changed. Dokholayan *et al.* (2000) investigated the protein folding nucleus of 46-mer off-lattice protein model using MD. The simulations revealed that a few well-defined contacts were formed with high probability in the TS that drives the protein into its folded conformation.

Monte Carlo (MC) dynamic method has been widely used to understand the basic folding principles of on-lattice simple protein models chains (Dinner *et al.*, 1996; Socci *et al.*, 1998; Pande and Rokhsar, 1999; Li *et al.*, 2000; Klimov and Thirumalai, 2000). The statistical analysis of hundreds of folding in MC trajectories of lattice model proteins performed by Pande and Rokhsar (1999), revealed a classical dominant folding pathway which was composed of on-pathway intermediates. Klimov and Thirumalai (1998) investigated the TS structure and folding nuclei of 27-mer and 36-mer lattice models. The analysis of individual trajectories showed that the polypeptide chain reaches the native state upon the formation of critical contacts, which is consistent with the nucleation-collapse mechanism. The effect of non-native contacts on the folding mechanism was analyzed by Li *et al.* (2000) using MC techniques. They concluded that the specific non native interactions in the TS would give a rise to non classical Φ-values ($\Phi > 0$ or $\Phi < 0$).

They also demonstrated that the specific residue, Ile 34 in src SH3 domain which has been shown to be kinetically, but not thermodynamically important is, universally conserved.

The relation between evolutionary pressure on folding rate and the classification of native state has been investigated by using simple protein models. The results suggests that topologically simple structures are expected to fold faster than the complicated structures (Baker, 2000). Thus, the native topology is a key determinant of folding mechanisms. The outcome of this striking result is that the basic physics underlying the protein folding problem could be relatively simple.

In the picture of the previous studies, the aim of the present study is to explore the "full" kinetics of protein folding using a simple model protein. This will be achieved by three steps:

(i) Generating the complete sets of the model chain as self-avoiding walks on a square lattice,

(ii) Constructing the full microscopic transition matrix with respect to two criteria: intramolecular energy barrier and frictional effect,

(ii) The exact solution of the matrix using a master equation formalism that has been previously used in earlier studies of folding kinetics (Leopod *et al.*, 1992; Zwanzig, 1995; Ye *et al.*, 1999; Munoz *et al.*, 1998).

Go model, which is the principal theoretical model for exploring microscopic steps in two-state protein folding kinetics, is adopted (Ueda *et al.*, 1975). Thus the folding is driven by the attractive potentials assigned to native contacts. Go models were used in earlier folding kinetic studies (Hoang and Cieplak, 2000; Pande and Rokhsar, 1999) and analytical methods for investigating the folding of proteins (Erman and Dill, 2000). Crucial to the present study is the application of a general, rigorous, and unambiguous method to identify all microtrajectories of the landscape.

The plan of the present thesis is as follows: In the following section, the classical and new view of protein folding will be explained briefly and the basic theoretical and experimental approaches for investigating the folding kinetics will be discussed. The

theoretical background of the master equation formalism and the model and parameters will presented in the third chapter. In the result section, these subjects will be discussed:

(i) time evolution of the native contacts and coupling between the native contacts,

(ii) the fluxes and the transition between macroconformations (subsets of conformations) that give insight about the folding pathway(s),

(iii) the kinetic scheme of folding,

(iv) the landscape mapping method that enables to understand the physics underlying the energy landscape,

(v) the results from Φ-value analysis, and

(vi) three-dimensional energy landscape surface plots.

In the final section, conclusions and recommendations for future work will be presented.

2. PROTEIN FOLDING

2.1. Classical View: Folding Pathways

Solving how a protein folds from its denatured state to its native state poses an intellectual challenge that is far more complex than solving classical chemical mechanisms. In protein folding, the whole molecule changes in the structure. Thousands of weak non-covalent interactions are made or broken, unlike in simple chemical reactions. However the basic strategy for the analysis of protein folding is the same. That is to characterize all the stable and metastable on and off reaction pathways and the transition state that connects them.

Ideas about protein folding were dominated by two interrelated concepts: the Levinthal paradox and the supposed necessity of folding intermediates. Levinthal argued that because of the astronomical conformations, it would take an absurdly long time for even a small protein to explore all the accessible conformations for folding. Yet, many proteins fold to their native conformations in less than a few seconds. Therefore, the classical view of protein folding suggests that the search for the native state through the immensity conformations flows through the predetermined pathways defined by discrete intermediates and barriers (Rumbley *et al.*, 2001).

Folding pathways are explained by three basic mechanisms (Figure 2.1.). The *framework model* (Ptitsyn and Rashin, 1975) proposed that protein folding starts with the formation of secondary structural elements independently from the tertiary elements. These elements then assemble into tightly packed native tertiary structure by *diffusion and collision mechanism* (Karplus and Weaver, 1976). In the *diffusion collision mechanism*, the secondary structural elements would diffuse until they collide in order to coalesce into a structural entity with the native conformation. The *hydrophobic collapse model* for protein folding proposed that the initial event of the reactions is the collapse of the protein molecule, mainly driven by the hydrophobic effect and then rearrangement of the stable structures in the collapsed state (Dill, 1990). The *nucleation condenstation* mechanism argued that early formation of a diffuse folding nucleus catalyzes the folding. The nucleus

primarily consists of adjacent neighbour residues that have some correct secondary structure interactions but it is stable with the assistance of the tertiary interactions (Fersht, 1997).

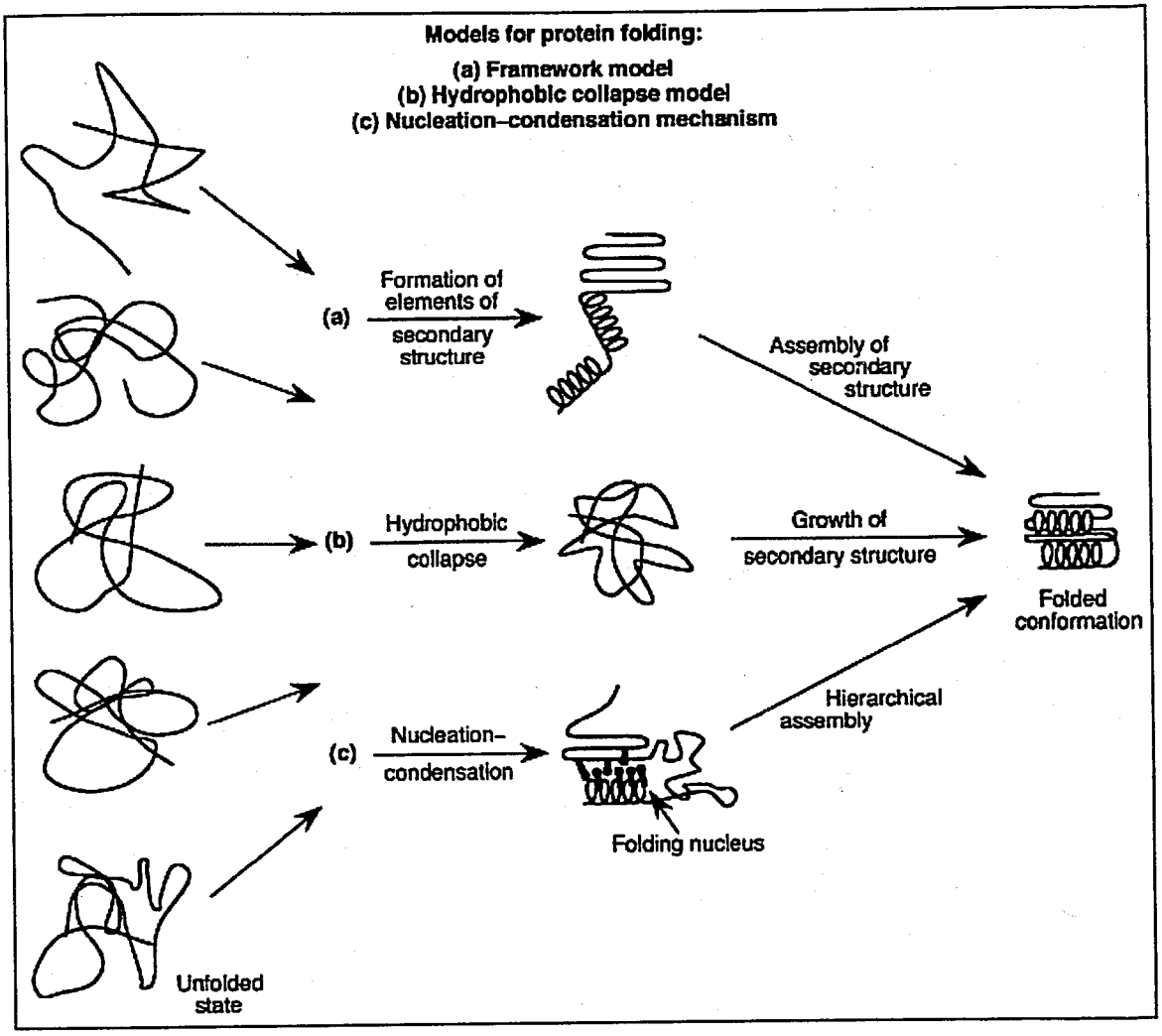

Figure 2.1. Schematic representation of folding mechanisms (Nolting and Andert, 2000)

A contrary view, the jigsaw model, is that each protein folds by a different distinct path (Harrison and Durbin, 1985). All the three mechanisms described above may be extended to proteins that have intermediates and multiple transition states on their pathways which is closer to the new landscape view (Leopold *et al.*, 1992).

The Levinthal paradox presupposes that the search is unbiased so that the groups on the protein rotate around their single bonds at random without any stabilization of any

particular conformation until they are all in the right conformation. If there is a bias on the sequence toward the correct structure, then the paradox disappears (Finkelstein and Badretdinov, 1997).

2.2. New View: Energy Landscapes

The new view of the protein folding introduces the funnel landscape, which provides an alternative to the classical view that there must exist a single pathway for the folding with well-defined intermediates (Figure 2.2). Within this view, protein folding is a collective self-organization process that generally does not occur by an obligate series of intermediates, "a pathway" but by a multiplicity of routes down a folding funnel (Onuchic, 2000). Thus the new view envisions folding as representing the ensemble average of a process that is microscopically more heterogenous (Chan and Dill, 1998)

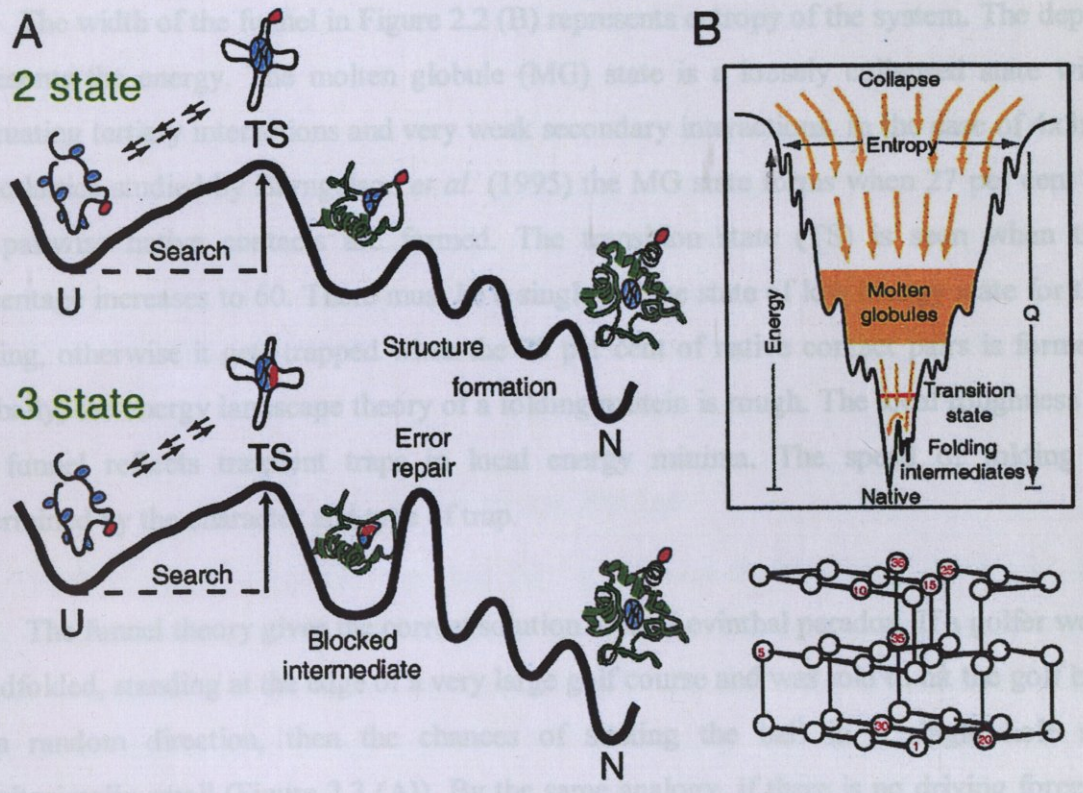

Figure 2.2. (A) Classical pathway consisting a single path, intermediates and the transition state structure (TS), (B) New view of folding funnel (Rumbley *et al.*, 2001)

In the landscape view, each individual protein molecule may follow its own trajectory, but they all may eventually reach the same point at the bottom, the native state. At the top of the funnel, the protein molecules exist in a large number of conformations that have relatively high free enthalpy and entropy that is called denatured state. There is a competition between the entropy keeping the protein as random as possible at the top and the minimization of the enthalpy dragging the protein down the funnel. Progress down the funnel is accompanied by an increase in native-like structures, meanwhile the routes to native state decrease. The folding can be described by a Brownian type of motion between the conformations that are geometrically similar and follows a general drift from higher energy to lower energy conformations. When folding or unfolding kinetics involves transient accumulation of partially folded or unfolded intermediates then energy landscapes are bumpy with kinetic traps. In contrast, when folding and unfolding each involve only a 2-state kinetics (denatured and native state), then energy landscapes can be represented as smoother funnels, with no substantive traps.

The width of the funnel in Figure 2.2 (B) represents entropy of the system. The depth represents the energy. The molten globule (MG) state is a loosely collapsed state with fluctuating tertiary interactions and very weak secondary interactions. In the case of 4x3x4 cubic lattice studied by Bryngelson *et al.* (1995) the MG state forms when 27 per cent of the pairwise native contacts are formed. The transition state (TS) is seen when the percentage increases to 60. There must be a single unique state of low energy state for the folding, otherwise it gets trapped when the 70 per cent of native contact pairs is formed. Globally, the energy landscape theory of a folding protein is rough. The local roughness of the funnel reflects transient traps in local energy minima. The speed of folding is determined by the character and type of trap.

The funnel theory gives the correct solution to the Levinthal paradox. If a golfer were blindfolded, standing at the edge of a very large golf course and was told to hit the golf ball in a random direction, then the chances of sinking the ball in a single hole are infinitesimally small (Figure 2.3 (A)). By the same analogy, if there is no driving force to push the protein in the direction of folding when the protein exists in a large number of non-native states of equal energy and has just one native state of the lowest energy, that is the Levinthal paradox. However, if the golf course sloped down from all directions to the

hole, the gravity would funnel the ball to the hole and the golfer would always score a hole in one (Figure 2.3 (B)).

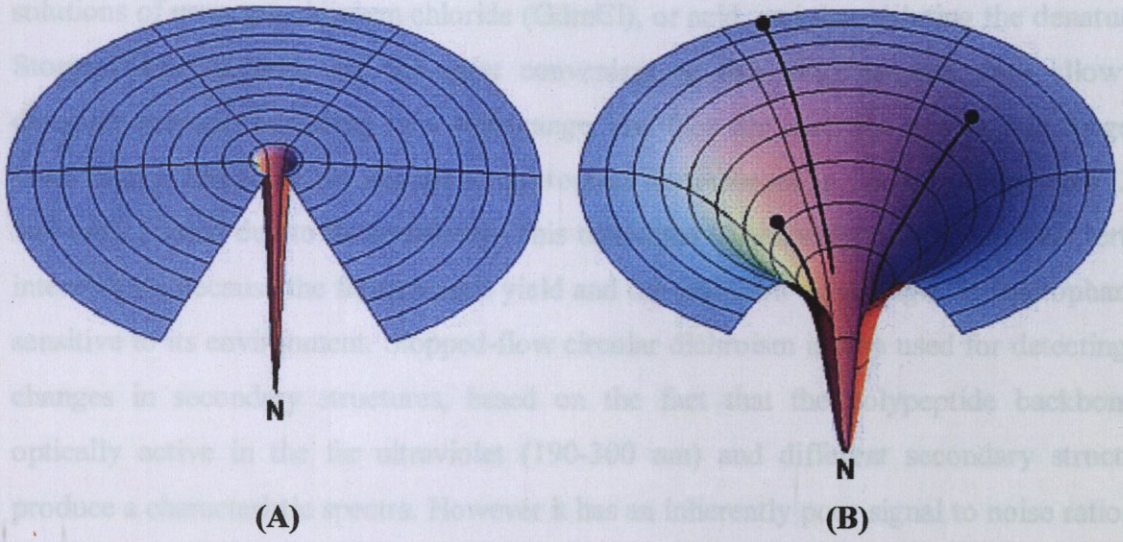

Figure 2.3. The schematic representation of golf course funnel (A) and smooth landscape funnel (B) (Chan and Dill,1998)

The energy landscapes of the new view and the pathways of the classical view can be united by using the ideas of ensembles. The folding of small proteins have energetically downhill folding starting from the denatured state ensemble that consists of many different conformations to the more restricted state (the transition state) and finally the native state. Some of the states (intermediates) that accumulate due to bumpy structure of the funnel are determined by experimentalists.

2.3. Kinetics of Folding

The fundamental questions in protein folding concerns the kinetics of folding which is not yet understood in detail. The discovery of the reversible denaturation of the small proteins ribonuclease A and staphylococcal nuclease (Anfinsen, 1973) has enhanced many experimental and theoretical approaches for understanding the mechanism of kinetic processes. The kinetics of folding and unfolding appear to be a very complex process, but these processes are governed by the basic rate laws. The basic principles and powerful experimental and theoretical approaches investigating the protein folding will be discussed in this chapter.

2.3.1. Experimental Approaches

The folding of proteins is usually studied in *vitro* by first denaturating them in solutions of urea, guanidinium chloride (GdmCl), or acid, and then diluting the denaturant. Stopped-flow methods are the most convenient to this aim, because they allow for detecting changes occurring on a wide range, and they are ideal for mixing two reagents. Fluorimetry, following the change in tryptophan fluorescence in the near ultraviolet (290-300 nm), is used due to its sensitivity. This technique is generally used to monitor tertiary interactions, because the fluorescence yield and the emission wavelength of tryptophan are sensitive to its environment. Stopped-flow circular dichroism is also used for detecting the changes in secondary structures, based on the fact that the polypeptide backbone is optically active in the far ultraviolet (190-300 nm) and different secondary structures produce a characteristic spectra. However it has an inherently poor signal to noise ratio and requires considerably more protein than does fluorescence (Gruebele, 1999).

Helices usually form in a few hundred nanoseconds and β-turns in a few microseconds in model proteins (Fersht, 1999). Short loops in proteins form with an upper limit of about 10^6 s^{-1}. Thus, a lower limit for the initial collapse of a denaturated protein is about 1 μs. Conventional rapid mixing methods are limited to a time scale of milliseconds and greater, but specialized continuous-flow apparatus have been used for tens of microseconds.

The time involved in mixing places a limit on the dead time of flow techniques. The only way to increase the time resolution is to cut out the mixing by using a pre-mixed solution of reagents that can be perturbed in some way to allow a measurable reaction to occur. A classic method from physical chemistry is flash photolysis, in which a particular bond is cleaved by a pulse of light so that reactive intermediates are formed. An alternative method of overcoming the time delay of mixing is to use a relaxation method. An equilibrium mixture of reagents is preincubated and the equilibrium is perturbed by an external influence. Then the relaxation to equilibrium is measured. The most common procedure for this is temperature jump. A solution is incubated in an absorbency of fluorescence cell and its temperature is raised in less than a microsecond. The system will proceed to its new equilibrium via a series of relaxation times if the equilibrium involves

an enthalpy change. The temperature jump method is ideal for denaturing proteins by jumping from ambient to elevated temperatures. This method has been adopted to study rapid events on time scales of nanoseconds to microseconds in the folding of a cold denatured protein (Ballew *et al.*, 1996). However, rapid mixing techniques provide a much more dramatic change in the folding equilibrium constant, allowing detection of the full reaction.

The diverse methods briefly discussed above provide information on folding rate, signal the accumulation of intermediates, and give insights as the role of particular amino acids or estimate about the average parameters of the main chain. However, hydrogen exchange (HX) is another method that can be used to identify the structures of intermediates in both kinetic and equilibrium modes and quantify their thermodynamic stability. HX information defines the structure by identifying the amino acids that have been slowly exchanging; presumably hydrogen-bonded main chain amides. It first became possible to define transiently formed submolecular structures by the use of hydrogen-deuterium (H-D) exchange labelling together with stopped-flow methods and high resolution proton NMR. The protein, unfolded in GdmCl and deuterated by exchange in D_2O solvent is diluted into H_2O to initiate refolding. After various experimental folding times, a brief pulse to a higher pH is used to promote fast D to H exchange and label with H the main chain amides that are not yet protected by hydrogen bond formation. The protein folds to its native state, trapping the H-D labelling profile imposed during the labelling pulse, which can then be read out by two-dimensional NMR. It is also possible to identify the MG states and secondary interactions by HX method (Rumbley *et al.*, 2001) Yet, HX at equilibrium cannot be used to determine pathways because equilibrium measurements only give information on the thermodynamic properties of intermediates and not the pathway between them (Fersht, 1999). HX measurements at equilibrium have been related to the amplitude of fluctuations near folded state (Bahar *et al.*, 1998).

Refolding is generally found to proceed by a series of exponential phases. Some of these exponentials are related to the *cis-trans* isomerization of peptidyl-prolyl bonds. The equilibrium constant for the normal peptide bonds in proteins favors the *trans* conformation, however the peptidyl-prolyl bond assures the *cis* form with 2-20 per cent probability. This exceptional character of peptidyl-prolyl bonds enables us to see the

proline isomerization in the folding kinetics, since all the prolines in the *trans* conformations in the native structure will equilibrate upon denaturation to give a mixture of *cis* and *trans* transforms (Kyte, 1995).

2.3.2. Theoretical Approaches

There are various theoretical methods for analyzing protein folding ranging from atomic analysis to simple model polymer chains. Molecular Dynamic (MD) simulations are used for exploring the intermediates and transition states visited during the unfolding process, and the energetics of folding. These simulations are based on Newton's law of motion for every atom including the solvent molecules (Wong *et al.*, 2000). Such "all atom" representation suffer from computational time and memory calculations. Generally, experimentally determined structures are taken as the initial structure. So the unfolding or earls stage of unfolding are explored. Folding process cannot be explored by conventional MD simulations.

In order to increase efficiency and ensure complete coverage of conformational space, proteins can be more simplified into strings of beads that are arranged in two- or three-dimensional lattices. Each bead of the string is spherical and has no side chains. The specificity can be added by assigning polar or hydrophobic characters to each bead or assuming interactions at certain bead pairs. Off-lattice simulations, on the other hand use a coarse-grained structure of the polypeptide chain, which offer a realistic representation of the secondary and tertiary structures. The most significant advantage of simplified models is that folding can be started from a random coil and well-formulated questions about the general principles of folding kinetics can be quantitatively answered (Thirumalai and Klimov, 1999). The Monte Carlo (MC) techniques are common methods used to observe the folding pathways starting from random coil structures. The possible moves such as tail flip, corner flip and crankshaft are determined in a probability range by random walks. Moves are accepted according to energy criteria. Different pathways of different simulations are then analyzed statistically (Li *et al.*, 2000). Another approach is the spin glass theory, which was originally formulated to analyze the orientation of spins of ferromagnetics. The spins favor being antiparallel to each other even they cannot all satisfy this requirement. One can think of the proteins as heteropolymers having random spin

coupling constants (Brygelson and Wolynes, 1987). A residue in a protein is frustrated if it wants to be in more than one conformation. The most stable structure is minimally frustrated compared with misfolded states. The insights that have come from these computational methods can be summarized as: (i) Protein folding involves a serious of structures with decreasing energy levels so that the final state has distinctly lower energy than others, (ii) Simplified models fold much better than random sequences and give reasonable results comparable with experimental results (Fersht, 1999).

2.3.3. Transition State in Protein Folding

Since the earliest experimental work on protein folding kinetics (Ikai and Tanford, 1971; Tsong *et al.*, 1971; Dill and Chan, 1997), experimental folding has been described by simple mass-action kinetics models, generally involving two states unfolded (U) and native (N) as,

$$U \rightleftharpoons N \tag{2.1}$$

or three states including an intermediate (I) that may be on-pathway,

$$U \rightleftharpoons I \rightleftharpoons N \tag{2.2}$$

or off-pathway,

$$\begin{array}{ccc} U & \rightleftharpoons & N \\ \uparrow\downarrow & & \\ I & & \end{array} \tag{2.3}$$

An extension of the scheme II is the sequential transition

$$U \rightleftharpoons I_1 \rightleftharpoons I_2 \rightleftharpoons \cdots \rightleftharpoons N \tag{2.4}$$

The intermediate accumulation may be non-productive (2.3), or productive but acting as a kinetic trap (2.2 or 2.4). The advantage of such a description in terms of two or more states is its simplicity. Due to their simplicity, studies of two-state proteins have been widely exploited for understanding the nature of the folding transition state (TS).

TS structure is a saddle point on the free energy surface, at a maximum of energy along the reaction coordinate (Figure 2.4). For a reaction involving a series of intermediates, the reaction of each intermediate involves a separate TS. The conventional description is that the highest energy TS is referred to as the TS for the overall reaction. The TS of protein folding can be described by statistical mechanics and Newton's equation. The TS is an ensemble of states that differ slightly from each other in energy around the saddle point. This concept of energy surface can lead to TS structures occupying at very wide and long saddle positions with many small dips in the profile. The energy surfaces in protein folding also have very high entropy components because of the large changes in configurational entropy and the change in the entropies of hydration (Matouschek *et al.*, 1989).

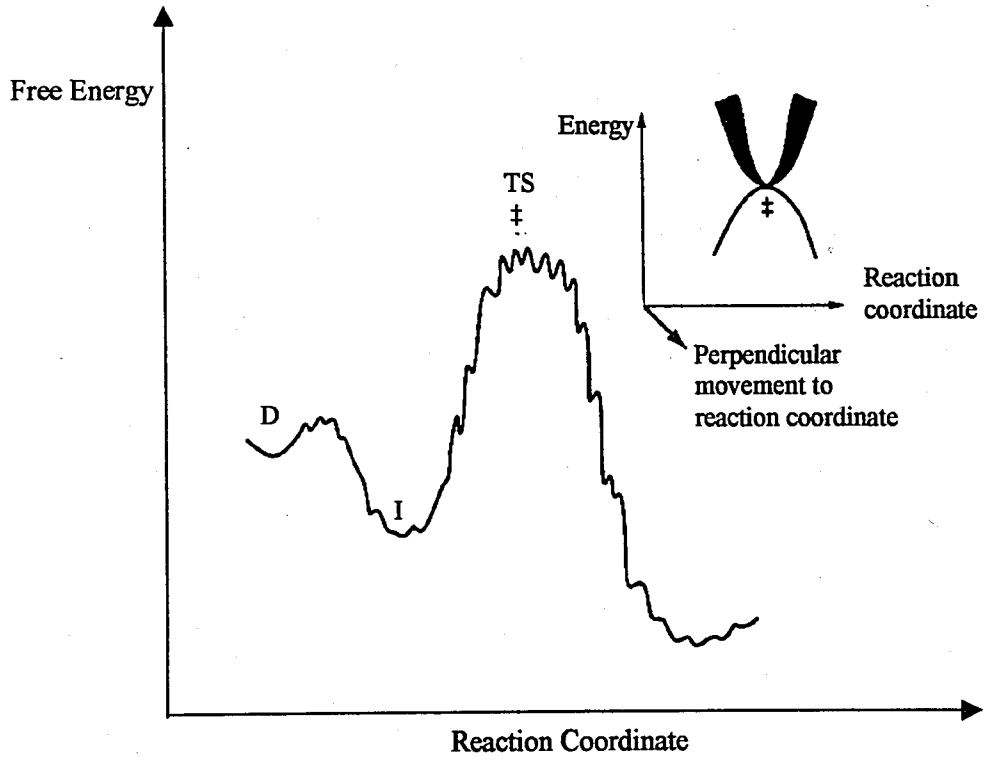

Figure 2.4. Sketch of a reaction coordinate and a saddle point (Fersht, 1999)

A simple way to derive an equation for a chemical reaction at the TS is to consider that the TS and the ground state are in thermodynamic equilibrium. Suppose that the difference in Gibbs free energy between the TS, $X^{\ddagger}$ and the ground state X is $(\Delta G^{\ddagger} - \Delta G_D)$.

Then the fractions of molecules in the transition state can be defined by using the Boltzman equation

$$\frac{[X]}{[X]^{\ddagger}} = \exp[(\Delta G^{\ddagger} - \Delta G_D)/RT] \qquad (2.5)$$

where R is the gas constant and T is the temperature. Then, the folding rate will be:

$$k_f = \nu\kappa \exp(-(\Delta G^{\ddagger} - \Delta G_D)/RT) \qquad (2.6)$$

ν is the characteristic vibrational frequency along the reaction coordinate at the saddle point and κ is the transmission coefficient.

Equation (2.5) is occasionally used for the estimation of the barrier heights assuming that κ is 1.0 and $\nu = k_B T/h$, where k_B is the Bolztman constant and h is the Planck's constant. The main use of Equation (2.5) is however calculation of change in the energy difference between the TS and ground state, $\Delta(\Delta G^{\ddagger} - \Delta G_D)$. In this case the front terms $\nu\kappa$ cancel out. For example, if a mutant folds with a rate constant $k_f`$, compared with the folding rate of wild type k_f, then the change in $(\Delta G^{\ddagger} - \Delta G_D)$ upon mutation is,

$$\Delta(\Delta G^{\ddagger} - \Delta G_D) = RT \ln (k_f`/k_f) \qquad (2.7)$$

2.3.4. The Hammond Postulate and Brønsted Theory

A useful guide in the analysis of TS is the Hammond postulate. It states that if there is an unstable intermediate on the reaction pathway, the TS for the reaction will resemble the structure of this intermediate based on the assumption that the unstable intermediate will be in the local minimum at the top of the reaction coordinate. This helps to predict the TS structure and the types of stabilization. It is really difficult to apply the Hammond postulate to biomolecular reactions, since these involve two molecules condensing to form

one transition state and a large part of the Gibbs free energy change is due to entropy. Hammond postulate applies mainly to the energy differences, so it works best with unimolecular structures.

Substrates, intermediates, and products can be viewed as sitting at the bottom of U-shaped energy wells before the reaction occurs. The shapes at very bottom are parabolic for small changes in the reaction coordinate. We can see the basis of the Hammond postulate by drawing a reaction coordinate diagram as the intersection of two parabolic curves (Figure 2.5). The energy curves of the substrate (S) and product (P) intersect at the TS on the reaction coordinate. Destabilizing the substrate raises its energy by amount ΔΔG. This makes its curve shift upward so that it intersects with the product curve at a higher energy. Thus the point of intersection moves closer to the original locus of the energy well of the substrate and as a result, the TS is reached earlier in the reaction. Clearly, the TS approaches the structure of substrate. The movement along the reaction coordinate is known as the Hammond effect (Fersht, 1999).

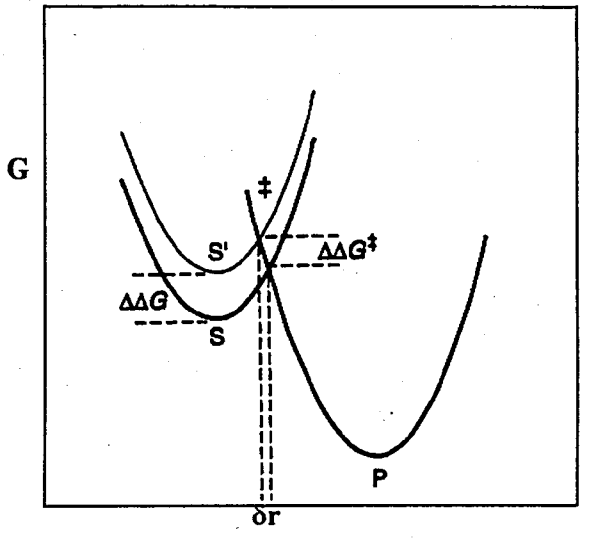

Figure 2.5. Schematic representation of Hammond postulate (Fersht, 1999)

The change in the equilibrium energy of substrate, ΔΔG thus leads to a change in the Gibbs free energy of activation, $\Delta\Delta G^{\ddagger}$. The value of $\Delta\Delta G^{\ddagger}$ depends on location of the intersection and the shape of the curves. The ratio $\Delta\Delta G^{\ddagger}/\Delta\Delta G$ is called Brønsted β value

which is between 0 and 1. β-value is used as indication of the extend of bond formation and dissociation in the TS.

2.4. Thermodynamics of Folding

The thermodynamic hypothesis says that the native conformations of proteins are global free energy minima, and the experimental evidence of reversible folding and unfolding reactions supports this view. Folding of a protein can be regarded as thermodynamically driven transition from a state of high free energy (denatured state) to a state of low free energy (native state).

Folding of a random-coil polypeptide chain into a unique conformation, which involves a tremendous decrease in the order of the system, can be thermodynamically viewed as a significant decrease in the entropy of the system. An entropy decrease is thermodynamically unfavorable, so this entropy effect should be compensated by the energy gained as a result of a redistribution of various intramolecular interactions between the protein groups and the environment. Folded proteins can be unfolded easily by a slight change in the environmental conditions. This indicates that the folding process must alter only various non-covalent interactions between the protein groups. On the other hand, the uniqueness of the native state implies that this molecule is a highly cooperative system (Creighton, 1992; Schulz and Schirmer, 1979).

Figure 2.6 illustrates the free energy surfaces for a hypothetical protein under thermodynamic (A) or kinetic control (B). In reality, the free energy surface for a protein would be an extremely high dimensional space; for simplicity a two-dimensional surface plot is shown.

Figure 2.6 (A) depicts a situation where there is a single global energy minimum that is accessible from any point on the energy surface. The outcome of the folding reaction is independent of the starting configuration and the folding reaction would be under thermodynamic control seeking out the most stable state However, a more convoluted energy surface with multiple minima (Figure 2.6 (B)) implies that the reaction would strongly depend on the starting point Molecules starting on the left might be trapped in the

local energy minimum, whereas those on the right would rapidly reach the global energy minimum. The folding studies of different proteins reveal that the free energy barriers in polypeptide chain conformational space can have appreciable magnitude and the free energy landscapes can be considerably more complex than the thermodynamic view of Figure 2.6 (A) (Baker and Agard, 1994).

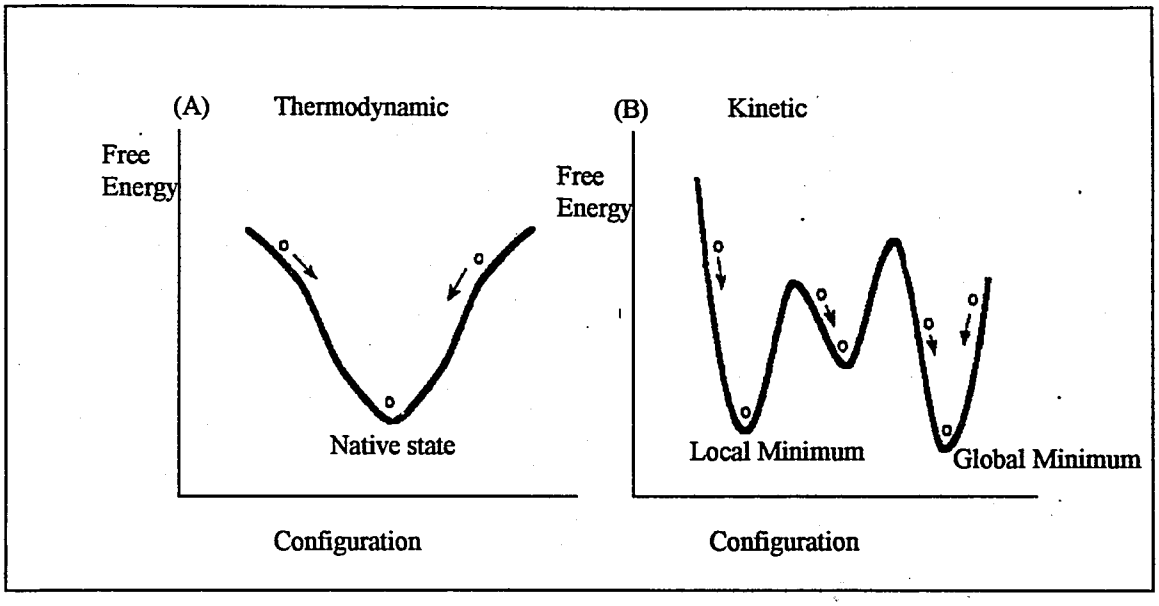

Figure 2.6. Schematic diagram of two-dimensional energy surface plot contrasting two extremes (A) Thermodynamic (B) Kinetic control (Baker and Agard, 1994)

Recent experimental and theoretical studies indicate that the fundamental physics underlying folding may be simpler than previously thought, and the topology of the protein's native state plays a crucial role on the features of folding free energy landscape. Thus protein folding mechanism can be predicted to some extent, using simplified models (Baker, 2000).

3. ANALYSIS OF FOLDING KINETICS USING THE MASTER EQUATION FORMALISM

3.1. Master Equation Formalism

3.1.1. Previous Studies of Master Equation Formalism

The master equation formalism was used in a number of earlier studies for classifying the states according to the actual kinetics of the underlying simple chain models. Leapod *et al.* studied the folding of simple lattice chains having different sequences by applying the master equation formulation (Leapod *et al.*, 1992). In this study, the transition matrix was such that only local transitions would govern the folding. They concluded that two sequences are considered as foldable and nonfoldable because one gives rise to a single large folding funnel leading to a native state and the other has multiple pathways leading to several stable conformational states. A transition matrix approach which is equivalent to the master equation formalism with a finite time approximation was used by Chan and Dill (Chan and Dill, 1993) for analyzing the macromolecular collapse dynamics. Their method is based on the transition probabilities at certain time intervals. As the time unit becomes infinitesimally small, the approach reduces to a standard master equation formalism. Chan and Dill enumerated all the conformations of a simple lattice model and applied the transition matrix approach to explore all the possible kinetic pathways for folding of the simple lattice model.

The folding of many proteins appears to be a two-state kinetics. A two state kinetic model is suitable, if protein molecules rapidly equilibrate between different unfolded conformations prior to complete folding. Zwanzig was the first to apply the master equation formalism to describe protein folding by two-state kinetics (Zwanzig, 1995; Zwanzig 1997). Ye *et al.* used the general Laplace transformation solution of master equation formalism to describe the folding kinetics of small portion of Staphyloccocal Protein A (Ye *et al.*, 1999). They concluded that the protein folds in a fast cooperative

process and neither the initial state nor the number of local energy minima affect the protein to reach the native state after a sufficiently long time.

The master equation for 12-monomer lattice heteropolymers has been solved numerically by Cieplak *et al.* and the time evolution of the occupancy of the native state has been determined (Cieplak, *et al.*, 1998).

3.1.2. Formulation of the Equation

Consider a model protein molecule having N accessible conformational states. The time evolution of these states is controlled by the *master equation*

$$\mathbf{dP(t)/dt = A\,P(t)} \tag{3.1}$$

where **P**(t) is the N-dimensional vector of the instantaneous probabilities of the N conformations, and **A** is the NxN *transition (*or *rate) matrix* describing the kinetics of the transitions between these conformations. By definition, the ijth off-diagonal element (A_{ij}) of **A** is the rate constant for the passage from conformation j into conformation i. From the principle of detailed balance, $A_{ij}\, p_j^0 = A_{ji}\, p_i^0$, where p_i^0 is the equilibrium probability of the ith conformation. The ith diagonal element of **A**, on the other hand, represents the overall rate of escape from conformation i. It is found from the negative sum of the off-diagonal elements in the same column, i.e. $A_{ii} = -\,\Sigma_j\, A_{ji}$ $(j \neq i)$.

Equation (3.1) represents a set of N equations, to be solved simultaneously. The formal solution can be cast into a tractable form by decomposing **A** as

$$\mathbf{A = B\,\Lambda\,B^{-1}} \tag{3.2}$$

where **B** is the matrix of the eigenvectors of **A**, Λ is the diagonal matrix of its eigenvalues λ_i ($\lambda_1 = 0$ and $\lambda_i < 0$ for $2 \leq i \leq N$) and $\mathbf{B^{-1}}$ is the inverse of **B**. The time-dependent probability of occurrence of the ith conformation (i.e. the ith element of **P**(t)) can be expressed in terms of the elements of **B**, Λ, $\mathbf{B^{-1}}$ and **P**(0) as

$$P_i(t) = \sum_{j=1}^{N}\sum_{k=1}^{N} B_{ik} \exp(\lambda_k t)[B^{-1}]_{kj} P_j(0) = \sum_{j=1}^{N} C(i,t \mid j.0) P_j(0) \qquad (3.3)$$

where C(i,t|j,0), denotes the *conditional* or *transition probability* of conformation i at time t, given conformation j at t = 0. For stationary processes, C(i,t|j,0) is independent of the initial time of observation, but depends on the time interval t, only, between two successive events, such that $C(i,t_2|j,t_1) = C(i,t|j,0)$ for $t = t_2 - t_1$. In matrix notation, Equation (3.3) reads

$$\mathbf{P}(t) = \mathbf{B} \exp\{\Lambda t\} \mathbf{B}^{-1} \mathbf{P}(0) = \mathbf{C}(t) \mathbf{P}(0) \qquad (3.4)$$

where $\exp\{\Lambda t\}$ is a diagonal matrix whose ith element is $\exp\{\lambda_i t\}$, and $\mathbf{C}(t)$ is the conditional or transition probability matrix. $\mathbf{C}(t)$ fully controls the stochastic process of NxN transitions. The *time-delayed joint probability* of conformations i at time t_2 and j at time t_1 is found from

$$P(i,t_2; j,t_1) = C(i,t_2-t_1|j,0)\, P_j(t_1) \qquad (3.5)$$

Combination of these probabilities in

$$P(A, t_2; B,t_1) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} C(i, t_2- t_1|j,0)\, P_j(t_1) \qquad (3.6)$$

yields the time-delayed joint probability $P(A, t_2; B, t_1)$ of specific conformational subsets (or macroconformations) *A* and *B* of interest. N_A and N_B denote the numbers of conformations in these subsets.

It is worth noting that equitation (3.3) may be reorganized to express $P_i(t)$ as a sum of exponentials

$$P_i(t) = \sum_{k=1}^{N} a_k \exp(-\lambda_k t) \tag{3.7}$$

where $(-\lambda_k)$ is the frequency of the kth mode of motion, and a_k is the corresponding amplitude factor. a_k is an equilibrium characteristic of the kth mode; it is related to the eigenvectors of **A** and the initial distribution of conformations as

$$a_k = \sum_{j=1}^{N} B_{ik} [B^{-1}]_{kj} P_j(0) \tag{3.8}$$

This equation follows from comparison of Equations (3.3) and (3.7). The frequencies are usually organized in ascending order, such that $\lambda_1 = 0$, and $-\lambda_2$ is the frequency of the slowest mode of conformational motion. The latter can also be viewed as the *global* folding mode, while the high frequency modes refer to *local* structure formation or conformational fluctuations.

3.2. Model and Method

3.2.1. Models and Native Conformations

3.2.1.1. Model. In the present study, short model chains (9-mers and 16-mers) generated on square lattices are considered. These models show apparent two-state kinetics according to the criterion that folding and unfolding kinetics can be approximated by a single-exponential. These models present the advantage of exploring the time evolution of the complete ensemble of NxN conformational transitions, thus providing exact and detailed information on the mechanism or pathway(s) of folding. The accessible conformations consist of all self avoiding walks generated on a square lattice, including both the extended conformations that dominate the denatured state, and compact forms confined to 3x3 (or 4x4) lattices. Exhaustive enumeration yields N = 740 and 802,075 distinct conformations for the 9-mers and 16-mers, respectively, excluding the conformers that are related by symmetry or rigid body rotation. The analysis of the complete ensemble enables us to

capture any microscopic detail, any sequence-kinetics relationship, or any structural aspect of how the chains actually fold.

3.2.1.2. Native Conformation. There exist three maximally compact conformations for a 9-mer on a square lattice (Figure 3.1(a)-(c)). Each have four intramolecular contacts, labeled as A-D, E-G and I-L, respectively. Calculations are performed for each of these conformations selected as the native state. The conformation (a) is analyzed in more details, since this model involves both local (between monomers i and i+3) and non-local contacts, grouped in two sequentially separate domains.

The 16-mer, on the other hand, has 31 maximally compact conformations having nine native contacts each (confined to 4x4 lattice). Among these, the conformation shown in Figure 3.1(d) is selected here as the native structure. This structure may be viewed as a simplified model for a protein comprising two domains, an α-helical and a β-sheet. Contacts A-C are representative of helical contacts, G-I are β-strand contacts. These six may be viewed as local and non-local intradomain contacts, while D-F are inter-domain contacts that assemble these secondary structures. The analysis of the time evolution of these contacts should provide insights about the hierarchical formation, if any, of different types of contacts during the folding process.

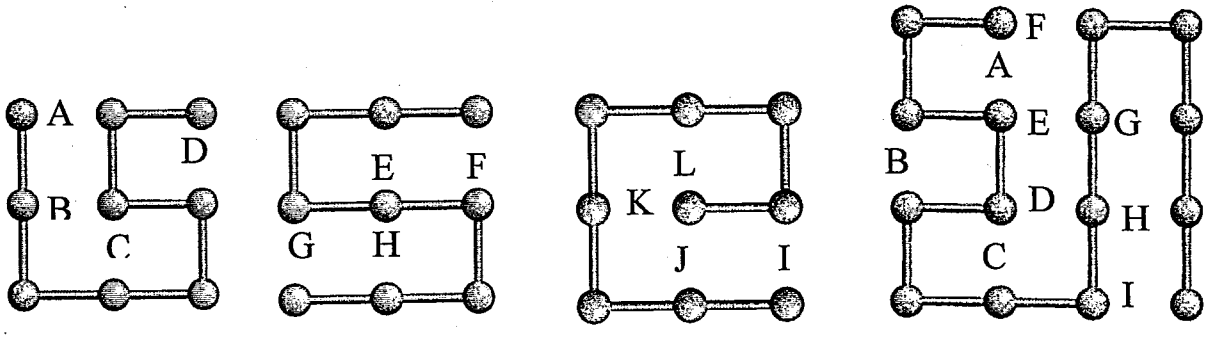

Figure 3.1. The lattice model of the native conformation previously explored, for 9-mer and 16-mer

Boğazici Üniversitesi Kütüphanesi

3.2.2. Energetics and Parameters

The passage from a uniform distribution of all conformations (fully denatured state) into a predefined native conformation, i.e. the folding process, is analyzed by the master equation formalism described in Chapter 3.1.2. The classical Go model is adopted (Ueda *et al.*, 1975). Folding is driven by attractive potentials assigned to pairs of monomers making native contacts. The native contact formation is accompanied by an energy decrease ε, and its dissociation by energy increase ε; all non-native contacts have zero energy, i.e. they do not involve enthalpy change. Go models are used because they have the same large conformational search as proteins; they have a unique lowest energy 'native' state, and they have two-state kinetics. The full energy landscape is not yet known for more atomically detailed models but it is possible to explore the details of the landscape using a Go model. Go Models were used in earlier folding kinetic studies (Hoang and Cieplak, 2000; Pande and Rokhsar, 1999), and the results were comparable to those obtained with full model (Klimov and Thirumalai, 1999).

Conformational transitions are assigned rate constants according to two criteria: intramolecular energy barrier and frictional effect. These are respectively enthalpic and entropic in origin. The energy barrier slows down the transition among the different conformations. The energy barrier height is taken as zero for passages to an equal or lower energy conformation, which favors the transition to more native-like conformations. The barrier is taken as the energy difference between the initial and final conformations in the case of passages to higher energy conformations. The frictional effect, on the other hand, accounts for geometric accessibility between conformations. It slows down, or practically hinders, the passage between highly dissimilar conformations. It scales with the root-mean-square (rms) deviation between the bead positions of communicating conformations. The rate constant A_{ij} is therefore given by

$$A_{ij} = \exp\{-\Delta G_{ij}/RT\} = \exp\{-\nu \langle(\Delta r_{ij})^2\rangle^{1/2}\} \exp\{-(q_i - q_j)\varepsilon\ H(q_i, q_j)/RT\} \qquad (2.9)$$

for $q_i < q_j$, where ΔG_{ij} is the free energy change accompanying the transition, q_i is the number of native contacts occurring in conformation i, $H(q_i, q_j)$ is the Heavyside step

function, equal to one for $q_j > q_i$, and zero otherwise. $<(\Delta r_{ij})^2>^{1/2}$ is the rms deviation betwen the conformations i and j evaluated after optimal superposition of the two conformations, and ν is a proportionality constant dependent on the strength of the frictional effect. In the absence of viscous effects, this parameter equates to zero. The frictional resistance could instead be included through an inverse proportionality on macroscopic viscosity, following Kramer's rate expression, in conformity with the modeling of protein folding as a diffusional process (Jacob and Schimid, 1999). But it is chosen to resort to the present explicit form that discriminates between individual transitions on the basis of the 3-dimensional similarities/differences of the communicating states.

The values $\varepsilon = -5$ RT and $\nu = 0.5$ were adopted for the 9-mers, and $\varepsilon = -2.3$ RT and $\nu = 1.0$ for the 16-mers. Bonds have unit length. Relatively weaker potentials and higher frictional resistances were used in the 16-mers due to physical and technical reasons: (i) physically, the moderate driving potential for folding enables us to examine the time evolution of contacts and possible accumulation of intermediates, (ii) technically, the computational overflows arising from the exceedingly large time scale difference between the fast and slow processes are avoided.

3.2.3. Initial Conditions and Equilibrium Distribution and Unit Time Steps

Calculations for the 9-mers are performed using a uniform distribution of all conformations, i.e. $P_i(0) = 1/N = 1/740$ for all i as initial conditions. This represents the infinite temperature limit. The ensemble converges to the Boltzmann distribution at 300 K at long times. The equilibrium probability of the native conformation (n) is $P_n(\infty) = 0.9848$ using $\varepsilon = -5$ RT. Therefore, the stochastic process of folding to the native state starting from a uniform distribution of conformations is observed to investigate the different pathways.

In the case of the 16-mers, the Boltzmann distribution at 500K is preferred over the infinite temperature limit as the initial condition. The net effect is to reduce the high probability of conformations having no contacts or one contact at the initiatial stage of

folding. The equilibrium probability of the native conformation is 0.002 at T=500 K, and 0.837 at T = 300 K using ε = - 2.3 RT. The equilibrium probability of the energetically nearest conformation making eight native contacts instead of nine, contact I at chain terminus being disrupted is 0.086. Thus the total equilibrium probability of the two conformations account for more than 92 per cent of the observed molecules at equilibrium.

Master equation formalism allows us to explore folding processes occurring at different time scales. Time steps Δt of different sizes can be conveniently used, depending on the time scale or the stochastic process of interest. Steps of $\Delta t = 0.01$ time units were used, for example, for examining the initial folding processes in the 9-mers, while the later stages were examined with 4-5 orders of magnitude larger time steps, consistent with the observed distributions of frequencies (eigenvalues of **A**). A broader distribution spanning about five orders of magnitude is operated in the case of 16-mers. This way, it is possible to observe both the local structure formation and global folding processes in proteins, which have a large time scale difference.

3.3. Reducing the Size of the Transition Rate Matrix

In order to visualize folding mechanism, we analyze the results in terms of subsets of conformations called "macroconformations". The conformations are grouped into 13 subsets, according to their number and types of native contacts: Subset O comprises the conformations having no native contact; subsets A, B, C, D contain those having only one (A, B, C or D) native contact, as indicated by their name; AB, AC, BC, BD have two native contacts; ABC and BCD have three contacts, and finally ABCD is the native conformation (having four contacts).

The important question that is addressed is whether a reduced 13 x 13 model of macroconformations can accurately describe the kinetic process extracted from the 740 x 740 matrix of microconformations for 9-mer. The reduced model is much faster to compute. Reducing the size of **P**(t) by one order of magnitude indeed increases the computational efficiency by two orders of magnitude, as the computation time scales with N^2. This enables us to analyze the longer chain models, which are computationally very

expensive. The experimental results are based on ensemble averages, so the results of the reduced model are comparable with the experimental results.

A reduced 13 x 13 transition matrix **A'** is constructed, describing the rates of passages between the macroconformations. The element of **A'** accounting for the passage from macroconformation B to A, for example, is found by double summing the elements A_{ij} of **A** over $1 \leq i \leq N_A$ and $1 \leq j \leq N_B$, similarly to Equation (3.6). Calculations were repeated for the reduced (13-d) probability array. The time evolutions of the native structure and the native contacts identically reproduced the results obtained from the 740-d analysis.

The calculations for the 16-mer is carried out in such a reduced space of 257 macroconformation, which includes all possible distributions of native contacts except for ten having ≤ 1 native contacts. Table 3.1 shows the total number of macroconformations and microconformations for structures having the same number of native contacts, whereas Table 3.2 presents the most important macroconformations and the corresponding number of microconformations.

Table 3.1. The total number of microconformations (W_{mic}) and macroconformations (W_{mac}) for the conformations having the same number of native contacts (m)

m	W_{mic}	W_{mac}
1	258453	9
2	81992	35
3	21024	68
4	3889	76
5	522	50
6	94	20
7	14	6
8	1	1

The rms deviation $\langle(\Delta r_{ij})^2\rangle^{1/2}$ between macroconformations i and j is found on the basis of the average radii of gyration of the conformations belonging to the two macroconformations. This approximation is tested with the 9-mers and verified to lead to insignificant changes in the observed results.

Table 3.2. The dominant macroconformations having different numbers of native contacs (m) and corresponding number (W_{mic}) of microconformations

	m=2	W_{mic}
1	*AB*	13004
2	*AC*	10391
3	*AG*	9545
4	*BC*	9291
5	*GH*	6207
6	*CG*	5264
7	*BG*	5189

	m=3	W_{mic}
1	*ABC*	1867
2	*GHI*	1558
3	*ABG*	1405
4	*AGH*	1241
5	*ACG*	1139
6	*BCG*	1025
7	*BCD*	691

	m=4	W_{mic}
1	*BCDE*	503
2	*AGHI*	315
3	*ABCG*	207
4	*ABGH*	181
5	*CGHI*	178
6	*BGHI*	172
7	*ACGH*	141

	m=5	W_{mic}
1	*ABCDE*	51
2	*ABGHI*	47
3	*ACGHI*	71
4	*BCDEG*	38
5	*BCGHI*	35
6	*ABCGH*	7
7	*ABCDG*	16

	m=6	W_{mic}
1	*ABCDEF*	39
2	*ABCGHI*	7
3	*BCDEHI*	5
4	*BCDGHI*	5
5	*CDEGHI*	4
6	*ABCDEG*	3
7	*ABCDEI*	3
8	*ABCDGI*	2
9	*ADEFGI*	1
10	*ADEGHI*	1
11	*CDEFHI*	1
12	*ACDEHI*	1
13	*ABDGHI*	1
14	*ABCDGH*	1
15	*DEFGHI*	1
16	*ACDEGH*	1
17	*CDEFGH*	1
18	*ADEFGH*	1
19	*ACEFHI*	1
20	*ACDFGI*	1

	m=7	W_{mic}
1	*BCDEGHI*	5
2	*ABCDEFG*	3
3	*CDEFGHI*	1
4	*ADEFGHI*	1
5	*ACEFGHI*	1
6	*ABCDEHI*	1

	m=8	W_{mic}
1	*ABCDEFGHI*	1

3.4. Dispersion and Shapes of Modes from Reduced Transition Rate Matrix

The analysis of the reduced transition rate matrix **A'** provides the additional advantage of visualizing the type and relative time scale of the individual modes of motion that contribute to the folding process. The eigenvalues λ_k' ($2 \leq k \leq 13$) of **A'** are representative of these modes' frequencies, and the associated eigenvectors describe the shapes of the individual modes. Thus the dispersion and shapes of individual modes associated with the transitions between macroconformations can help us to understand the folding process in more detail.

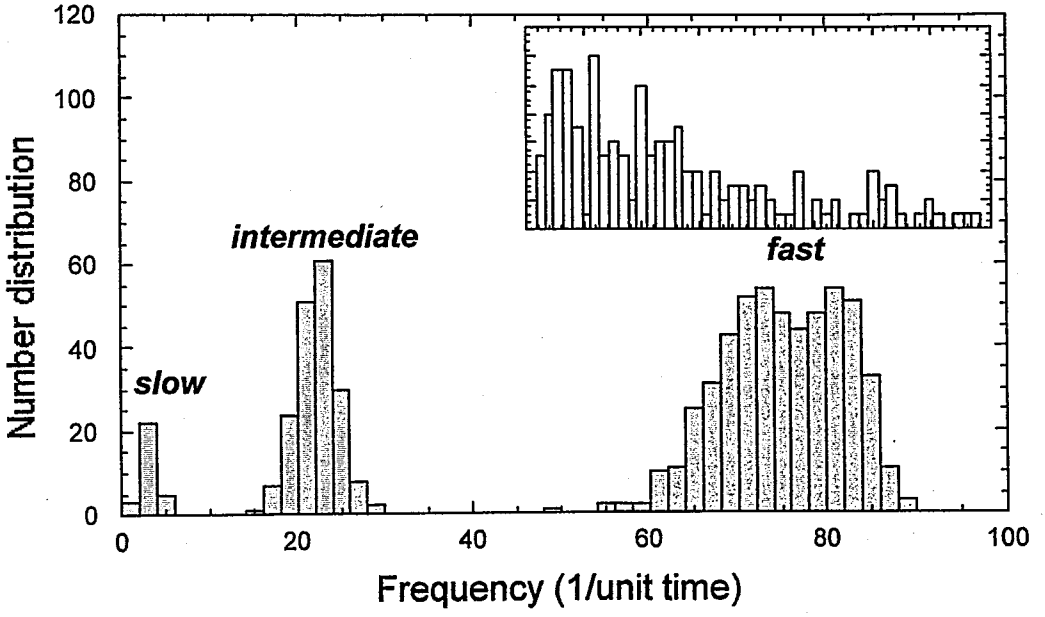

Figure 3.2. Eigenvalue distribution of the 9-mer and the 16-mer (inset)

The decomposition of the transition rate matrix A of the 9-mer gives a trimodal distribution that is presented in Figure 3.2. The trimodal distribution could be classified as: (i) a burst stage at $t < 0.03$ time units, approximately, (ii) intermediate times $0.03 \leq t < 2$ time units, and (iii) long times $t \geq 2$ time units. At the burst stage, only the fastest modes operate. This stage will be observed to correspond to a rapid decay of the conformations having no native contacts. At the other extreme case of long times, on the other hand, only one mode, the slowest, effectively contributes; there is an apparent single exponential accumulation of native conformation. At intermediates times, there is a superposition of

multiple modes giving rise to a more complex scheme with multiexponential time dependence.

The decomposition of the reduced transition rate matrix of 16-mer is smoother and broader, as shown in the inset of Figure 3.2. The three regimes are not distinctive anymore. Approximately five orders of magnitude difference 's observed in the time-scale of the fast and slow processes. This is consistent with the large time scale difference between *local* structure formation and *global* folding in proteins.

The eigenvalues $\lambda_{k'}$ ($2 \leq k \leq 13$) of **A'** (13 x 13) for the 9-mer represent the frequencies of the modes in the space of macroconformations, and the eigenvectors describe the transitions driven by that specific mode. There are 12 nonzero eigenvalues . Each element of a given eigenvector is associated with a given macroconformation, the latter being indexed from 1 (macroconformation *0*) to 13 (*ABCD*) The minima or maxima indicate the macroconformations with the highest activity (or transition probability).

The slow and fast modes of the 9-mer in the reduced space of transitions are found to differ by 4-5 orders of magnitude in their frequencies . The dispersion obtained for the 740 x 740 transitions on the other hand, varies over 2-3 orders of magnitude. Comparison of the two sets shows that the slowest modes (k = 2) have about the same frequencies (~ 0.28/unit time), whereas the fastest modes differ. This difference can be explained as follows: The fast transitions observed in the space of macroconformations reflect the cumulative contribution from the *multiple* transitions, or multiple pathways of relaxation, simultaneously operating at the initial stages of folding, in conformity with the energy landscape view of folding starting from an ensemble of denatured conformations. For example, subset *O* disappears - and subsets *C* and *D* form - by multiple mechanisms, via transitions between several conformations at the initial stage of the folding process, hence the apparent fast transitions (high frequencies) observed at short times in the space of macroconformations.

Figure 3.3 illustrates the shapes of a few modes (k = 2, 4, 7, 11-13). These are simply the eigenvectors plotted against macroconformation index. The extrema indicate the macroconformations that are most strongly influenced by the action of a given mode. Positive and negative values refer to changes in opposite direction, i.e. one

macroconformation being formed (or accumulated) while the other is disrupted (or depleted). The uppermost curve (k = 2) describes the slowest mode, and the lowermost (k = 13), the fastest. The latter reveals, for example, that the fastest mode decreases the population of subset *O*, while increasing those of subsets *D* and *C*. The second fastest mode (k = 12) describes the communication between subsets *C* and *D*. The third (k = 11) reveals the depletion of *C* and *D*, and concurrent accumulation of *A*, *B*, and *CD*. These three modes lie all in the fast transitions regime (Figure 3.2). The curve k = 7, on the other hand, reflects an intermediate time process, mainly an equilibration in favor of subset *CD* between all subsets of conformations involving two native contacts. Finally, the uppermost two curves reflect the increase in the population of conformations having three native contacts (k = 4) and the stabilization of the native structure at the expense of subsets *ABC* and *BCD* (k = 2).

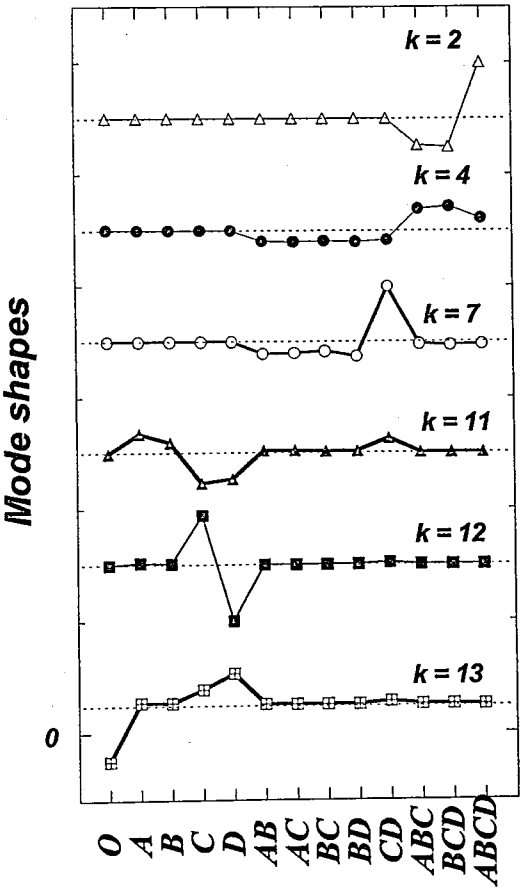

Figure 3.3. Modal shapes of eigenvectors

3.5. Effect of Energy Parameters

The change in the folding condition can make changes in the folding process. It may give rise for example, to the accumulation of some intermediates. In order to understand the effect of energy parameter ε, the time evolution of the macroconformations *CD* and *BCD* and the native (N) conformations of the 9-mer were computed with respect to different ε values.

The result is illustrated in Figure 3.4. Part (a) shows the accumulation of macroconformation *CD* of 9-mer at short times, which is diminished at higher temperatures or lower ε values. This subset of conformations is stabilized at short times. Its probability reaches the value of 0.39 during the burst stage of folding kinetics using ε/RT = -5, which is significantly higher than its original and equilibrium populations. Such a transient accumulation might be attributed to being trapped in a local minimum along the folding pathway. The escape from this subset is easier and faster at higher temperatures, as expected. Likewise, the subset *BCD* accumulates in the intermediate time regime, as shown in part (b), and decreasing ε values give rise to an earlier and weaker tendency to accumulate. Part (c) presents the gradual stabilization of the native state. The equlibrium probability of the native state increases with increasing the energy parameter ε.

The general conclusion deduced from the plot is that the qualitative features of the folding kinetics do not vary with the folding conditions. The intermediates that are accumulated during the folding process still exist even though their population is decreased or their accumulation time is shifted. The native conformation still shows an apparent two-state kinetics. However, it should be investigated whether the change in attractive potential of a given native contact speeds up or slows down the folding process. This will be discussed in Chapter 4.

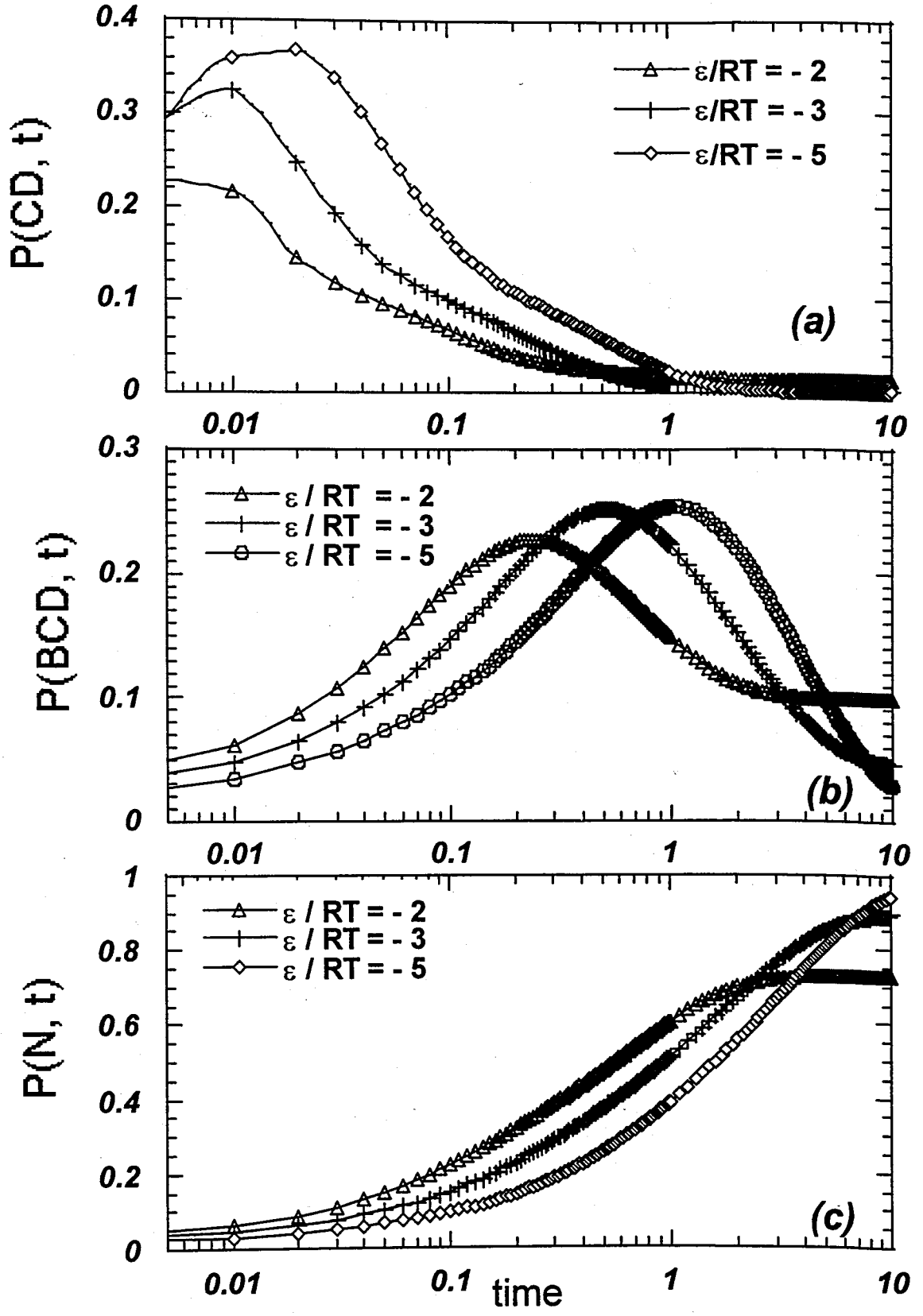

Figure 3.4. Time evolution of conformations with respect to different energy parameter

4. RESULTS AND DISCUSSION

4.1. Time Evolution of Native Contacts

The time evolution of native contacts can give us insights as to the sequence of event during the folding process. The time evolutions of native contacts are computed for the three different native structures of the 9-mers, that are presented in Figure 4.1 The ordinates represent the time-dependent probabilities of contacts, P(X, t), where X designates a given contact (*A, B, C* or *D*).

A general feature emerging from the examination of the curves in Figure 4.1 is that the *innermost* local native contacts stabilize first, succeeded by the local contacts occurring on relatively exterior regions. This qualitative observation can be consolidated by the characteristic times $\tau(X)$ for the *stabilization* of each contact found from the integral

$$\tau(X) = \int [P(X, t) - P(X, \infty)]/[P(X, 0) - P(X, \infty)]\, dt \qquad (4.1)$$

over $0 < t < \infty$. The results are indicated on the figure for each type of contact.

The following simple rules can be deduced:

(i) The contact stabilized the fastest is invariably a contact made by the central site of the 3 x3 lattice, in each of the three examined cases. This site may be viewed as the *core* of the structure. Thus, core contacts exhibit the strongest tendency to be stabilized first, amongst all native contacts, irrespective of the sequential position of the core residues. The relatively fast formations of contact *C* in case (a), *E* and *H* in (b), and *L* in (c) show the tendency for the core residues to be stabilized first.

(ii) Among the different native contacts that the core residue can make, the most *local* ones - i.e. those involving the closest monomers along the sequence - are the ones that are most likely to form first. For example *C* is preferred over *B* in case (a), *L* is preferred over *K* and *J* in case (c). In particular, a hierarchical characteristic time

$\tau(L) < \tau(K) < \tau(J)$ is observed in the latter case, confirming the increase in time with sequential distance.

It is interesting to note that in part (a) the contact *D* shows a tendency to form rapidly, at the burst stage of folding. This indicates the high propensity of local interactions at chain termini at short times. However, early formation of this contact does not ensure that it is stabilized and conducive to the native state. At intermediate times *D* competes with *C*, and the latter eventually turns out to be the most rapidly stabilized contact: $\tau(C)$ is indeed about twice shorter than $\tau(D)$. We note that even the contact *B* is stabilized faster than *D*.

This observation suggests the following third rule:

(iii) Helical contacts at chain termini, despite having a tendency to form at early stages of folding, can be rapidly reverted (or opened) such that their effective folding (or stabilization) time is longer than that of inner helical contacts. β-strand or interdomain contacts, on the other hand, accumulate steadily, and may eventually exhibit a shorter characteristic time compared to the reversible helical contacts. The recent stopped-flow kinetic studies of β-lactoglobulin also indicate that substantially more helices are formed at early times than is present in the final native state (Baldwin, 2001). This is in consistency with our results that the helical contacts at chain termini are easy to form but have marginal stability.

It is of interest to see if the rules (i)-(iii) deduced above from the 9-mers are also applicable in the case of the 16-mers. Our analysis confirms the same hierarchical pattern for the 16-mer structure (d) shown in Figure 3.1. The time evolution of native contacts is shown in Figure 4.2. The long-time behavior is displayed for clarity, while the inset shows the short-time behavior. The calculated characteristic times obey the order $\tau(C) \approx \tau(G) < \tau(D) < \tau(I) < \tau(A) < \tau(H) < \tau(E) < \tau(B) < \tau(F)$.

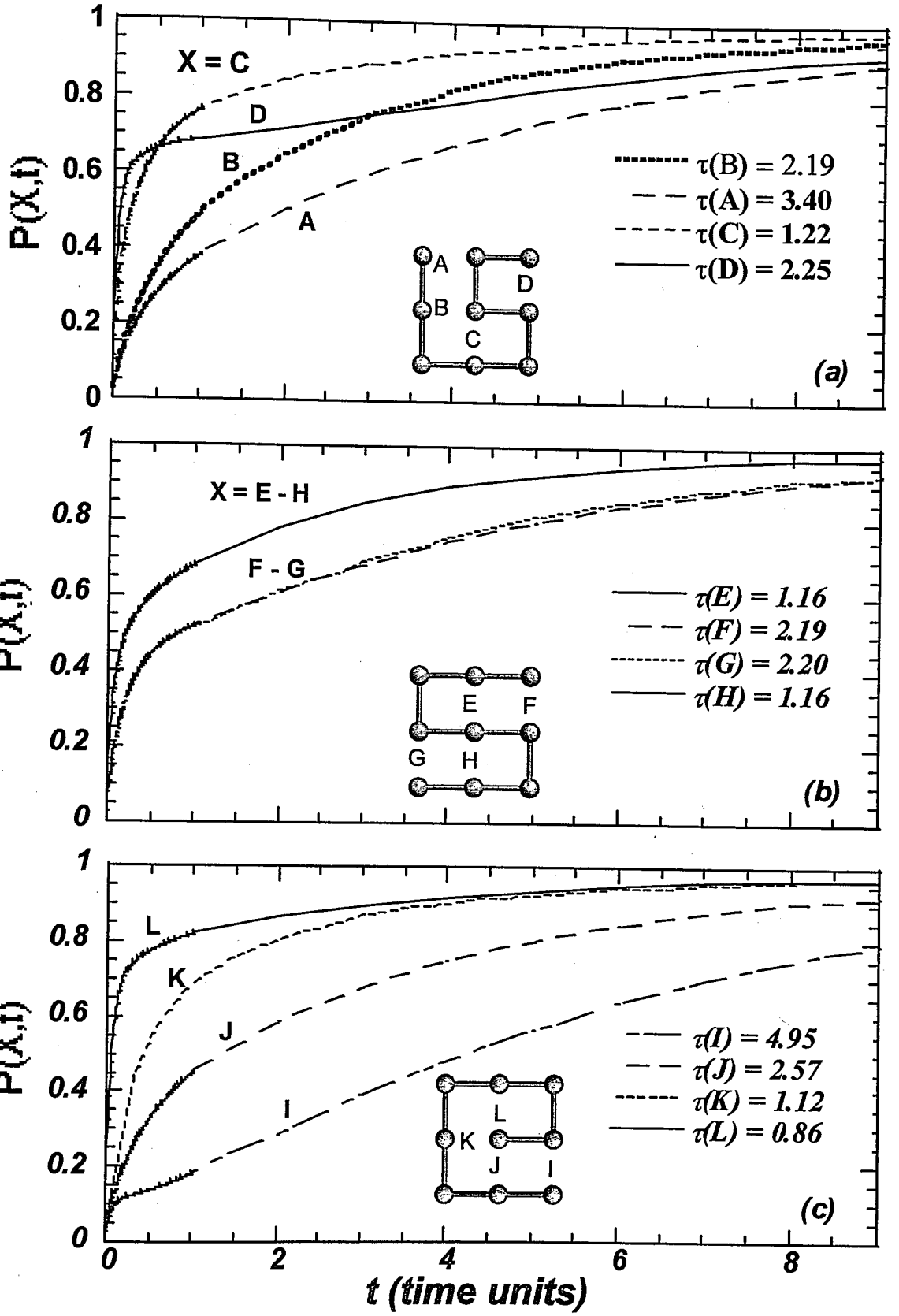

Figure 4.1. Time evolution of native contacts for three 9-meric native structures

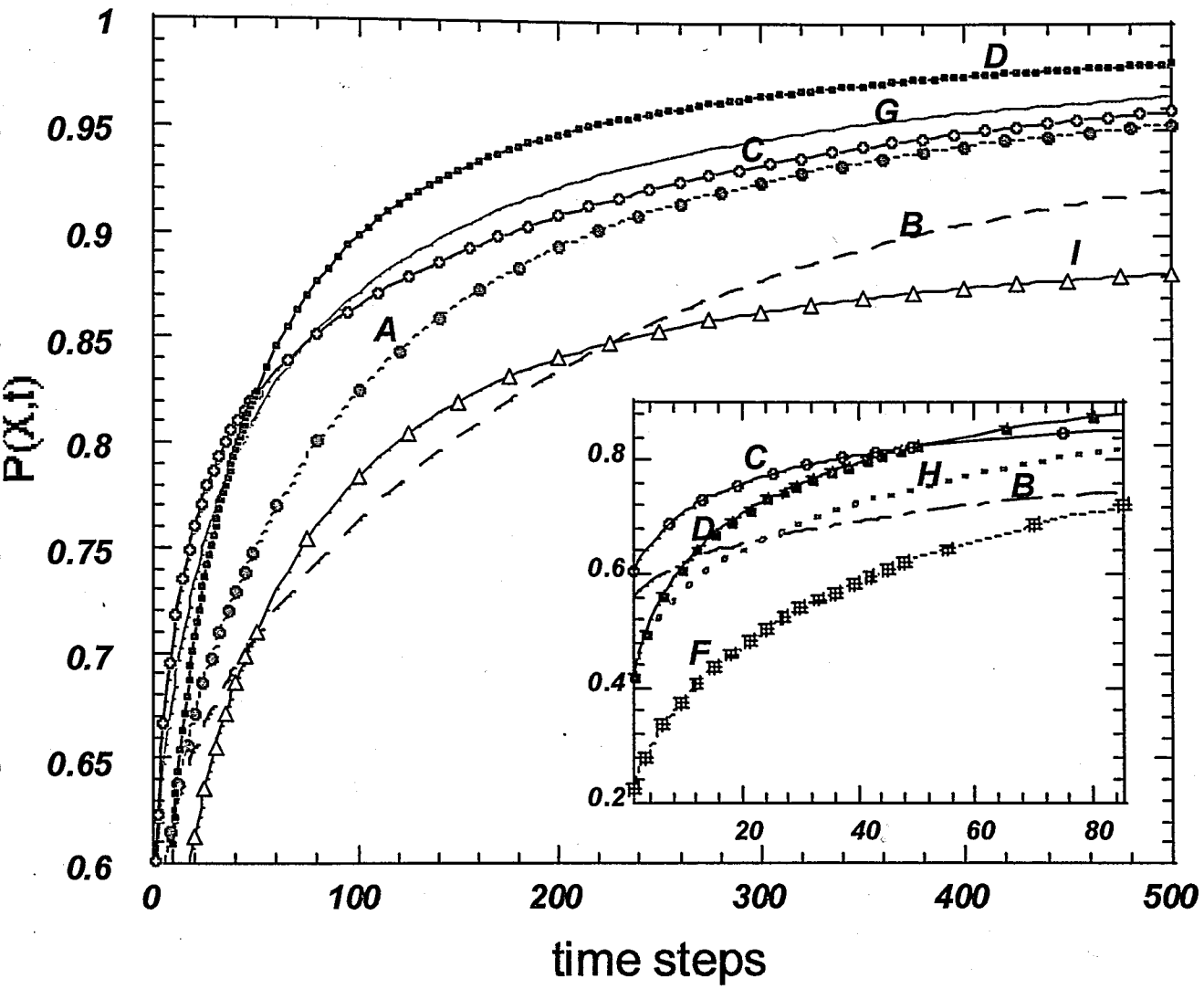

Figure 4.2. Time evolution of native contacts for 16-mer

The native contact that are stabilized first, *C* and *G*, are again contacts involving a core residue which is consistent with rule (i). Here core residues are those occupying the four central positions of the 4x4 lattice. Actually, six contacts (*C, D, E, A, H* and *G*) involve (at least) one core residue. Especially *C, G, A* are the most local ones *i.e.* between residues i and i+3. Thus, two of these most *local* contacts made by one or more *core* residues appear to be formed at the earliest folding stage, that obeys *rule II.* The third, A, is a local contact made at the chain terminus, which, in consistency with rule *III*, can be easily reversed, and as a result $\tau(A)$ is relatively long.

In the inset of Figure 4.2, the short-time behavior is enlarged. The curves for contacts *C* and *B* illustrate the high propensity of helical contacts at the burst stage (contact *A*, not shown, lies between *C* and *B)*. The interdomain contacts, on the other hand, are the least probable contacts at the burst stage, as illustrated by the limiting curve, while β-sheet contacts exhibit an intermediate behavior (see *H*). However, this order is soon reversed, because the respective rates of accumulation obey the opposite order: β-sheet contacts approach their equilibrium probabilities faster than the α-helical contacts, and more strikingly, the innermost interdomain contact *D* surpass both the α-helical and β-strand contacts, suggesting that the latter is rather stable or irreversible, once formed. The competition between the innermost α-, β- and interdomain contacts *C, G* and *D*, respectively, can also be discerned in the uppermost curves of the main figure. On the contrary, the surface exposed interdomain contact *F* is significantly more sluggish compared to all other contacts, suggesting that *F* is also the first contact to be disrupted upon unfolding. It is interesting to recall that the disruption of tertiary interactions between the helix and a two-stranded portion of the β-sheet was the primary unfolding event in the extensive MD study of CI2 unfolding by Lazaridis and Karplus (1997).

All these results lead to the conclusions that the folding process is a hierarchic process in which the folding begins with the formation of local contacts of marginal stability (Takada, 1999). Thus formation of *key* tertiary native contacts forms the extended folding nuclei (Fersht, 2000) and drives the system to the native state.

4.2. Coupling Between Native Contacts

For the 9-mer show in Figure 4.1 (a), contacts *C* and *D* are the most probable contacts at the burst stage of folding, while *A* and *B* form at the later stages. It is of interest to see which contact, *A* or *B*, is more readily driven by the original contacts *C* and *D*. Figure 4.3 displays the conditional probabilities of formation of a second native contact subject to the condition that native contacts *D* (part (a)) or *C* (part (b)) have already formed. In both cases, there is a strong driving potential for successive or concurrent formation of the contacts *C* and *D*. This is consistent with the scheme of Figure 3.3. However, contact *B* is unambiguously seen to be the next stabilized contact, in both cases, as also revealed in Figure 4.2. This supports the view that formation of contact *B* is the key

native contact for the transition state and stabilization of the native conformation *ABCD*. The formation of *B* immediately drives contact A. This was clearly seen from the conditional probability curves for the native contacts given that the contact *B* has formed (data not shown).

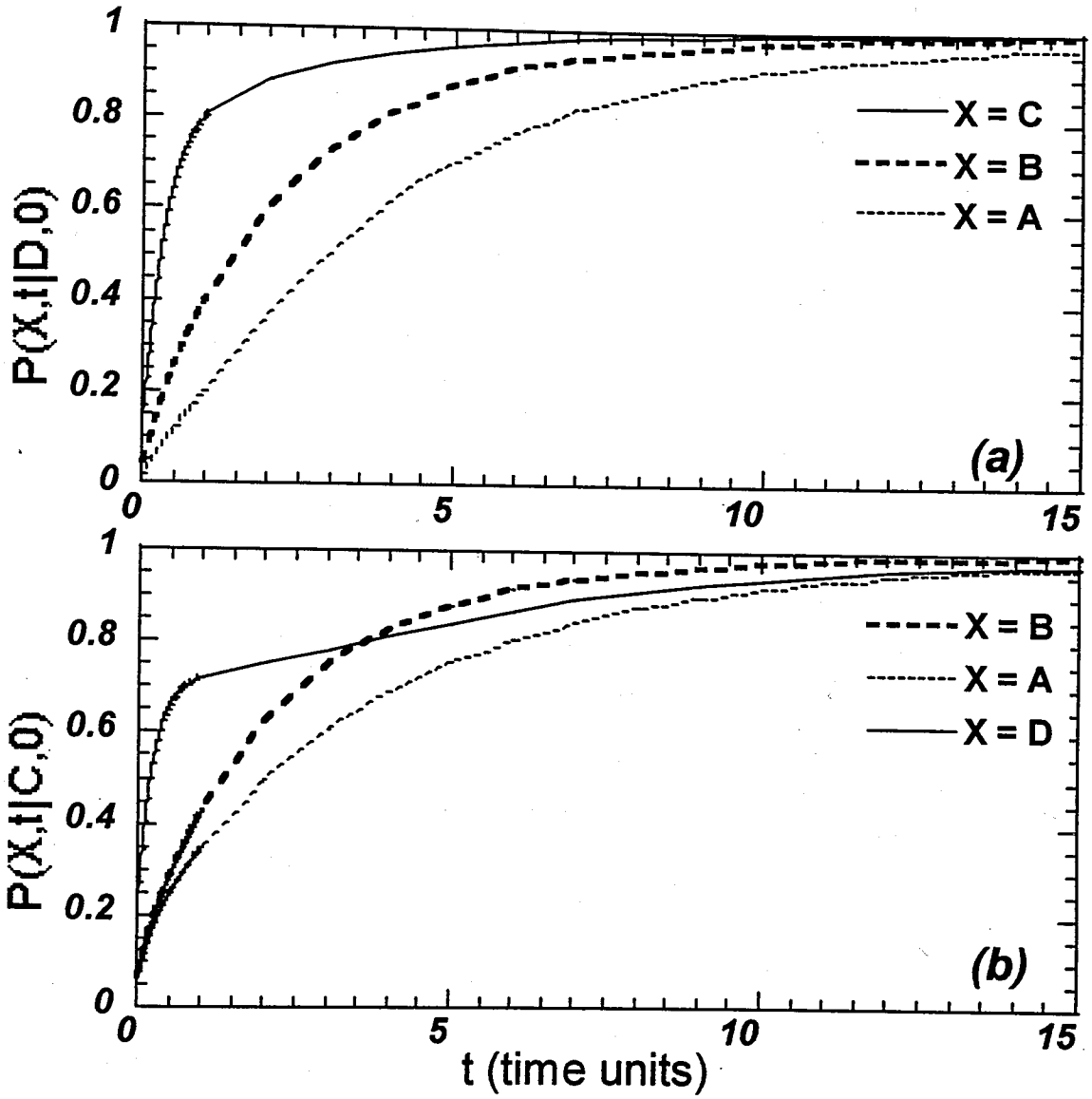

Figure 4.3. Conditional probability curves of the native contacts for 9-mer

The latest mutational study of Vendruscolo *et al.* (2001) support this view in which three key residues of the 98 residue acylphosphates form a critical contact network in the transition state.

Another question that should be answered is whether there is a set of native contacts that accelerate the formation of native conformation. For this purpose, the conditional probabilities of the native conformation, subject to the condition that some sets of native contacts have formed, are computed. Figure 4.4 present the conditional probability curves with the singlet probability of the native conformation for 16-mer. The structure of the native conformation is shown in the inset of Figure 4.4.

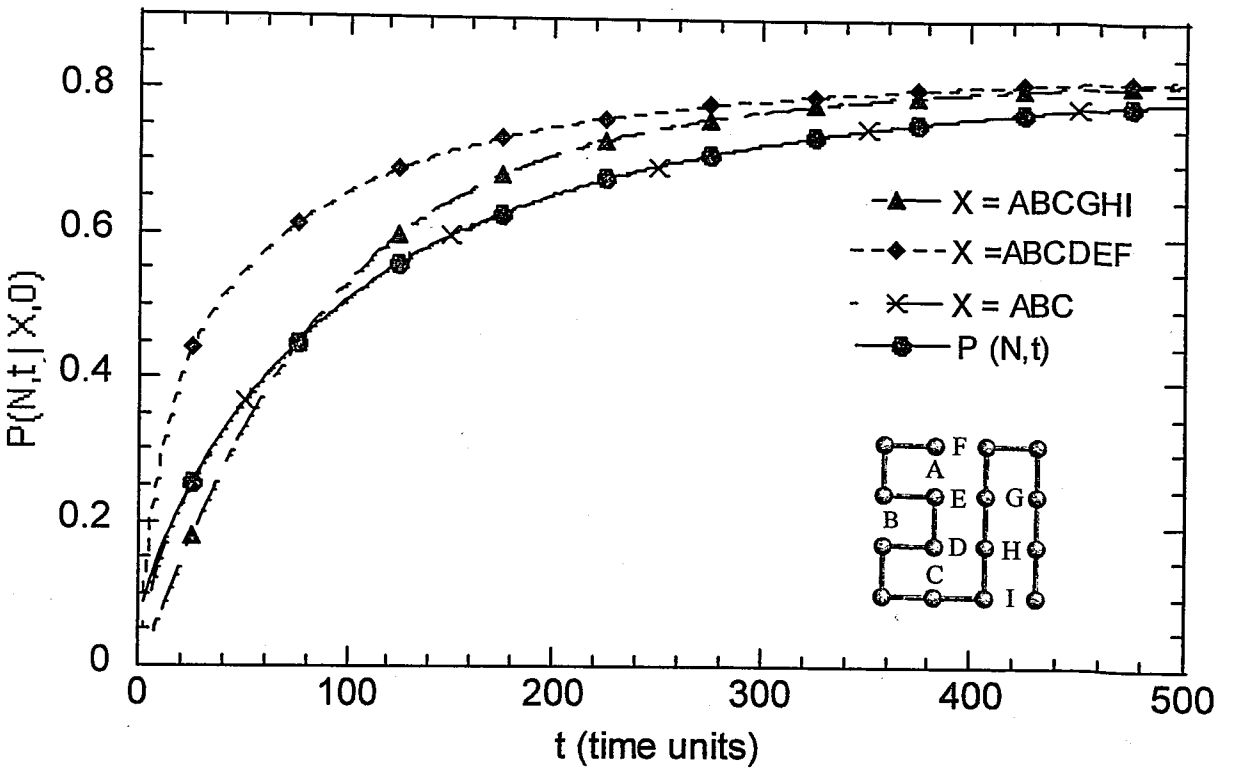

Figure 4.4. Comparision of the conditional probability curve of the native conformation with its singlet probability curve

The conditional probability curve of the native conformation, when the helical contacts *ABC* have formed a t = 0 is exactly the same as its singlet probability curve. The conditional probability curve subject to the condition that native contacts of *GHI* (β-strand) have formed also shows the same trend. The data is not shown for clarity. This result indicates that the earlier formation of secondary structural native contacts does not necessarily speed up the folding process. On the other hand, the two other conditional probability curves for the condition that the two different sets of six native contacts have already formed, are faster than the singlet probability curve of the native conformation.

Comparing the two conditional probability curves, it is seen that the formation of the helical and intra domain (tertiary) contacts drives folding faster than the formation of the native contacts of α-helix and β-strand. One can ask if these six native contacts are a transition state structure/nucleus for the folding of this model protein. Klimov and Thirumalai (2000) defined the folding nucleus as a set of native contacts that (i) include a minimal number of stable contacts, and (ii) result in rapid assembly of the native conformation. Fersht defines the folding nucleus as a native like structure being composed of partly or well formed secondary structures that are stabilized by tertiary interactions (Fersht, 2000). The presence and identity of folding nuclei in the presently investigated 16-mer will be further explained below.

4.3. Dominant Folding Pathway

4.3.1. Fluxes between Macroconformations

A protein quickly and reliably finds its native structure, an exponentially small region of its total phase space. The issue that has been addressed here is whether a protein folds via a unique pathway, or it folds through an ensemble of pathways. In a strict sense, there cannot be a single pathway by which a protein folds. If an ensemble of denatured proteins all must pass through a single narrow pathway in their phase space, then there must be a large reduction in entropy upon entering this path. This step would consequently be very unlikely and rate limiting. It is much more likely that proteins fold via many different pathways. Such a mechanism would allow analysis of protein folding dynamics through general equilibrium and non-equilibrium statistical mechanics.

In this picture, the transition state should be composed of a broad ensemble of structures rather than one particular structure. This does not mean that the transition state is completely random. The transition state may be characterized by a partial structure in the form of stable pieces of secondary structure or partially correct backbone shape (Brooks *et al.*, 1998).

In the present work, a simple model is explored, which shows a 2-state kinetics, in the sense that there are two macrostates, denatured and fully folded, predominantly

populate the ensemble of conformations. The folding process exhibits an apparent single exponential time evolution that can be deduced from the time evolution of the native conformation (P(N,t)) in Figure 4.4. The model chains follow a broad ensemble of micropaths during the folding process and the joint probabilities (P(X, t_1; Y, t_2) of the macrofonmations at various times are computed to visualize the folding pathways. These probabilities in a way reflect the fluxes, or communications, between the states and the diagonal terms (P(X,t_1;X,t_2)) reveal the important states during the folding process. By this way, it is possible to see the dominant states that accumulate during folding. The analysis demonstrates that there is a *dominant* macropathway, even though each chain may explore a large ensemble of possible microscopic routes. The folding macropath can be described in terms of a particular sequence of events in which local interactions generally precede more nonlocal contacts. Figure 4.5 parts (a)-(f) indicate the following time evolutions (see also the schematic representations next to the joint probability map): (a) Originally, the ensemble predominantly consists of conformations having no contact (subset *O*). The initial probability of this subset is P(*O*, 0) = 0.71. By the end of the time interval $0 \leq t < 0.1$ the distribution is changed in favor of subset *D* and (to a lower extent) subset C, at the expense of subset *O*. The flux from subset *O* into subset *D* is revealed by the off-diagonal red region. (b) The subset *D* remains as the most highly populated subset, while some exchange between subsets *C* and *D* is observable from the off-diagonal elements. More importantly, subset *CD* emerges as the first subset of conformations having two native contacts (yellow spot). The subset *CD* apparently grows as an extension of subset D. (c) There is a gradual increase in the population of subset *CD*. (d) *CD* is the dominant subset while the conformations having one native contact (subset *D*) practically disappear. Interestingly, the native conformation *ABCD* starts to be stabilized, although *ABC* and *BCD* are not discernible. Thus, contacts *A* and *B* form almost simultaneously. The extreme cooperativity between contacts *A* and *B* suggests that the formation of one of them immediately drives the other, and at the same time the complete folding into the native structure. Although which contact initiates this cooperative mechanism cannot be discerned in this scheme, the analysis of coupled transitions (Figure 4.3) indicates that B precedes A, as also suggested in Figure 4.2 (a). The formation of contact *B* may thus be viewed as a critical step that precipitates overall folding. (e) A new equilibrium is reached,

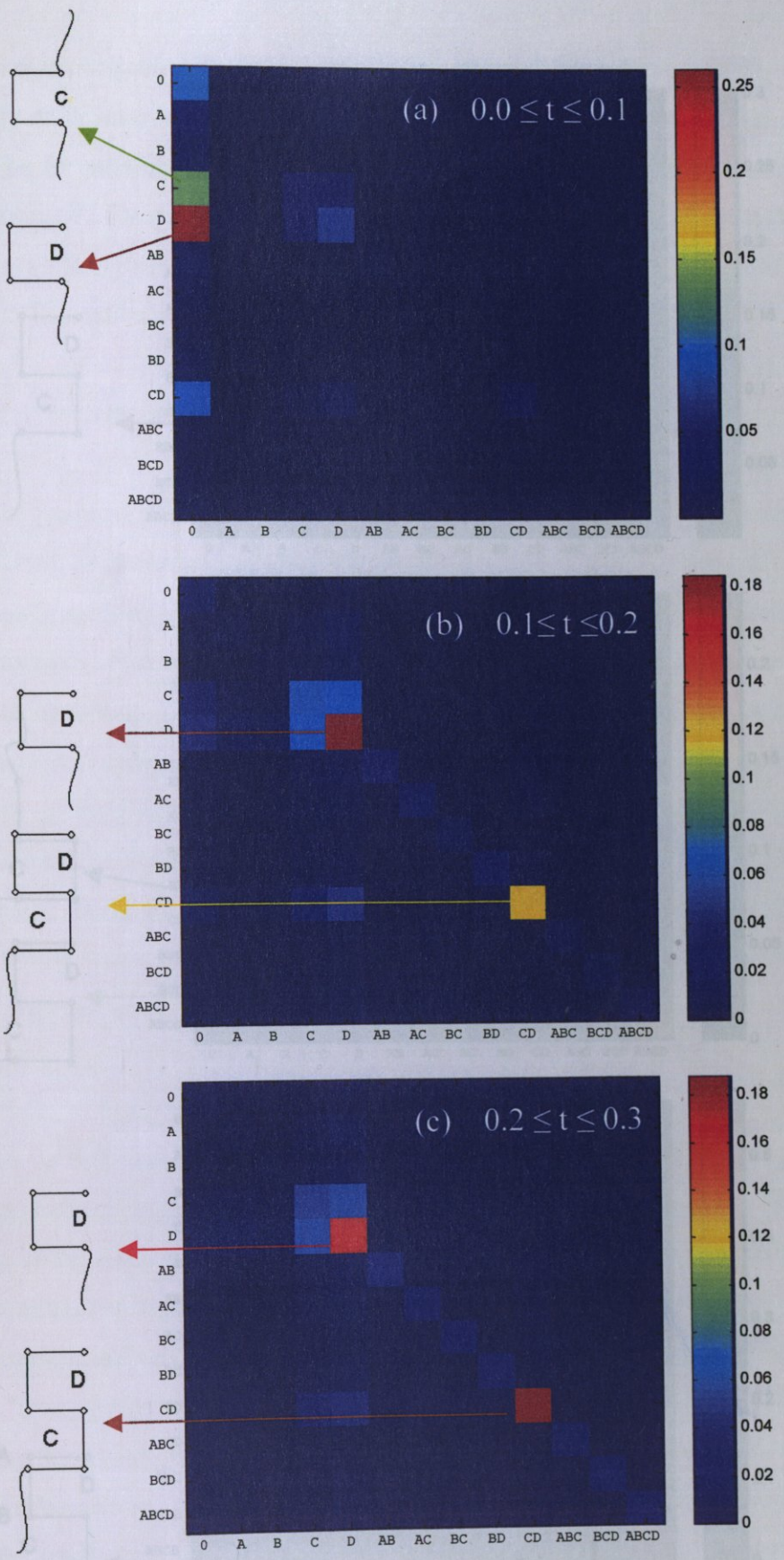

Figure 4.5. Joint probabilities of macroconformations at various stages of folding and their schematic representation (a)-(c)

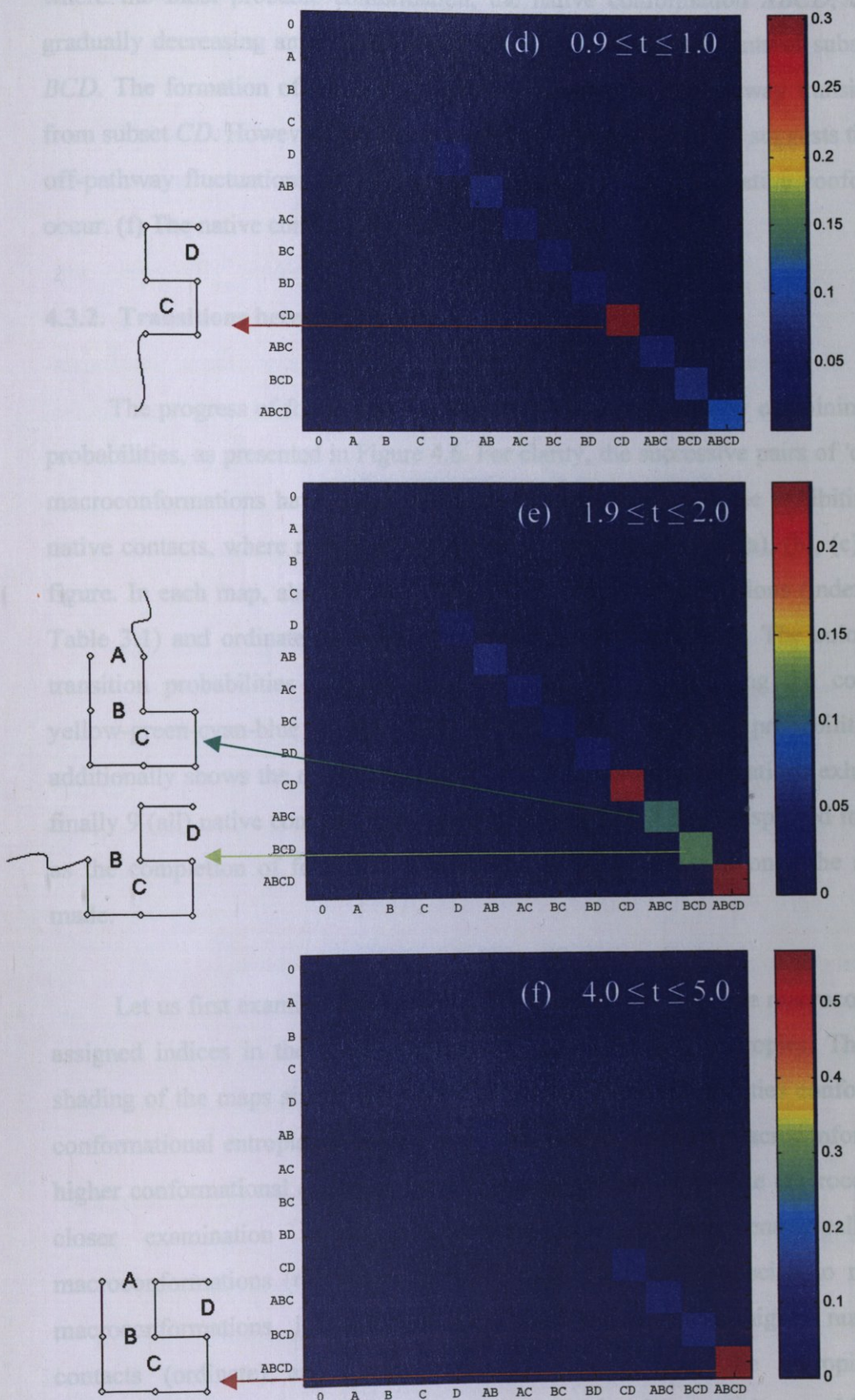

Figure 4.5. Joint probabilities of macroconformations at various stages of folding and their schematic representation(d)-(f)

where the most probable conformation, the native conformation *ABCD*, coexists with gradually decreasing amounts of subset *CD*, and increasing amounts of subsets *ABC* and *BCD*. The formation of subset BCD can be attributed to on-pathway transitions starting from subset *CD*. However, the concurrent formation of subset *ABC* suggests thatoccasional off-pathway fluctuations (loss of terminal contacts) away from native conformation also occur. (f) The native conformation is equilibrated.

4.3.2. Transitions between Macroconformations

The progress of folding for the 16-mer has been observed by examining the transition probabilities, as presented in Figure 4.6. For clarity, the successive pairs of 'communicating' macroconformations have been considered in each map, i.e. those exhibiting m and m+1 native contacts, where m = 2, 3, 4 and 5 in the respective maps (a), (b), (c) and (d) of the figure. In each map, abscissa refers to the initial macroconformations (indexed as listed in Table 3.1) and ordinate represents the final macroconformations. The colors describe the transition probabilities between these macroconformations, using the code red-orange-yellow-green-cyan-blue in the order of decreasing transition probabilities. Map (d) additionally shows the results for the transitions to macroconformations exhibiting 7, 8 and finally 9 (all) native contacts. The transitions beyond m > 6 are displayed in the same map, as the completion of folding is almost spontaneously achieved once the sixth contact is made.

Let us first examine the maps shown in parts (a) and (b). The macroconformations are assigned indices in the order of decreasing conformational entropies. The smooth color shading of the maps simply indicates that the transition probabilities conform closely with conformational entropies: The transitions are in favor of those macroconformations having higher conformational entropy especially among the less nativelike macroconformations. A closer examination further reveals that most of the entropically unfavorable macroconformations (right portions of the maps) are not conducive to more native-like macroconformations, i.e. the new macroconformations with higher number of native contacts (ordinate) are usually produced starting from the entropically favorable macroconformations. The dominance of conformational entropy is unambiguously seen in maps (a) and (b). In part (b) a greater screening effect is discerned compared to part (a), i.e.

the number of macroconformations that evolve into more native-like macroconformations is reduced, as indicated by the broadening of the blue and thinning of the red regions.

The accessible transitions (red regions) are confined to an even smaller number of macroconformations in the case of the passage to macroconformations having 5 native contacts, displayed in part (c). In the same map, some relatively probable transitions that do not necessarily conform with the order dictated by conformational entropy can be distinguished. For example, careful examination of the rows in the map (c) reveals that transitions to macroconformations 2, 3 and 5 are favored. These are macroconformations in which the β-domain is fully structured, and two helical contacts (out of 3) are made. On the other hand, the 9th column (or macroconformation *BCGH*) appears to be disposed to evolve into more native-like macroconformations. Therein the two core contacts C and G of α- and β-domains are formed, and each domain has one additional native contact.

Map (d) in Figure 4.6 illustrates the transitions of the macroconformations having five contacts to more native-like conformations. Although macroconformations having high conformational entropy are again favored, several macroconformations with relatively low entropy can be distinguished which efficiently fold into native-like structures. A specificity in the folding pathway, not necessarily dominated by conformational entropies, is thus observed at this stage.

These results clearly show that the conformational entropy plays an important role in the folding mechanism. The transitions involving macroconformations with the highest number of conformational entropy are more probable, since there are many direct and indirect kinetic microroutes to those macroconformations.

The present analysis reveals that:

(i) the demand to the lowest energy leads to macroconformations down to the funnel where the native state is,

(ii) certain pathways are more dominant due to their high conformational entropy.

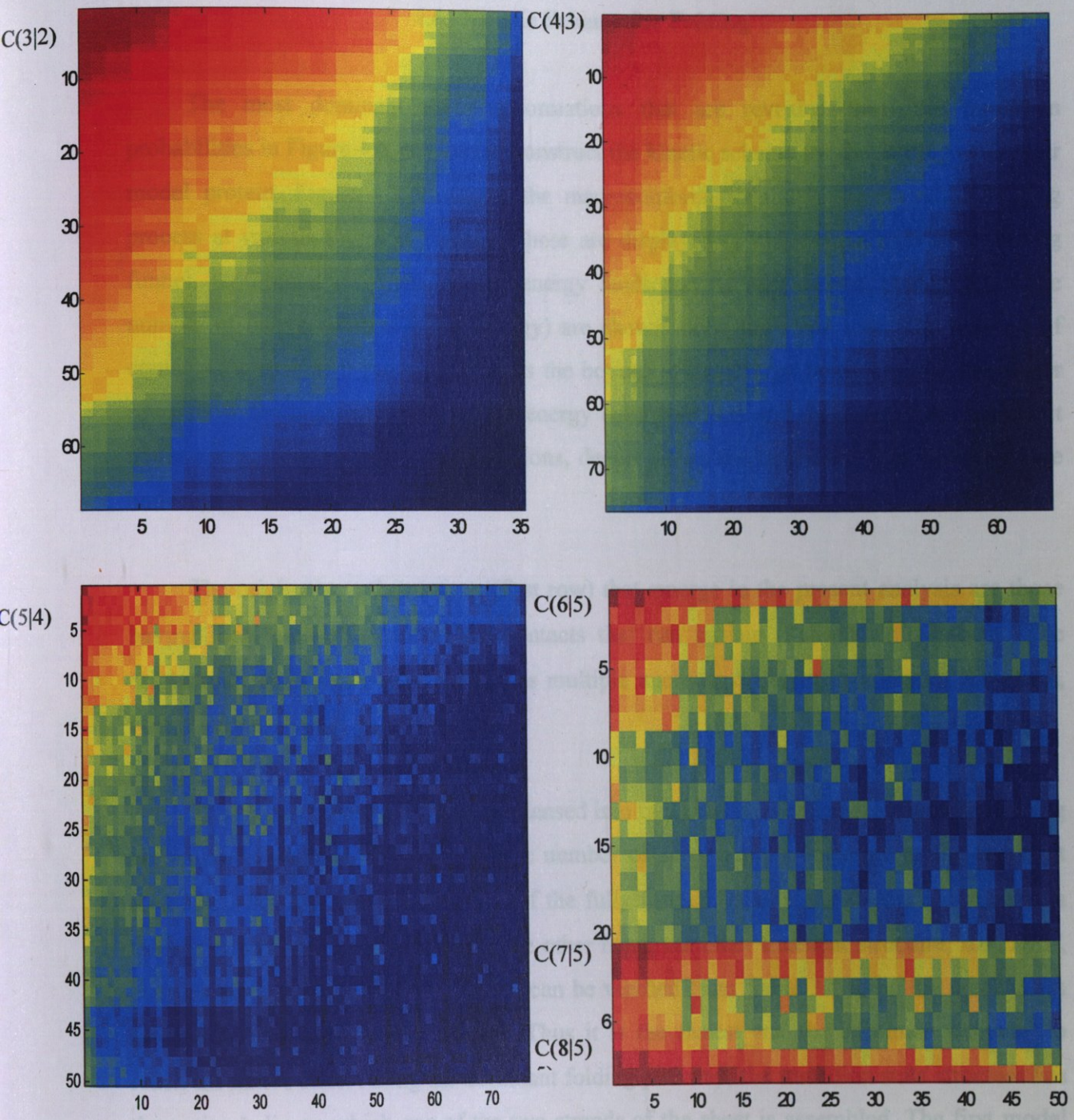

Figure 4.6. Transition between macroconformations

4.4. Kinetic Scheme for Folding

The most dominant macroconformations that are revealed from the transition probabilities in Figure 4.6, are used to construct the kinetic scheme for the folding of 16-mer model protein. Figure 4.7 illustrates the macroconformations that dominate the folding process at various stages of folding. These are displayed at different heights of a folding funnel, representative of the overall energy landscape. Conformations having the same number of native contacts (same energy) are shown along the same row. The number of contacts increase as it proceeds towards the bottom of the energy landscape. So, the upper rows show the high entropy and high energy conformations and the lower rows represent lower-energy lower-entropy conformations, deeper down the landscape. The bottom is the native (N) conformation.

The originating substructures (first row) that emerge in the present analysis are those having local contacts, such as the contacts that can initiate β-sheets or α-helices. The originating substructures can be seen as multiple nuclei at different locations of the chain, which initiate folding.

The macroconformations are condensed into two kinetic intermediate structures having six native contacts (fifth row). A large number of substructures converge to the transient structure *ABCGHI,* that is comprised of the fully formed α-helix and β-strand domains, in the absence of tertiary structure. On the other hand, the other transient structure, *ABCDEF*, evolves via a separate pathway, which can be verified from Table 3.1 to be the mechanism favored by conformational entropies. Thus it is the fastest and the dominant pathway to reach the native state. Along the dominant folding pathway, the transition state structure has the native helix on which one of the two strands of the sheet is assembled. The first model resembles the diffusion collision model of Karplus and Weaver (Karplus and Weaver, 1976) and the latter reminds the nucleation collapse model of Fersht (Fersht, 1995; Fersht, 2000).

These two intermediates, and the succeeding two macroconformations having seven contacts, exhibited a slight tendency to accumulate before complete folding, indicated by the peaks observed in their time evolution curves at relatively long times (Figure 4.8).

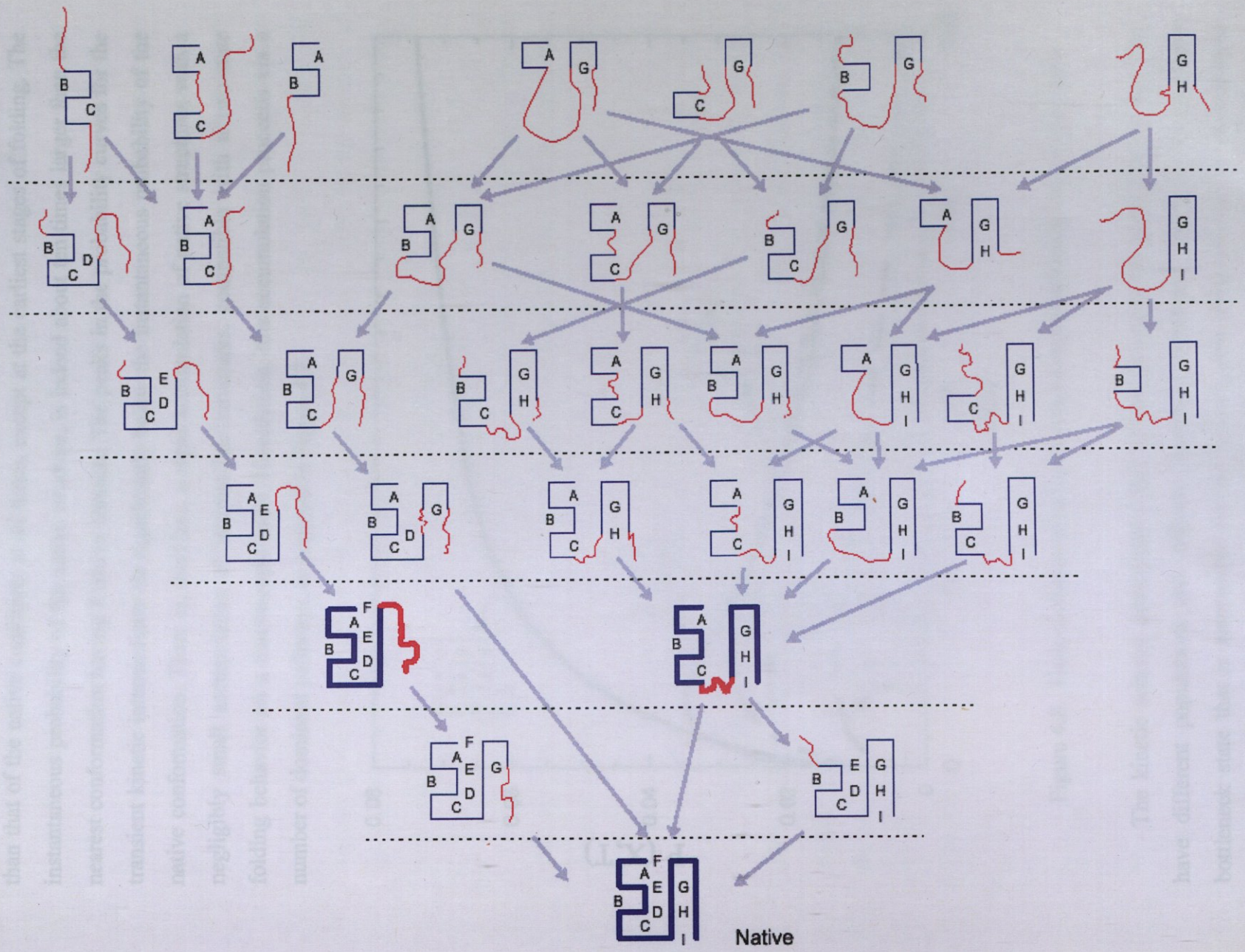

Figure 4.7. Kinetic scheme for folding

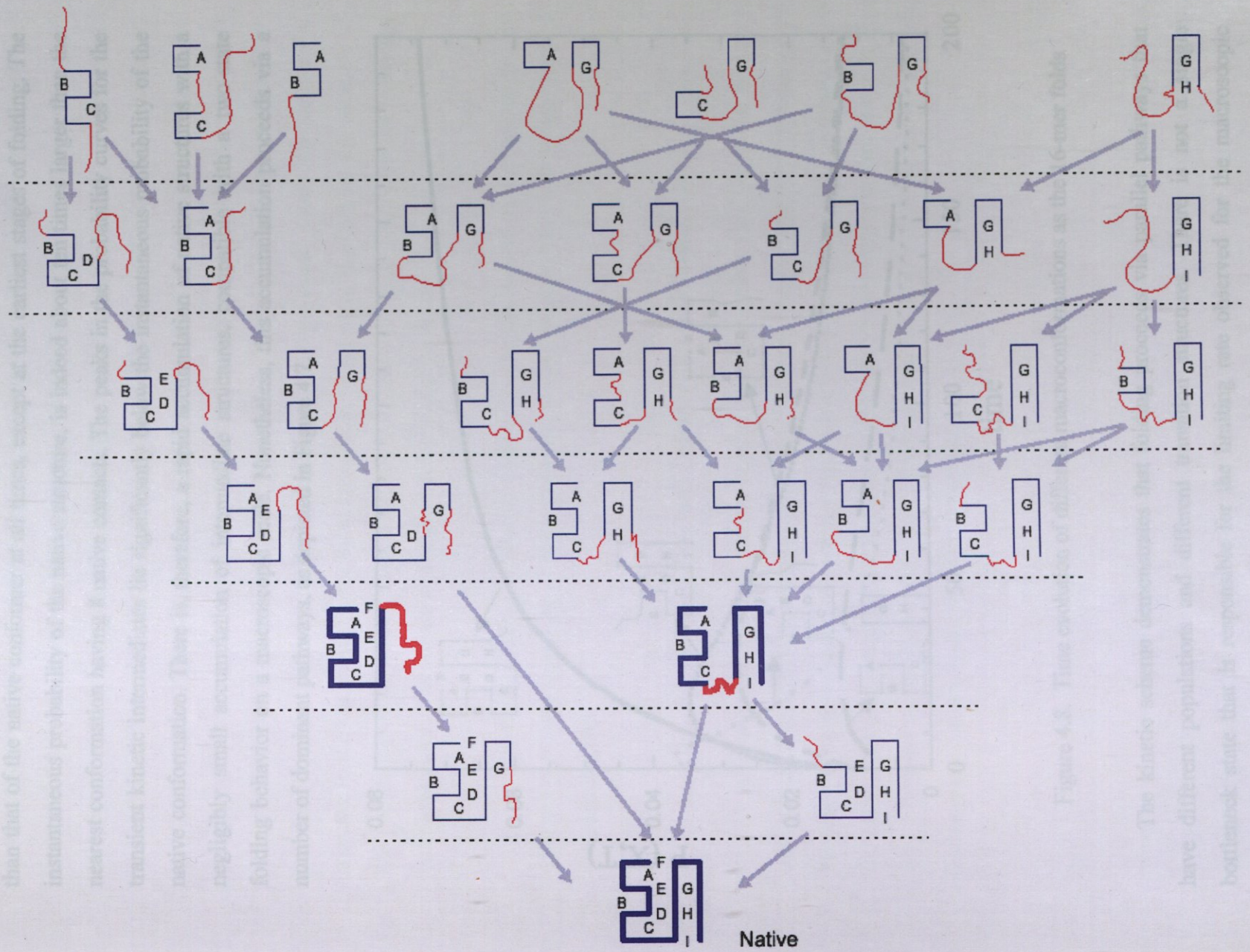

Figure 4.7. Kinetic scheme for folding

Yet, it is worth noting that the population of all conformations is significantly lower than that of the native conformer at all times, except at the earliest stages of folding. The instantaneous probability of the native structure, is indeed about ten times larger than the nearest conformation having 8 native contacts. The peaks in the probability curves for the transient kinetic intermediates lie significantly below the instantaneous probability of the native conformation. There is, therefore, a rapid accumulation of native structures with a negligibly small accumulation of intermediate structures, compatible with a two-state folding behavior on a macroscopic scale. Nonetheless, this accumulation proceeds via a number of dominant pathways, as depicted in Figure 4.7.

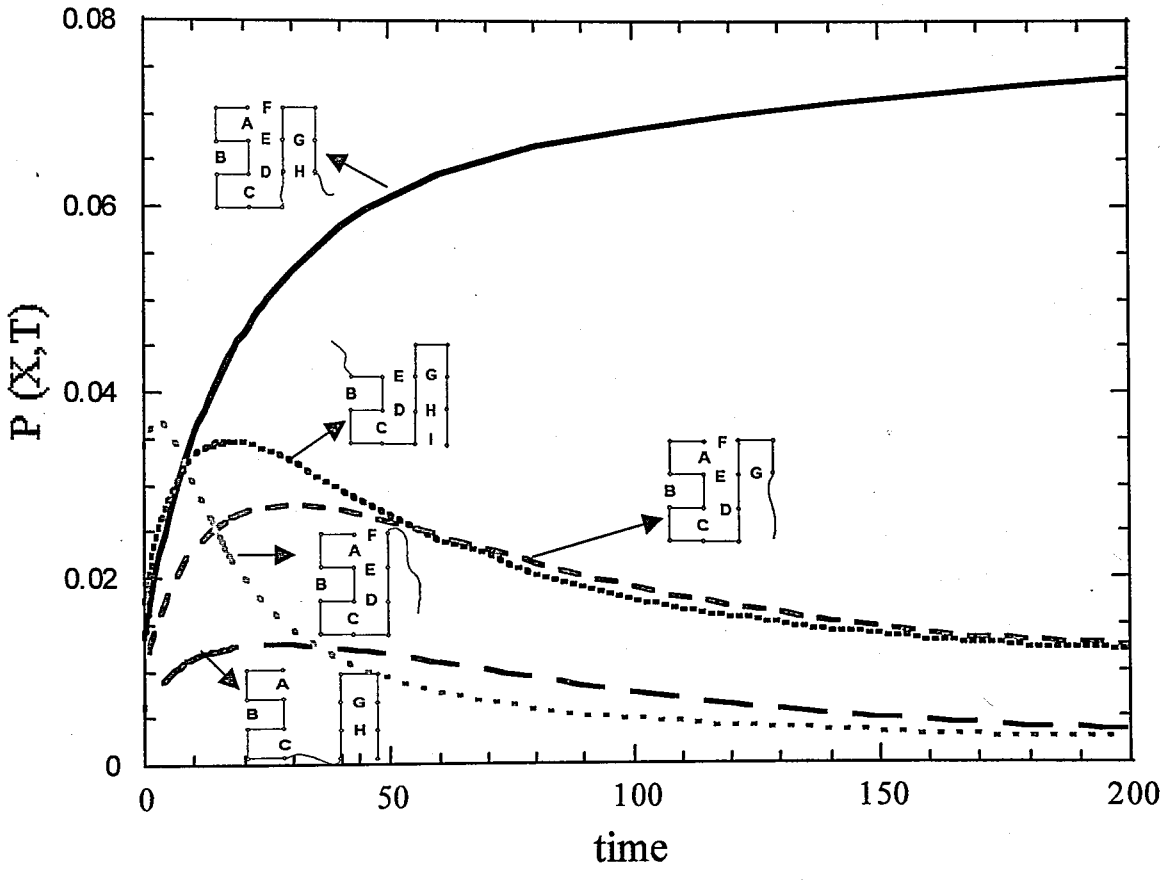

Figure 4.8. Time evolution of different macroconformations as the 16-mer folds

The kinetic scheme demonstrates that folding proceeds via parallel pathways that have different populations and different transition structures. There is not a single bottleneck state that is responsible for the limiting rate observed for the macroscopic process like in the classical theory. Thus the folding time to reach the native state is neither

the same as the folding time of the longest pathway nor the same as the shortest pathway. The experimental example of parallel flows involving a rapid helix formation and a slower β-sheet formation is lysozyme (Matagne *et al.*, 1997; Matagne. *et al.*, 1998). A variety of techniques, including quenched flow, hydrogen exchange labelling, stopped-flow absorbance and circular dichroism, has been used to investigate the refolding kinetics of hen egg lysozyme over a temperature range from 2 °C to 50 °C and at all temperatures, the fast (about 25 per cent) and slow (about 75 per cent) population of refolding is observed. It is found that the rate of formation of lysozyme depends on the microscopic rate constant and the population of the α-domain intermediates (Matagne *et al.*, 2000). The substantial increase in the rate constant due to an increase in temperature is offset by the decrease in the population of the intermediate. Thus, the population of TS structures on the dominant pathways plays an important role in the folding rate of proteins.

The important question is whether altering the populations of conformations on the different pathways speeds up or slows down the folding process. This will be analyzed in the next chapters.

4.5. Effect of Average Contact Order on Folding Time

It is known that it takes a shorter time to reach the native state when the folding is initiated from a nativelike macroconformation. However, the average contact order of the starting macroconformation can change the folding time(τ_N). Here the folding time to reach the native state was computed starting from a series of different macroconformations at time $\tau = 0$, in order to investigate the effect of average contact order <CO> on the folding rate. The average contact order (<CO>) is defined as

$$< CO > = 1/m \sum_{j} \sum_{i} (j-i)\, \delta\,(R_{ij} - d) \qquad (4.2)$$

where m is the number of contacts in the particular macroconformations and, $\delta(R_{ij} - d)$ is the delta dirac function equal to one if $R_{ij} = d$ and zero otherwise. R_{ij} is the distance between monomers i and j, and d is the lattice spacing, Figure 4.9 presents the plot of the average contact order of the 257 macroconformations of 16-mer versus the folding time.

The correlation coefficient is 0.44. The weak linear correlation points that the folding time tends to increase with increasing contact order, although this tendency is rather weak. The macroconformations *ABC* (helix) and *GHI* (β-strand) exhibit relatively shorter folding times compared to some of macroconformations having five or six native contacts. This indicates that native contacts that are close in sequence bring the other contacts into spatial proximity.

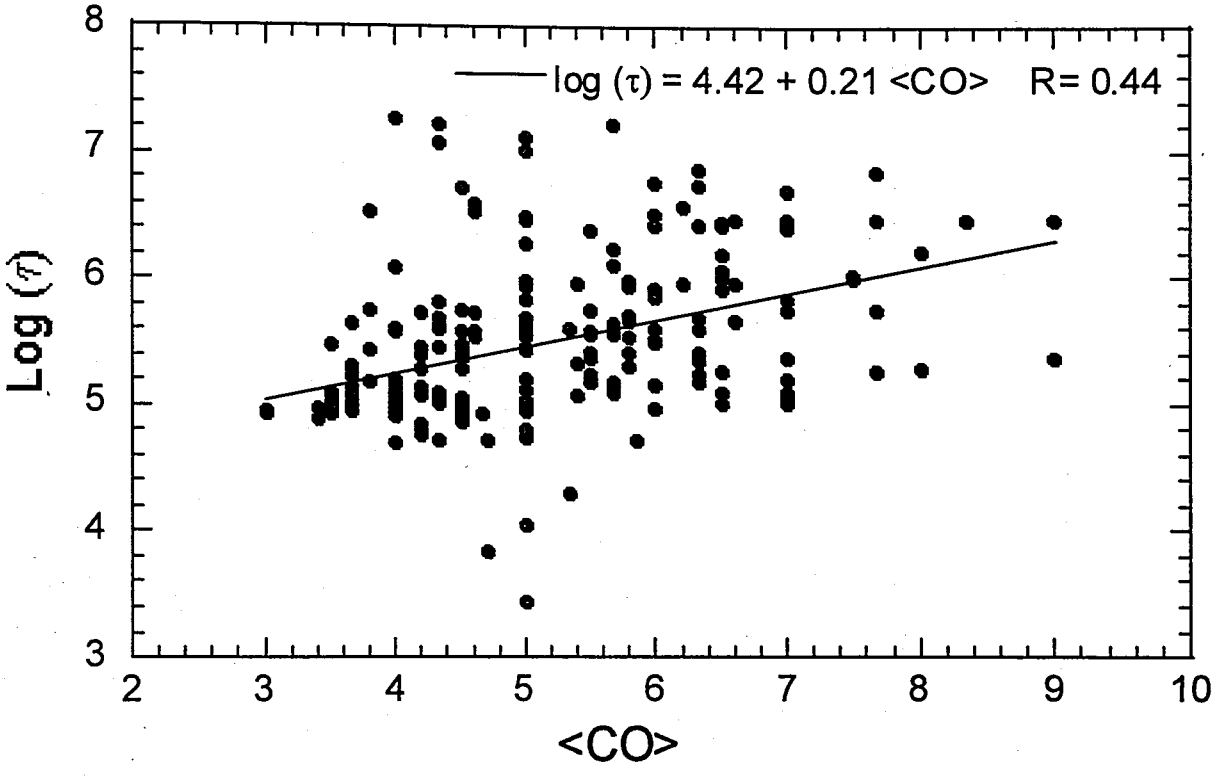

Figure 4.9. Correlation between the folding times and average contact order

4.6. Energy Landscape Mapping

A general computational experiment, called landscape mapping, is devised to identify the physical meaning of dominant pathways. If one could initiate a classical chemical reaction from any specific point along its reaction coordinate and measure the time required to reach the product from that point, it would give an unambiguous measure of reaction progress. By fixing molecular structures into specific conformations, then starting the reaction and measuring the time-to-product, one could map out the kinetic distances between conformations. This approach is applied to our folding model. At time

$\tau = 0$, folding is turned on with an ensemble of all the conformations having a particular set of native or non-native contacts already formed. Then the time (τ_N) required to reach the native state from that ensemble is computed. The degree to which a particular contact reduces τ_N defines the degree to which that contact is a folding nucleus.

Figure 4.10 shows τ_N, *for different starting points along specific pathways*. The times required for the passage between 'successive macroconformations' are calculated, and this successive times are summed up to find the folding time to reach the native state starting from that macroconformation and following those transitions. Two macroconformations are 'successive' if they differ by one additional contact only. So a point-to-point passage along a macropath is taken into account. For example, {GH} and {GHI} on folding scheme of Figure 4.7 are two successive macroconformations. The conditional probability of transition to {GHI} starting from {GH}, as a function of time is computed. The master equation formalism allows the calculation of all transition (or conditional) probabilities, between all pairs of macroconformations, as a function of time.

The resulting time evolution curve for each passage yields a characteristic time, say τ_{mn}, simply found from the best fitting single exponential. In principle, this time is comparable to $1/k_{mn}$. But it is *not* exactly equal to it, because the conditional probabilities include all direct and indirect passages between the initial and final states, even if the initial and final states are two successive (direct) states along the reaction scheme. Precisely, τ_{mn} can be approximated by a series, the leading term of which is $1/k_{mn}$. Suppose the characteristic time for the passage {GH} → {GHI} is designated as τ_2, the subscript referring to the number of native contacts in the original macroconformation (τ_2 is the difference between the ordinate values of {GH} and {GHI}). Likewise, one can obtain the times, τ_3, τ_4, τ_5,....., τ_8, for all the succeeding steps along the same macropath. The ordinate value for {CGHI} for example, is obtained by summing up τ_4, τ_5, τ_6, τ_7, and τ_8.

The sigmoidal shape of the curves indicates that the first few native contacts form relatively quickly, as do the last few contacts. The slowest transitions occur in the neighbourhood of the rate limiting conformations. These can be viewed as the slowest steps along the two pathways.

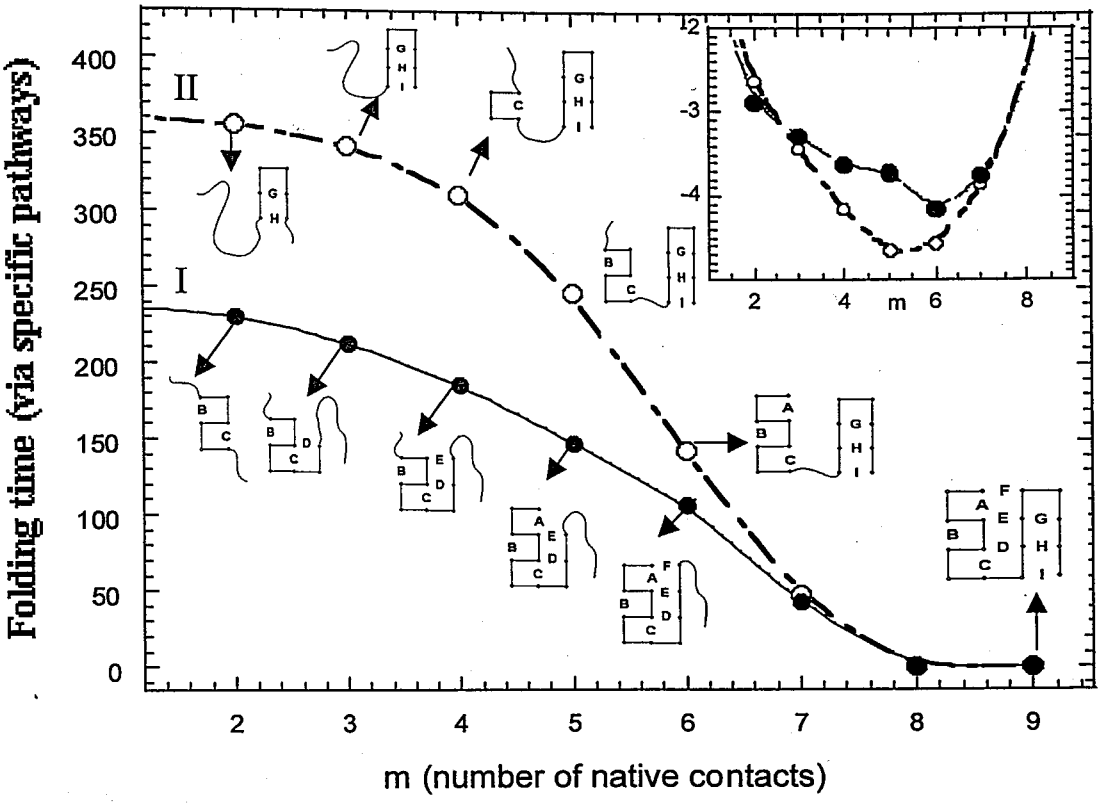

Figure 4.10. Folding time to reach the native state via specific macropathways

Consider also the elapsed time τ_m for each consecutive step m → m+1. An effective stepwise macroroute rate constant is evaluated, $k_m = 1/\tau_m$. A stepwise activation energy can be defined as $E_{act,\,m} = -RT \ln k_m$ (see the inset of Fig.4 10). A minimum in Figure 4.10 identifies the slowest macrosteps. The rate limiting macrostates along channels I and II are those having m = 6 and 5 native contacts, respectively. The route shown in Fig. 4.10 indeed represents a single 'macropath', but there are multiple micropaths contributing to this macropath, and the rate at each step of the macropath is directly proportional to the number *N* of micropaths involved in that step. *N* is the product of the number of microscopic conformations in the original macroconformation with that in the succeeding macroconformation.

The overall folding rate is faster than any individual micropath because folding occurs along many individual micropaths in parallel. While Figure 4.10 shows the elapsed time along a specific macropath, Figure 4.11 presents the rate to reach the native state from a particular macrostate, now summed over all routes.

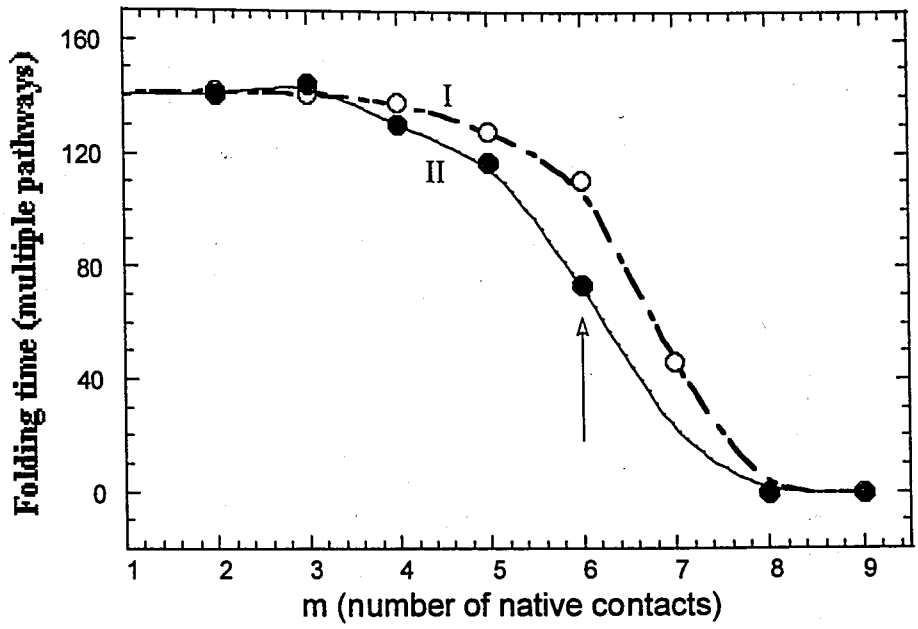

Figure 4.11. Folding time via multiple macropathways

This is achieved by computing the characteristic time to reach the native state when the folding is initiated from that specific macroconformation considering the all the transitions to all other macroconformations during folding. It is observed that: the time required to fold from open conformations ($m \leq 4$) is nearly independent of the conformation since there are many routes downhill at the top of the funnel and starting deeper on the funnel commits the flow to fewer and more specific routes.

These two different perspectives show that there is a kinetic barrier along the macroroutes, even though the landscape of microroutes is funnel-like. This results from a balance of two effects:

(i) The microscopic transition rates increase monotonically down the landscape,

(ii) The number of microroutes diminishes down the landscape.

The product of these two factors leads to an apparent barrier. This barrier results from a property of the landscape, not a property of a trajectory. The barrier is due to a reduction in density of routes, not an energetic problem along any one microroute.

The analysis has revealed a broad heterogeneity: often the native state is reached faster by some particular sets of two native contacts than by other particular sets of six

The analysis has revealed a broad heterogeneity: often the native state is reached faster by some particular sets of two native contacts than by other particular sets of six native contacts. The data can be rationalized using two ideas. First, as noted by many previous investigators (Fersht, 1997; Klimov and Thirumalai, 1998; Pande *et al.*, 1998; Dokholoyan *et al.*, 2000; Galzitskaya and Finkelstein, 1999; Fersht, 2000; Vendruscolo *et al.*, 2001), there are folding nuclei, i.e. certain sets of contacts that are relatively close to the native structure kinetically. Second, the sequences of folding events are zippers (Fiebig *et al.*, 1993, Fiebig *et al.*, 1993) the most local contacts form first, on average, and the least local contacts form later. The starting points that are kinetically closest to the native state are helical turns, or β-turns. The results are also consistent with the topology of the starting macroconformation on the folding rate that is presented in Figure 4.9. The statistical analysis of native contact formations for lattice model was performed by Tiana and Broglie (2001). They also found that the fast bonds are local bonds that form early in the folding process and nonlocal bonds form later involving the interactions with the amino acids already participating in the fast bonds.

4.7. Φ-value Analysis

4.7.1. Non-classical Φ-values

There have been many experimental and theoretical studies (Alexander *et al.*, 1992; Matthews, 1993; Kiefhaber, 1995; Kragelund *et al.*, 1995; Laurents and Baldwin, 1998; Englander, 2000) to identify the transition state structure (TS) which is considered as an ensemble of structures around a saddle point in an energy surface (Figure 4.12 (a)). The TS structures and intermediates can be analyzed by investigating the changes in the kinetics and equilibria of folding upon mutations. A key experimental strategy originated by Alan Fersht and his colleagues (Matouschek *et al.*, 1989; Fersht, 1995) and currently used by many theoreticians and experimentalists (Schindler *et al.*, 1995; Martinez and Serrano, 1999; Ternstorm *et al.*, 1999; Clementi *et al.*, 2000; Nymeyer *et al.*, 2000; Klimov and Thirumalai, 1998) is Φ-value analysis (Figure 4.12 (b)).

In Φ-value analysis, a particular amino acid in the protein is mutated. If the mutation destabilizes the protein by an amount $\Delta\Delta G$ (where $\Delta G = G_N - G_D$, N and D refer to native

and denatured states, respectively, and the first Δ refers to the change in stability that arises from the mutation), and if mutation changes the stability of the folding barrier by an amount $\Delta\Delta G^{\ddagger}$, the Φ-value is

$$\Phi = \frac{\Delta\Delta G^{\ddagger}}{\Delta\Delta G} \tag{4.3}$$

The interpretation of Φ-values derives from the Brønsted (β) theory and the Hammond postulate of classical chemical reactions which was explained in Chapter 2.3.4. The β and Φ are similar. Whereas β is used as an indication of bond formation or dissociation in the transition state. F- measures the non-bonded contacts formation and dissociation. Thus they are identical at the two extreme values of zero and one.

The rate constant for a conformational change can generally be described via Kramer's like equation,

$$k = k_o \exp[-\Delta G^{\ddagger}/RT] \tag{4.4}$$

where k_o depends on the reconfigurational diffusion coefficient and the geometric shape of the barrier. If the front term is insensitive to the specific amino acid sequence, then

$$\Delta G^{\ddagger} = - RT \ln (k_{mut}/k_{wt}) \tag{4.5}$$

where k_{mut}, and k_{wt} are the mutant and wild type folding rates, R is the gas constant and T is temperature. In the same manner, the change in the stability of a protein can be found as

$$\Delta\Delta G = -RT \ln (K_{mut}/K_{wt}) \tag{4.6}$$

where K_{mut}, K_{wt} are the mutant and wild type protein equilibrium constants of the folding curve.

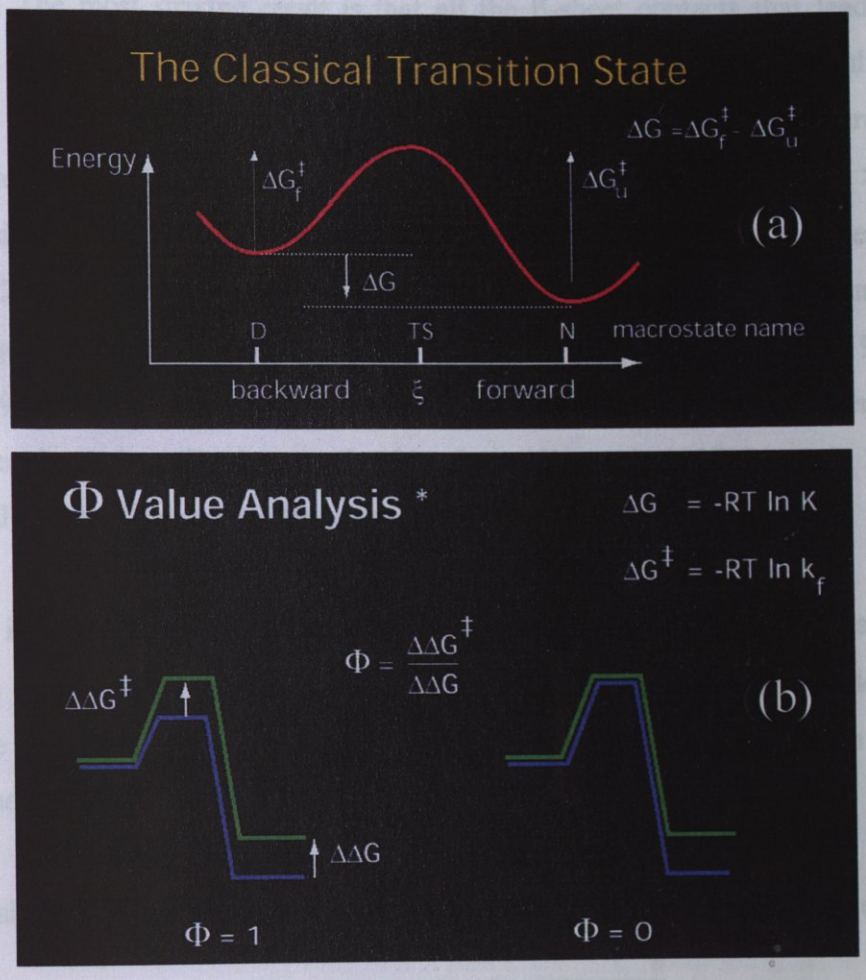

Figure 4.12 The schematic representation of classical reaction coordinate (a) and Φ-value analysis (b)

A Φ-value of 0 means that the mutation affects the energy of the transition state by the same amount that the denatured state is affected. Therefore at the mutation site, the TS structure resembles the denatured state. A Φ-value of 1 means that the free energies of the TS and native (N) states are equally affected by the mutation, which implies that the site of mutation assumes a native-like local conformation at the TS. According to the theory Φ-value should be between 0 and 1.

In this study, the native contacts are destabilized by reducing the corresponding attractive potentials by 30 per cent. The folding rates of wild type and mutants are found by the time evolution of native conformation, and the equilibrium constants of those are calculated using equilibrium probabilities. Table 4.1 shows the Φ-values of all native

contacts. The most striking result is that all the β-sheet contacts that lead to a slower pathway (the rate-limiting structure of partial or complete helix and the full sheet) have negative Φ-values (Figure 4.13 (a)). Within the framework of the theory, there is no model that currently explains Φ-values outside the normal range, i.e. $\Phi < 0$ or $\Phi > 1$. So when nonclassical values are observed, they are sometimes regarded as experimental artifacts. Yet, 10-20 per cent of the hundreds of measured Φ-values for protein folding are outside this range (Matouscheck *et al.*, 1992; Gay *et al*, 1994; Grantcharova. *et al.*, 1998; Martinez *et al.*, 1998; Nolting and Andert, 2000). Negative Φ -values have also been observed in computer simulations (Daggett *et al.*, 1996; Lazaridis *et al.*, 1997; Li *et al.*, 2000; Shea *et al.*, 2000), where they can be due to non-native contacts.

It is found that while classical Φ-values are restricted to systems having a single reaction coordinate, nonclassical Φ-values can arise from parallel coupled flows, for example in funnel-shaped energy landscapes (Baldwin, 1995; Bryngelson *et al.*, 1995; Chan and Dill, 1997).

Table 4.1. Φ-values resulting from a 30 percent destabilization of native contacts

Type of native contact	Φ-values
A	0.012
B	0.096
C	0.251
D	0.990
E	0.093
F	0.035
G	-0.357
H	-0.296
I	-0.085

Destabilizing the β-sheet contacts reduces the rate of β -sheet formation, causing a backing up and redirection of flow into the dominant pathway (I) where the rate-limiting step is when the native helix is formed and one of the two strands of the sheet is assembled. Since this pathway has a faster flow, the destabilizing mutation leads to an increase in the folding rate, hence a negative Φ-value. Figure 4.13 (b)-(c) presents the

schematic view of the wild type and mutant landscape and the backflow to the fast channel (I) upon destabilization of slowest pathway (I).

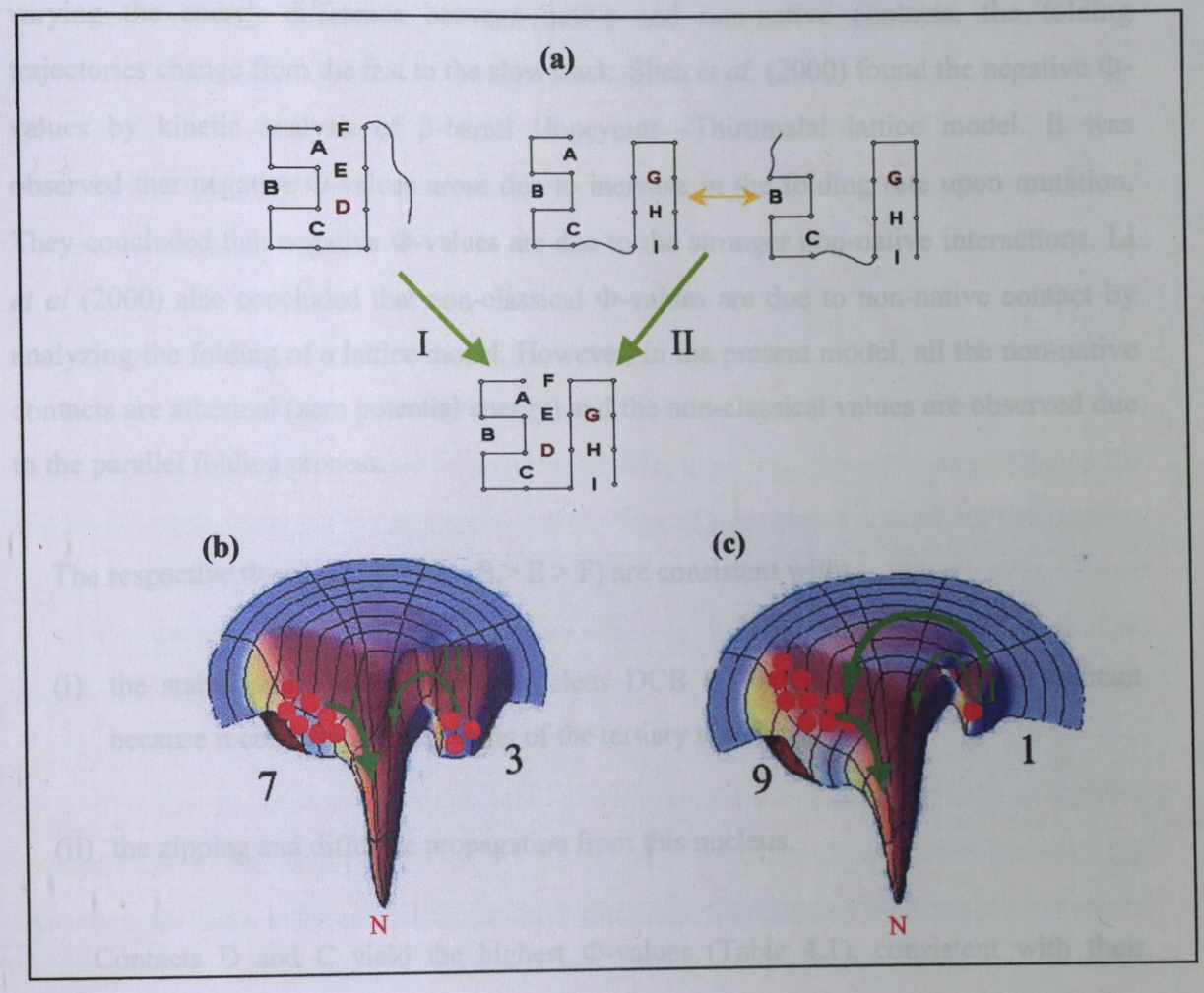

Figure 4.13. Schematic representation of Φ-values

Correspondingly, a mutation that destabilizes the helix (site D) blocks the fast pathway redirecting the flow into the slower pathway, decreases the overall folding rate. This mutation has $\Phi = 0.99$. It is interesting to note that the value of Φ depends also on the strength of the mutation: a stronger mutation, say reducing the attractive potential by 50 percent, at the same site gives $\Phi = 1.38$.

The increase in the folding rate was observed in the molecular dynamic (MD) study of Ladurner *et al.* (1998). In that study, it was found that the mutation on specific region of the single helix of CI2 speeds up the folding rate but it prevents the stabilizing of tertiary

interactions resulting in destabilization of the protein. Zhou and Karplus (1999) analyzed the folding kinetics of helical proteins using a three helix bundle like protein model. The MD results indicate that there are two main trajectories (fast and slow tracks) and by varying the energy difference between native and non-native contacts, the folding trajectories change from the fast to the slow track. Shea *et al.* (2000) found the negative Φ-values by kinetic analysis of β-barrel Honeycutt –Thirumalai lattice model. It was observed that negative Φ-values arose due to increase in the folding rate upon mutation. They concluded that negative Φ-values are due to the stronger non-native interactions. Li *et al* (2000) also concluded that non-classical Φ-values are due to non-native contact by analyzing the folding of a lattice model. However, in the present model, all the non-native contacts are athermal (zero potential energy) and the non-classical values are observed due to the parallel folding process.

The respective Φ-values (D > C > B > E > F) are consistent with:

(i) the stabilization of the folding nucleus DCB (D is the most important contact because it commits the beginning of the tertiary structure),

(ii) the zipping and diffusive propagation from this nucleus.

Contacts D and C yield the highest Φ-values (Table 4.1), consistent with their involvement in the folding nuclei. Their relative Φ-values are in accord with a zipping mechanism, starting from D, and propagating to C and E, and then to B. Contact A is relatively unimportant, its formation or dissociation being inconsequential for completion of folding, hence its relatively small Φ-value. The negative Φ-values for the sheet (G < H < I) result because G slows the overall folding by directing the flow into a slow folding channel.

Φ is often regarded as a "kinetic ruler" of the position along a reaction coordinate. But if folding landscapes have multiple microscopic reaction coordinates, then the interpretation of Φ -values cannot be this simple. For this purpose, the relation between the

average characteristic times of the native contacts and the Φ-values of those upon destabilization is investigated.

The time distribution function of each native contact is calculated by the time evolution of the native contacts presented in Figure 4.2. Then the average characteristic time $<\tau>$ is found for each native contact, Figure 4.14 (a) shows the plot of the Φ-values for each native contact versus the corresponding $<\tau>$. There is no observable correlation between $<\tau>$ and Φ-values.

In order to see whether Φ-values show the change in rate, it is decided to calculate the change in the average characteristic time of each native contact upon mutation. The average characteristic time of native contacts for wild type, $<\tau_{wt}>$ are already computed for Figure 4.14 (a). Then, the average characteristic time of each native contact for the mutant, $<\tau_{mut}>$ is computed by destabilizing the corresponding native contacts 30 per cent. Figure 4. 14 (b) presents the plot of $(<\tau_{wt}>/<\tau_{mut}>)$ values against the Φ-values. The correlation coefficient is 0.83.

The low correlation in Figure 4.14 (a) supports that Φ-values often do not give a kinetic ruler of the progress toward the native state: Φ-values do not correlate with τ_{wt}. However Φ-values indicate the changes in folding rate. Sites having high positive Φ-values indicate where mutations most strongly decelerate folding and high negative Φ-values indicate where mutations accelerate folding, by redirection into a faster channels or by destabilizing non-native contacts (Shea *et al.*, 2000).

Thus the Φ-value for a given contact measures the *change* in the effective rate of formation/stabilization of this particular contact during the folding process, rather than the hierarchical formation of this contact along the reaction coordinate (Figure 4.14).

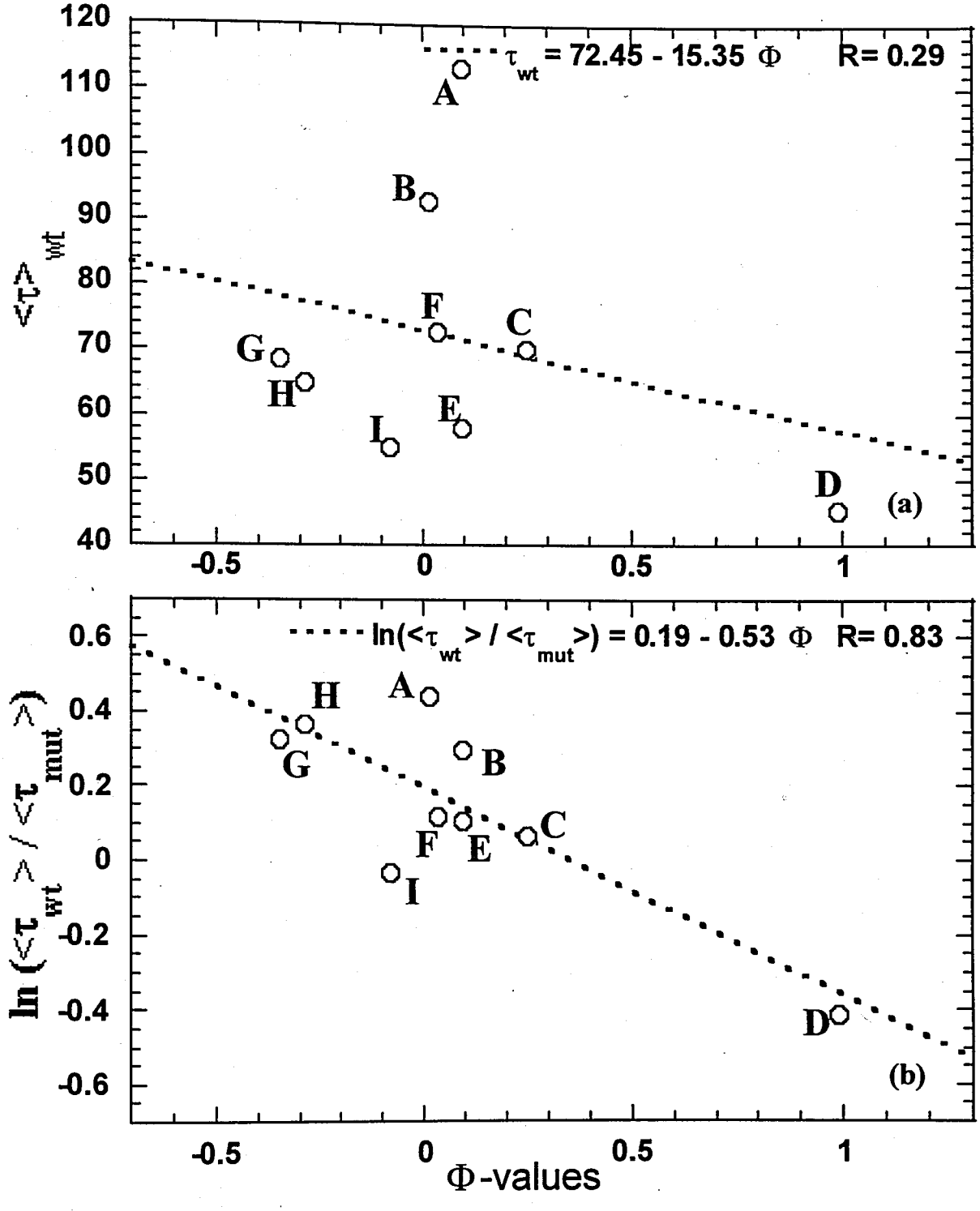

Figure 4.14. Correlation between Φ-values and $\langle\tau_{wt}\rangle$ for marked native contacts (a), Correlation between Φ-values and the change in average characteristic times (b)

4.7.2. Effect of Double Mutations

It is possible to detect the interaction of two native contacts by measuring the double mutant cycle consisting of the wild type protein model, the two single mutations and the double mutant (Horrovitz and Fersht, 1992). The pairwise coupling energy (Δ^2G_{int}) in the native state (N) relative to denatured state (D) can be defined by

$$\Delta^2G_{int(N\text{-}D)} = (G_N - G_D)_{wt} - \Sigma(G_N - G_D)_{single\ mutants} + (G_N - G_D)_{doublemutant} \quad (4.7)$$

The energy change between native and denatured states can be found using the equation 15. The pairwise coupling energies in the transition states are first calculated relative to folded state using the following equations:

$$\Delta^2G_{int\ (D\text{-}\ddagger)} = (G_D - G^{\ddagger})_{wt} - \Sigma(G_D - G^{\ddagger})_{single\ mutants} + (G_D - G^{\ddagger})_{doublemutant} \quad (4.8)$$

The coupling energies in the transition states are then found as

$$\Delta^2G_{int}^{\ddagger} = \Delta^2G_{int(N\text{-}D)} - \Delta^2G_{int(D\text{-}\ddagger)} \quad (4.9)$$

Table 4.2 shows the coupling energies for the native contact pairs where as Table 4.3 presents Φ-values that are obtained by destabilizing the two native contacts pairs at the same time.

In Table 4.3, the Φ-values for double mutations are shown with the single mutations Φ-values of those pairs. Among the 36 native contact pairs, only the sixteen native contact pairs are taken into consideration. This is due to computational overflows arising from the exceedingly large time scale difference between the fast and slow processes that were caused upon decreasing the attractive potentials.

Table 4.2. Coupling energies for native contacts

Type of native contact	$\Delta^2G_{int(N-D)}$	$\Delta^2G_{int(D-\ddagger)}$	$\Delta^2G_{int}^{\ddagger}$
AB	-0.327	0.010	-0.337
AC	-0.364	0.050	-0.414
AE	-0.367	0.010	-0.377
AH	-0.415	0.080	-0.495
BD	-0.366	-0.216	-0.15
BE	-0.358	-0.038	-0.32
BF	-0.115	-0.069	-0.046
CD	-0.633	0.030	-0.663
CG	0.006	0.148	-0.142
CI	-0.386	-0.030	-0.356
EG	-0.356	0.050	-0.406
EI	-0.367	-0.038	-0.329
FH	-0.356	-0.180	-0.176
GH	-0.533	0.116	-0.649
GI	-0.381	0.034	-0.415

The Φ-value of native contact pairs *CD* has a non-classical Φ-value, much more greater than 1. The native contacts *C* and *D* are key contacts for the fastest folding pathway. Destablizing these two contacts has a disruptive effect on that pathway. Thus blocking the fastest pathway upon mutations decreases the folding rate. Secondly, the extraordinary high Φ–value of *CD* and the coupling energies indicate that native contacts C and D work cooperatively for the folding process. On the other hand, *GH* has a negative Φ-value of –0.461. Native contacts *G* and *H* are the core contacts of the β-sheet. Decreasing the attractive potential of these contacts reduces the formation of β-sheet and thus the flux of the pathway where the transition structure is that of a partially folded α-helix and β-sheet. Therefore destabilizing both G and H redirects the flow to the dominant and faster pathway and increases the rate of folding, which is consistent with the single mutations of β-sheet contacts. Those double mutants having β-sheet contacts have negative Φ-values. The lowest coupling energies of *GH* and *CD* show that the native contacts, which are close in sequence and found in the core regions work cooperatively for the folding process.

Table 4.3. Comparison between single and double mutation

Type of native contacts (XY)	Φ_{XY}- values	Φ_X- values	Φ_Y- values	$\Phi_X + \Phi_Y$ values
AB	0.097	0.012	0.096	0.108
AC	0.299	0.012	0.251	0.263
AE	0.102	0.012	0.093	0.105
AH	-0.071	0.012	-0.296	-0.284
BD	0.372	0.096	0.990	1.086
BE	0.089	0.096	0.093	0.189
BF	0.006	0.096	0.035	0.131
BI	0.0243	0.096	-0.085	0.011
CD	2.58	0.251	0.990	1.241
CG	0.139	0.251	-0.357	-0.106
CI	0.080	0.251	-0.085	0.166
EG	-0.145	0.093	-0.357	-0.264
EI	-0.082	0.093	-0.085	0.008
FH	-0.439	0.035	-0.296	-0.261
GH	-0.461	-0.357	-0.296	-0.653
GI	-0.219	-0.357	-0.085	-0.442

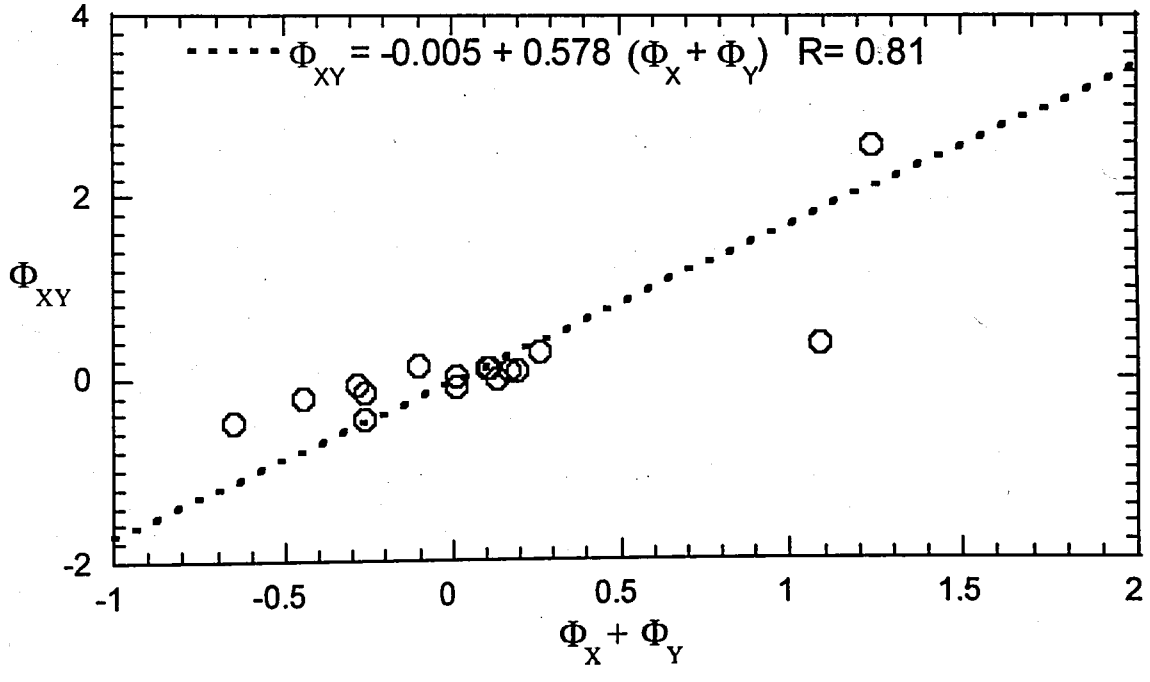

Figure 4.15. Correlation between the Φ-values of double mutations and the sum of corresponding Φ -value of single mutations

Figure 4.15 shows the correlation between the Φ-values of double mutations and the sum of the corresponding Φ-values from single mutations. It is interesting to see that the correlation coefficient is 0.81. One can think that the Φ-values are additive. However this result might be misleading since only 16 pairs of native contacts are considered among the total of 36 pairs. Among the 16 pairs, the strongest deviations from the correlation occurs at the pairs *BD* and *CD*. This is consistent with the observations that these three contacts are the folding nuclei (Vendruscolo *et al.*, 2001; Fersht, 2000) for the fast and dominant track of the landscape.

4.8. The Energy Landscape from SVD Analysis

Schematic and simulated energy surfaces enable us to compare the new view of folding with the more conventional classical picture such as pathways, transition states and intermediates. It is possible to visualize all these concepts in an ensemble context form by the help of energy landscape. For this purpose, the energy surface of the whole ensemble is analyzed by decreasing the dimensionality of the ensemble space. Each microconformation of 16-mer is represented by a 32-dimensional vector that is composed of the spatial x-, and y-coordinates of the individual residues in that particular microconformation. Thus, the M microconformations are organized in a 32xM matrix. The singular value decomposition of this matrix yields a new matrix of the same size, which is nothing else than the representation of the original matrix of conformations in the new (normal) space. Each column then designates the coordinates of a given conformation along the normal (principal) axes of the new frame. Using the dominant two directions, i.e. the first two rows of the M columns, the M microconformations are represented by single points in the two-dimensional spanned by the singular vectors. This way, the microconformations are located on a plane according to their structural similarities. The corresponding equilibrium energies determine the energy surface.

Figure 4.16 presents the energy surface plot of 522 microconformations having five or more native contacts. The native conformation is labelled as N on the surface. The shape of the landscape is really complicated even for this small subset of microconformations. This indicates that the folding cannot be described in terms of a single pathway such that the trajectories of all the microconformations are limited to a narrow region (Dinner *et al.*,

2000). We observe a deep channel, relatively further from the native state which can act as a trap. This reminds the second slowest channel of the kinetic scheme of macroconformation in Figure 4.7. The region next to the channel that slopes gently is closer to the native state. This is apparently a fast pathway to reach the native state.

Figure 4.17 shows the energy surface for the microconformations having more than five native contacts. The second minimum next to minimum of the native state corresponds to conformation having eight native contacts, *ABCDEFGH*. Time evolution of this conformation is shown in Figure 4.8 which has the second highest equilibrium probability. The microconformations having the native contacts *BCDEGHI* are also seen as minima on the energy surface next to the minimum of the native structure. The destabilization of native contact A in the native structure also destabilizes F and forms the microconformations having the native contacts *BCDEGHI*. Thus, it is reasonable to see such a minimum near the native conformation.

Maxima are seen on the landscape near microconformations having three or more native contacts (plot is not shown), which can be viewed as folding barriers. The two most pronounced maxima correspond to the microconformations having three β-sheet contacts, and three β-sheet contacts along with one α-helix contact. The result is consistent with the previous Φ-value analysis that the folding rate decreases when it starts from the formation of β-sheet contacts.

Magnifying the landscape around the native conformation by considering the most native-like microconformations, i.e. those of having seven or more native contacts (Figure 4.18), shows that the native conformation is at the global minimum and the landscape has a smooth funnel shape. The smooth funnel shape indicates that the folding process is fast and downhill provided that six native contacts are formed.

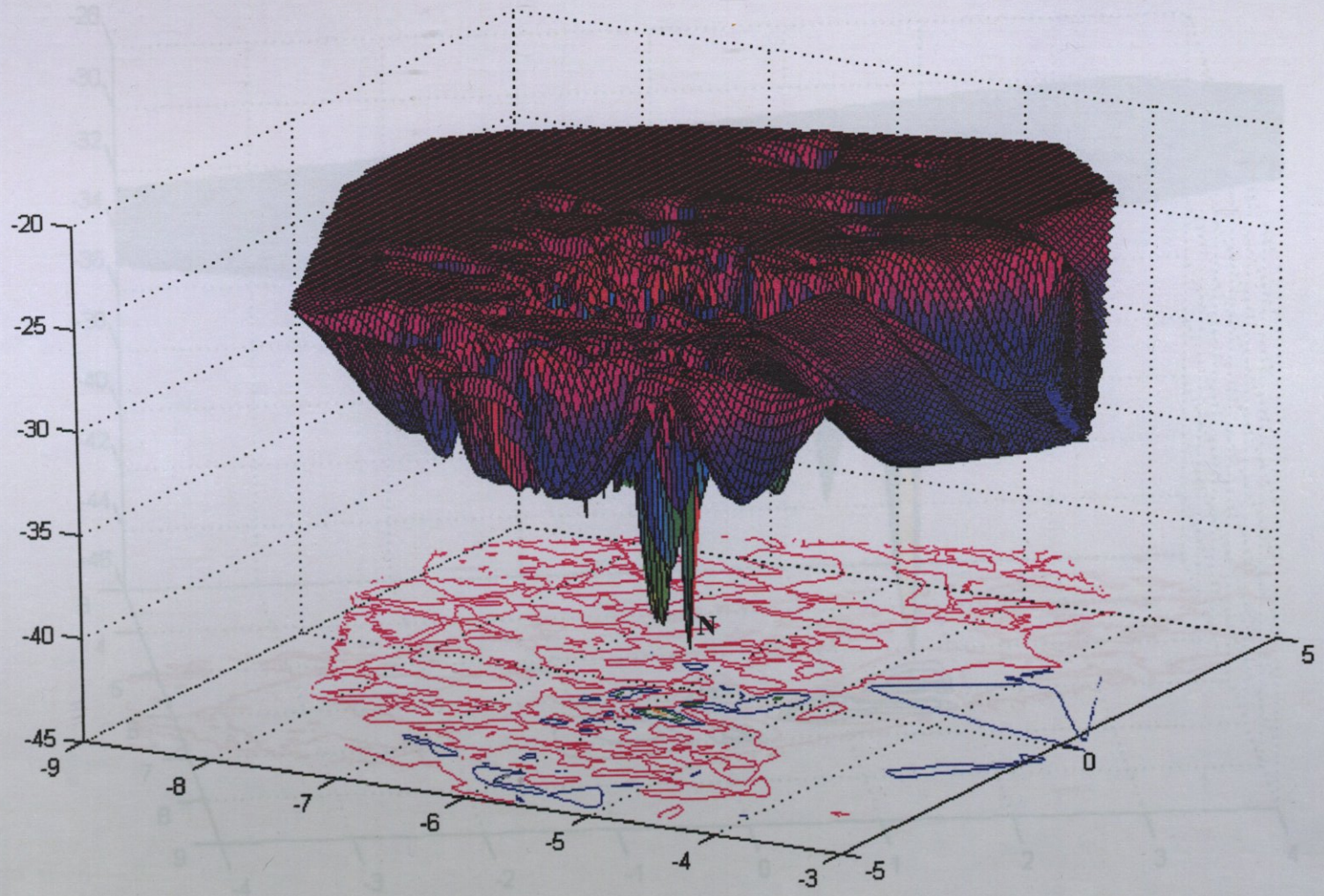

Figure 4.16. Energy surface map for the microconformation having more than four native contacts

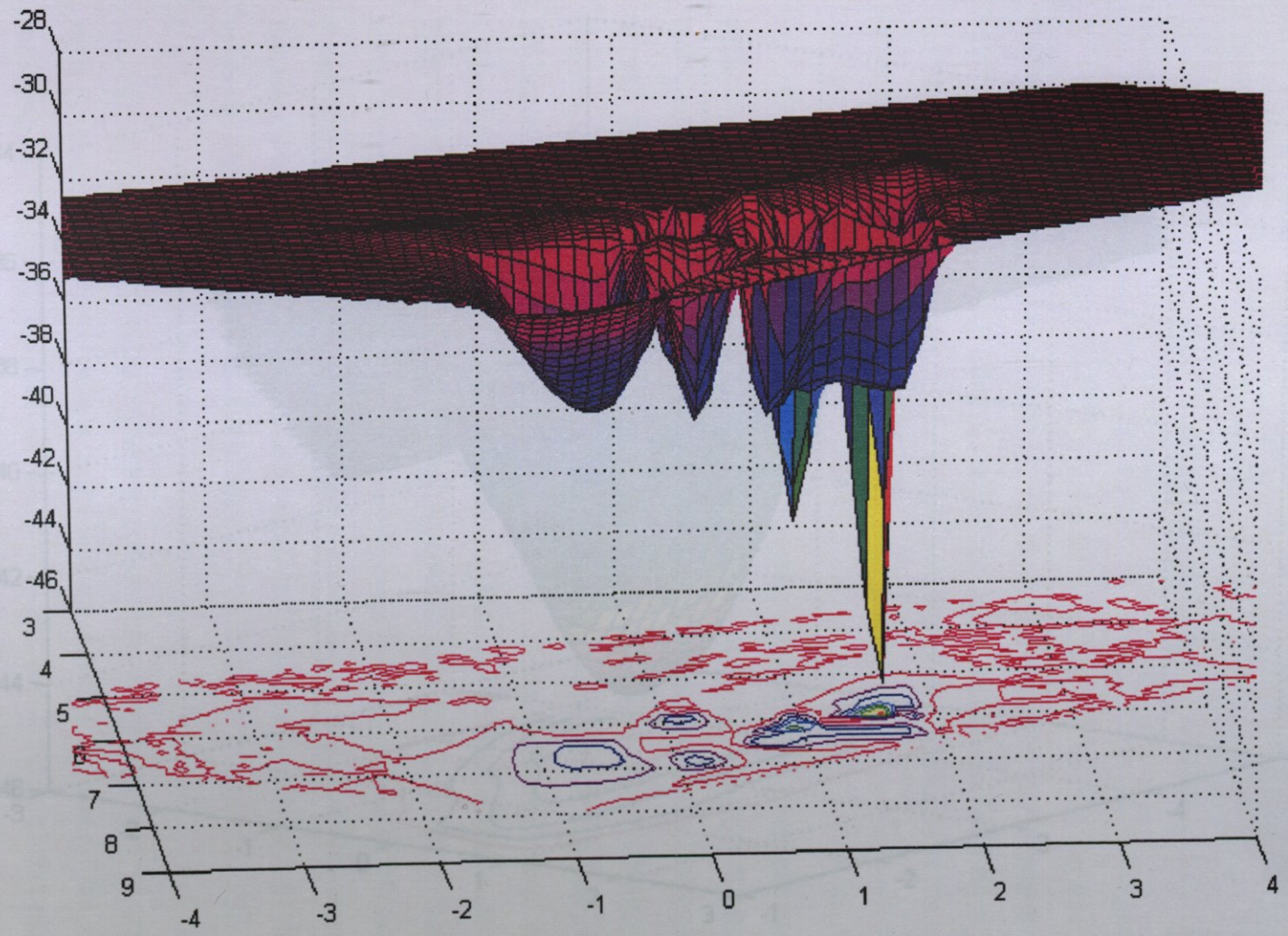

Figure 4.17 Energy surface map for the microconformation having more than five native contacts

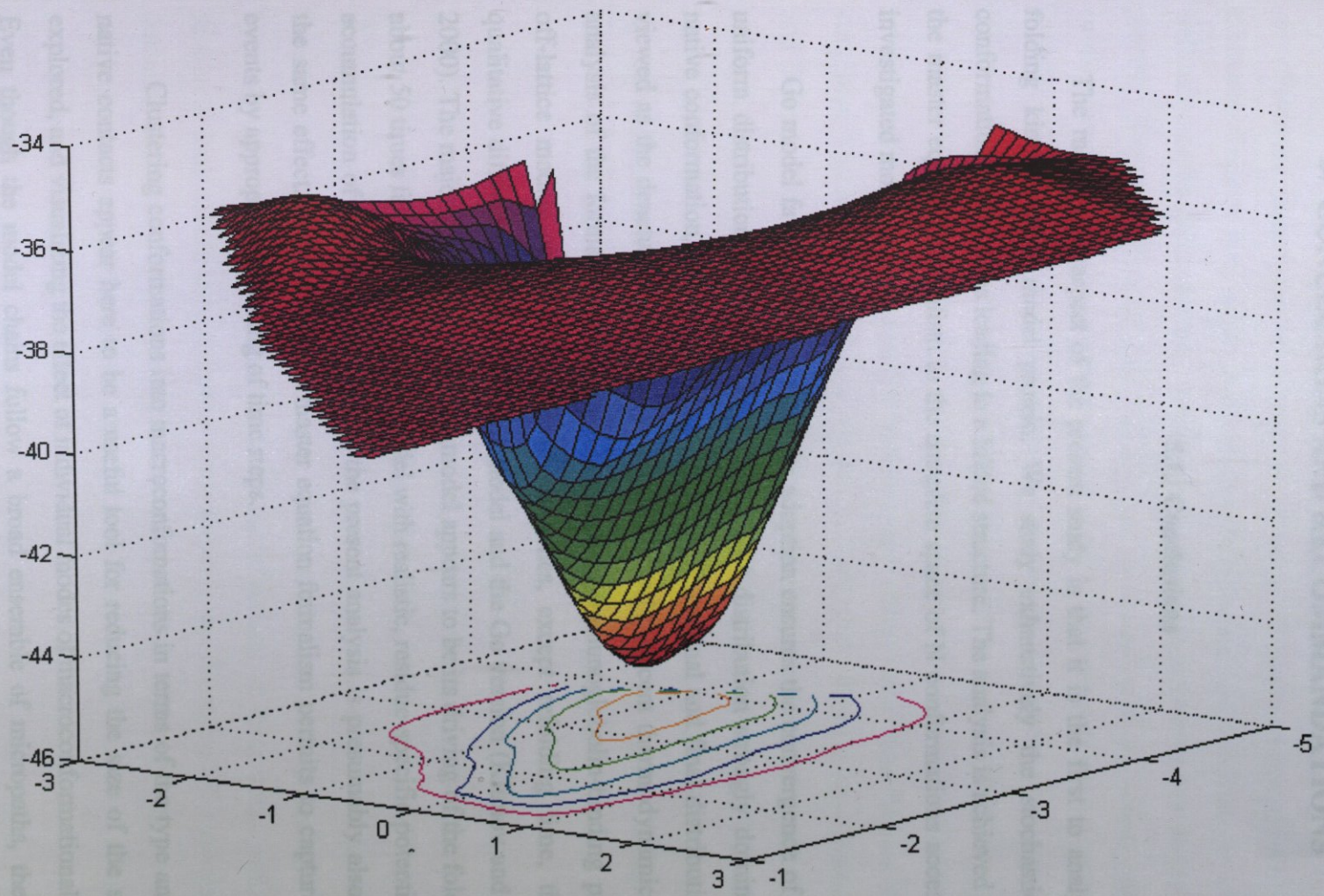

Figure 4.18 The Energy Surface Map for the micro conformations having more than six native contacts

5. CONCLUSIONS AND RECOMMANDATIONS

5.1. Conclusions

The most novel aspect of the present study is that it is the first to analyze the full folding kinetics of a model protein. We study exhaustively the stochastics of NxN conformational transitions leading to a folded structure. The analysis is achieved by applying the master equation formalism to the complete space of N conformations accessible to the investigated model chains.

Go model favors native contacts. Its adoption ensures the convergence of the original uniform distribution of conformations into a new distribution strongly dominated by the native conformation, following Boltzmann law. The original and final distributions are thus viewed as the denatured and native states, respectively. Recent thermodynamic and kinetic analysis of the formation of β-hairpin fragment of immunoglobulin-binding protein using off-lattice models showed that for all properties, except refolding time, there are no qualitative differences between the full model and the Go version (Klimov and Thirumalai, 2000). The major deficiency of the Go model appears to be its driving of the folding process about 50 times faster that in the full model with realistic, residue-specific potentials. The fast accumulation of the native structure in the present analysis is presumably also induced by the same effect. Nonetheless, the master equation formalism permits to capture all folding events by appropriate rescaling of time steps.

Clustering conformations into macroconfomations in terms of the type and number of native contacts appear here to be a useful tool for reducing the size of the system to be explored, and visualizing the effect of individual modes of macroconformational transitions. Even though the model chains follow a broad ensemble of micropaths, the statistically predominant pathway emerges when the ensemble context, "macroconformations" are taken into consideration. Previous MD simulations of CI2 unfolding also demonstrated that a dominant pathway, or a certain order of events in the folding process, is discernible when trajectories are analyzed in terms of evolution of native contacts (Lazaridis and Karplus, 1997).

From an alternative standpoint, examination of macroconformations reveals the existence of one or more substructures preferentially formed and stabilized at early stages, as well as on-pathway kinetic intermediates that accumulate at later stages during the folding process. The early forming substructures are reminiscent of secondary structures, or cooperative intradomain contacts, propagating off a contact made by a core residue. It is possible to identify one or more critical contacts, whose formation drives the formation of others at different levels. Whereas *C* and *G* effectively drive the α-and β-domain formations in the examined 16-mer, *D* plays a major role in stabilizing the overall 16-meric structure. Interestingly, a large number of substructures merge into well-defined intermediate structures. The intermediate macroconformation *ABCGHI,* where the two domains α–helix and β- strand are fully structured in the absence of interdomain contacts, illustrates this situation. On the other hand, another intermediate, *ABCDEF*, where the α-helical and interdomain contacts are fully formed, emerges by a dominant macropathway that is favored by the conformational entropic factor. This is evident by the high W values (Table 3.2) associated with each of the macroconformation along this macropath.

The formation of native contacts obeys a preferred order in which local interactions generally precede more nonlocal contacts, resembling a zippers process. First, there is a propensity of *local* contacts at short times and among these, the core contacts are first stabilized at early stages of folding, then these are succeeded by local contacts at chain termini. The latter, although forming at the burst stage, exhibit a tendency to fluctuate between open and closed conformations, which persists at long times and results in a relatively long effective stabilization time. Second, the formation of nonlocal or tertiary contacts are observed at intermediate and longer times. Thus folding mechanism is a hiearchic mechanism that begins with the formation of local structures of only marginal stability which then interact with tertiary contacts to be stabilized.

The observed hierarchy of native contact formation/stabilization suggests that the nucleus is composed of partly or well formed of secondary structural elements (local contacts) that are stabilized by tertiary interactions. The formation of the innermost tertiary contact apparently plays a key role in the progress of folding (Fersht,1997; Fersht, 2000).

The transition state is usually described as a saddle point between two minima in two-dimensional energy maps. Here, this is a metastable state characterized by a macroconformation having a high population of a number of native contacts, which upon formation of an additional contact, is immediately folded into the native structure. The macroconformation *CD* and the contact *B* appear to play this role in the dominant pathway of the 9-mer model. Near the transition state, the ensemble of chains consists of the subset *CD* in the first place, and the fully folded chains *ABCD*, the populations of the intermediate subsets *ABC* and *BCD* being relatively low. The low population of these two macroconformations persists until formation of contact *B* that immediately drives contact A, and thereby leads to the completion of folding. In the case of 16-mers, the elapsed times between successive macroconformations reveal that the rate limiting macrostates along two dominant pathways are those having six (*ABCGHI*) and five (*BCDEF*) native contacts, respectively. Thus, folding is a parallel process involving different pathways. These different pathways can lead to different folding times if they are not able to efficiently communicate. The fastest pathway or more exactly the slowest step of the fastest pathway determines the observed folding kinetics.

There is a fast accumulation of native state, with minimal transient accumulation of kinetic intermediates. Thus, even a slight increase in the complexity of the model leads to an immense diversity in the routes navigated by the denatured conformers, during the folding process, in accord with the new view of protein folding. The rapid accumulation of native conformation, in consistent with one dominant native basin of attraction. Such a behavior has been pointed to be typical of sequences having low σ values, where $\sigma = (T_\theta - T_F)/T_\theta$, T_θ and T_F being the equilibrium collapse and folding temperatures; whereas those having moderate to large σ values would obey a kinetic partitioning mechanism, i.e. a fraction of molecules attain the native state by off-pathway processes that involve trapping in misfolded structures (Veitshans *et al.*, 1996). No effective trapping in intermediate structures is detected in the present analysis, except for a transient accumulation of the significantly structured (native-like) conformations, close to the completion of folding. However, their population remains significantly small compared to that of fully folded molecules.

Experimentally observed nonclassical Φ-values, i.e., $\Phi > 0$ and $\Phi < 0$ may arise from parallel microscopic flow processes, such as in funnel-shaped energy landscapes. Negative Φ-values result when a mutation destabilizes a slow flow channel, causing an overflow into a faster flow channel. Φ-values greater than one occur when mutations redirect a fast flow into a slower channel. Φ-values are not simple kinetic rulers of the progress along a reaction coordinate, but a measure of the type and extent of 'changes in rates' due to mutations. Sites having positive Φ-values indicate where mutations decelerate folding and negative Φ-values indicate where mutations accelerate folding. In as far as the absolute values are concerned, |Φ|-values define the degree to which a contact is a *gatekeeper* site, controlling the folding flow. Sites having $|\Phi| >> 0$ are gatekeepers; sites with $\Phi \sim 0$ have little flow control.

Finally, the energy surface of a subset of microconformations shows that the trajectories of different micrconformations passs through very broad regions of conformational space even in this simple lattice model. The folding can be described as a parallel folding of a set of arrows that has a tree structure: many branches at the top, narrowing down to a few arrows at the bottom.

5.2. Recommendations

The analysis of simple model proteins whose native structure is composed of α-helix and β-strand domains shows that (i) the dominant pathway, that is favored by conformational entropy happens to be the fastest pathway. There are direct and indirect microtrajectories choosing this pathway. (ii) The TS structure is composed of partially folded α-helix and interdomain contact that reminds the TS structure of CI2 (Fersht, 1999). It would be complimentary to analyze the folding pathways of the native structures that are in the forms of only β-strand or α-helix. This model enables us to see the differences in the folding processes in comparison to the differences in the structural class of native structures. We can gain insights as to the folding times of different native structures.

Reducing the size of the system by clustering conformations into macroconfomations in terms of the type and number of native contacts gives insight in visualizing the folding process in an ensemble context. This approach can be applied to real proteins. The complete

set of conformations can be generated using off-lattice models. The generated conformations can be screened using different criteria (Ozkan and Bahar, 1998). Then the kinetic analysis can be performed using different subsets of ensembles.

In the present study, Go models are used for analyzing the folding kinetics of simple proteins. Go models are considered as minimally frustrated systems, unlike real proteins. In order to understand the role of non-native contacts during the folding process, the folding kinetics of the systems should be re-analyzed by assigning some attractive potential to non-native contacts. The effect of the sequence of the monomer on folding can be explored by a simple approach such as assigning a HP sequence to the monomer chain.

REFERENCES

Alexander, P., J. Orban and P. Bryan, 1992, "Kinetic Analysis of Folding and Unfolding the 56 Amino Acid IgG-binding Domain of Streptococcal Protein G", *Biochemistry* Vol. 31, pp.7243-7248.

Anfinsen, C., 1973, "Principles that Govern the Folding of Protein Chain", *Science*, Vol. 123, pp. 223-227.

Bahar, I., A. Wallqvist, D. Covell, and R. L. Jernigan, 1998, "Correlations Between Hydrogen Exchange from Native Proteins and Cooperative Residue Fluctuations from a Simple Model", *Biochemistry*, Vol. 37, pp. 1067-1075.

Baker, D. and D. A. Agard, 1994, "Kinetics versus Thermodynamics in Protein Folding", *Biochemistry*, Vol. 33, pp. 7505-7509.

Baker, D., 2000,"A Surprising Simplicity to Protein Folding", *Nature*, Vol. 405, pp. 39-42.

Baldwin, R. L., 1995, "The Nature of Protein Folding Pathways: the Classic versus the New View", *Journal of Biomolecular NMR,* Vol. 5, pp. 103-109.

Baldwin, R. L., 2001, "Folding Consensus", *Nature Structural Biology*, Vol. 8, pp. 92-94.

Ballew., R. M., J. Sabelko and M. Gruebbele, 1996, "Direct Obsevation of Fast Protein Folding: The Initial Collapse of Apomyoglabin", *Proceedings of the National Academy of Science USA,* Vol. 93, pp. 5759-5764.

Bender, M. L., R. J Bergeron and M. Komiyama, 1984, *The Bioorganic Chemistry of Enzymatic Catalysis,* John Wiley, New York.

Brooks, C., M. Gruebele, J. S. Onuchic and P. G. Wolynes, 1998, "Chemical Physics of Protein Folding", *Proceedings of the National Academy of Science USA,* Vol. 95, pp. 11037-11038.

Bryngelson, J D. and P.G. Wolynes, 1987, "Spin Glasses and the Statistical Mechanics of Protein Folding", *Proceedings of the National Academy of Science USA,* Vol. 84, pp. 7524-7528.

Bryngelson, J. D, J. N. Onuchic,N. D. Socci and P. G. Wolynes, 1995, "Funnels, Pathways and the Energy Landscape of Protein Folding: a Synthesis", *Proteins,* Vol. 21, pp. 167-195.

Chan, H. S and K. A. Dill, 1998, "Protein Folding in the Landscape Perspective: Chevron Plots and Non-Arrhenius Kinetics", *Proteins,* Vol. 30, pp. 2-33.

Chan, H. S. and K. A. Dill, 1993, "Energy Landscapes and the Collapse Dynamics of Homopolymers", *Journal of Chemical Physics*, Vol. 99, pp. 2116-2127.

Cieplak, M., M. Henkel, J. Karbowski and J. R. Banavar, 1998, "Master Equation Approach to Protein Folding and Kinetic Traps", *Physical Review Letters,* Vol. 80, pp. 3654-3657.

Clementi, C., H. Nymeyer and J.S. Onuchic, 2000, "Topological and Energetic Factors: What Determines the Structural Details of the Transition State Ensemble and En route Intermediates for Protein Folding? An Investigation for Small Globular Proteins", *Journal of Molecular Biology*, Vol. 298, pp. 938-953.

Creighton, T., 1992, *Protein Folding*, W.H. Freeman, New York.

Creighton, T., 1995, "Disulphide-Coupled Protein Folding Pathways, Philosophy of Transactions Royal Society of London, Vol. 348, pp. 5-10.

Daggett, V., A. Li, S. L. Itzhaki, D. E. Otzen and A. R. Fersht, 1996, " Structure of the Transition State for Folding of a Protein Derived from Experiment and Simulation", *Journal of Molecular Biology,* Vol. 257, pp. 430-440.

Dill, K. A., 1990, "The Meaning of Hydrophobicity", *Science,* Vol. 250, pp. 297-298.

Dinner, A. R., A. Sali, L. J. Smith, C. M. Dobson and M. Karplus, 2000, "Understanding Protein Folding via Energy Surfaces from Theory and Experiment", *Trends in Biochemical Sciences*, Vol. 25, pp. 331-339.

Dokholyan, N.V., S. D. Buldyrev, H. E. Stanley and E. I. Shakhnovich, 2000, "Identfying the Protein Folding Nucleus Using Molecular Dynamics", *Journal of Molecular Biology,* Vol. 296, pp. 1183-1188.

Eaton, W. A., V. Munoz, S. J. Hagen, G. S. Jas, L. J. Lapidus, E. R. Henry and J. Hofrichter, 2000, "Fast Kinetics and Mechanisms in Protein Folding", Annual Reviews of Biophysics and Biomolecular Structure, Vol. 29, pp. 327-359.

Englander, S. W., 2000, "Protein Folding Intermediates and Pathways Studied by Hydrogen Exchange" *Annual Review Biophysical and Biomolecular Structure,* Vol. 29, 213-238.

Erman, B. and K. Dill, 2000, " Gaussian Model of Protein Folding", Journal of Chemical Physics, Vol. 112, pp. 1050-1056.

Fersht, A. R., 1995, "Characterizing Transition States in Protein Folding: An Essential Step in the Puzzle", *Current Opinion Structure Biology,* Vol. 5, pp. 79-84.

Fersht, A. R., 1999, *Structure and Mechanism in Protein Science*, Freeman, New York.

Fersht, A. R., A. Matouscheck and L. Serrano, 1992, "The Folding of an Enzyme, I Theory of Protein Engineering Analysis of Stability and Pathway of Protein Folding", *Journal of Molecular Biology*, Vol. 224, pp. 771-782.

Fersht, A. R., L. S. Itzhaki, N.F. Elmarsy, J. M. Matthews and D. E. Otzen, 1994, "Single versus Parallel Pathways of Protein Folding and Fractional Formation in the Transition State, *Proceedings of the National Academy of Science USA*, Vol. 91, pp. 10426-10429.

Fersht. A. R., 2000, "Transition-state Structure as a Unifying Basis in Protein Folding Mechanisms: Contact Order, Chain Topology, Stability and the Extended Nucleus Mechanism", *Proceedings of the National Academy of Science USA,* Vol. 97, pp. 1525-1529.

Finkelstein, A. V. and A. Y. Badretdinov, 1997, "Rate of Protein Folding Near the Point of Thermodynamic Equilibrium Between the Coil and the Most Stable Chain Fold", *Folding and Design*, Vol. 2, pp. 115-121.

Gay, G., J. Ruiz-Sans, B. Davis and A. R. Fersht, 1994, "The Structure of the Transition State for the Association of Two Fragments of the Barley Chymotyripsin Inhibitor 2 to Generate Native-like Protein: Implications for Mechanisms of Protein folding", *Proceedings of the National Academy of Science USA*, Vol. 91, pp. 10943-10946.

Grantcharova, V. P., D. S. Riddle, J. V. Santiago and D. Baker, 1998, "Important Role of Hydrogen Bonds in Structurally Polarized Transition State for the Folding of Src SH3 Domain", *Natural Structural Biology*, Vol. 5, pp.714-720.

Gruebele, M., 1999, "The Fast Protein Folding Problem", *Annual Review Physical Chemistry*, Vol. 50, pp. 485-516.

Harrison, S. C. and R. Durbin, 1985, "Is There a Single Pathway for the Folding of a Polypeptide Chain", *Proceedings of the National Academy of Science USA,* Vol. 85, pp. 4038.

Hoang, T. X. and Cieplak M., 2000, "Molecular Dynamics of Folding of Secondary Structures in Go-type Models of Proteins", *Journal of Chemical Physics,* Vol. 112, pp. 6851-6863.

Horrovitz, A. and A. R. Fersht, 1992, "Co-operative Interactions during Protein Folding", *Journal of Molecular Biology,* Vol. 224, pp. 733-740.

Ikai, A. and C. Tanford, 1971, "Kinetic Evidence for Incorrectly Folded Intermediate States in the Refolding of Denatured Proteins", *Nature,* Vol. 230, pp. 100-102.

Jacob, M. and F. X. Schimid, 1999, "Protein Folding as a Diffusional Process", *Biochemistry,* Vol. 38, pp. 13773-13779.

Jencks, W. P., 1987, *Catalysis in Chemistry and Enzymology,* Dover, New York.

Karplus, M and D. L. Weaver, 1976, "Protein-folding Dynamics", *Nature*, Vol. 5550, pp. 404-406

Kiefhaber, T., 1995, "Kinetic Traps in Lysozyme Folding", *Proceedings of the National Academy of Science USA*, Vol. 92, pp. 9029-9033.

Klimov, D. K. and D. Thirumalai, 1998, "Lattice Models for Protein Reveal Multiple Folding Nuclei for Nucleation-Collapse Mechanism", *Journal of Molecular Biology*, Vol. 282, pp. 471-492.

Klimov, D. K. and D. Thirumalai, 2000, "Mechanisms and Kinetics of ß-hairpin Formation", *Proceedings of the National Academy of Science USA*, Vol. 97, pp. 2544-2549.

Kragelund, B. B., C. V. Robinson, V. Knudsen, C. M. Dobson and F. M. Poulsen, 1995, "Folding of a Four Helix Bundle: Studies of Acyl-coenzyme A Binding Protein", *Biochemistry,* Vol. 34, pp. 7217-7224.

Kyte, J., 1995, *Structure in Protein Chemistry*, Garland Publishing, New York.

Ladurner, A. G., S. L. Itzhaki, V. Daggett and A. R. Fersht, 1998, "Synergy Between Simulation and Experiment in Describing the Energy Landscape of Protein Folding", *Proceedings of the National Academy of Science USA*, Vol. 95, pp. 8473-8478.

Laurents, D. V. and R. L. Baldwin, 1998, "Protein Folding: Matching Theory and Experiment", *Biophysical Journal,* Vol.75, pp.428-434, 1998.

Lazaridis, T. and M. Karplus, 1997, "New View of Protein Folding Reconciled with the Old View Through Multiple Unfolding Simulations" *Science,* Vol. 278, pp. 1928-1930.

Leopod, P. E., M. Montal and J. S. Onuchic, 1992, "Protein Folding Funnels: A Kinetic Approach to the Sequence-structure Relationships", *Proceedings of the National Academy of Science USA,* Vol. 89, pp. 8721-8727.

Li, L., L. A. Mirny and E. Shakhnovich, 2000, "Kinetics, Thermodynamics and Evolution of Non-native Interactions in a Protein Folding Nucleus", *Nature Structural Biology,* Vol. 7, pp. 336-342.

Martinez, J. C., M. T. Pisabarro and L. Serrano, 1998, "Obligatory Steps in Protein Folding and the Conformational Diversity of Transition State and Conformational Diversity of Transition State", *Natural Structural Biology*, Vol. 5, pp. 721-729.

Maskill, H., 1985, *The Physical Basis of Organic Chemistry,* Oxford University Press, New York.

Matagne, A., E. W. Chung, L. J. Ball, S. E. Radford, C.V. Robinson and C. M. Dobson, 1998, "The Origin of the a-Domain Intermediate in the Folding of Hen Lysozyme", *Journal of Molecular Biology,* Vol. 277, pp.997-1005.

Matagne, A., M. Jamin, E. W. Chung, S. E. Radford, C.V. Robinson and C. M. Dobson, 2000, "Thermal Unfolding of an Intermediate is Associated with Non-Arhenius Kinetics in the Folding of Hen Lysozyme", *Journal of Molecular Biology,* Vol. 297, pp. 193-210.

Matagne, A., S. E. Radford and C. M. Dobson, 1997, "Fast and Slow Tracks in Lysozyme Folding: Insight into the Role of Domains in the Folding Process", *Journal of Molecular Biology,* Vol. 267, pp.1068-1074.

Matouscheck, A., J. T. Kellis, L. Serrano and A. Fersht, 1989, "Mapping the Transition State and Pathway of Protein Folding by Protein Engineering", *Nature*, Vol. 340, pp. 122-126.

Matouscheck, A., L. Serrano and A.R. Fersht, 1992, "The Folding of Enzyme: IV Structure of an Intermediate in the Refolding of Barnase Analysed by a Protein Engineering Procedure", *Journal of Molecular Biology,* Vol. 224, pp. 819-835.

Matthews, C. R., 1993, "Pathways of Protein Folding", *Annual Review Biochemistry,* Vol. 62, pp. 653-683.

Munoz, V., E. R. Henry, J. Hofrichter, and W. A. Eaton, 1998, "A Statistical Mechanical Model for β-hairpin Kinetics", *Proceedings of the National Academy of Science USA*, Vol. 95, pp. 5872-5879.

Nolting, B. and K. Andert, 2000, "Mechanism of Protein Folding", *Proteins: Structure Functions and Genetics*, Vol. 41, pp. 288-298.

Nymeyer, H., N. D. Socci and J.S. Onuchic, 2000, "Landscape Approaches for Determining the Ensemble of Folding Transition State: Success and Failure Hinge on the Degree of Frustration", *Proceedings of the National Academy of Science USA*, Vol. 97, pp. 634-639.

Onuchic, J. N., H. Nymer, A. E. Garcia, J. Chahine and N. D Socci, 2000, "The Energy Landscape of Protein Folding: Insights into Folding Mechanisms and Scenarios", *Advances in Protein Chemistry*, Vol. 53, pp. 87-152.

Ozkan, B. and I. Bahar, 1998, "Recognition of native structure from complete enumeration of low resolution models with constraints", *Proteins,* Vol. 32, pp. 211-222.

Pande, V. J. and D. S D, Rokhsar, 1999,"Molecular Dynamics Simulations of Unfolding and Refolding of ß-hairpin Fragment of Protein G", *Proceedings of the National Academy of Science USA*, Vol. 96, pp. 9062-9067.

Pande, V. J., A. Y. Grosberg, T. Tanaka and D. S. Rokhsar, 1998, "Pathways for Protein Folding : Is a New View Needed", *Current Opinion in Structural Biology*, Vol. 8, pp. 68-79.

Pross, A., 1995, *Theoretical and Physical Principles of Organic Reactivity,* JohnWiley, New York.

Ptitsyn, O. B. and A. A. Rashin, 1975, "A Model of Myoglobin Self-organization", *Biophysical Chemistry,* Vol. 3, pp. 1-20.

Ptitsyn, O. B., 1995, "Molten Globule and Protein Folding", *Advances in Protein Chemistry*, Vol. 47, pp. 83-229.

Rumbley, J., L. Hoang, L. Mayne and S. W. Englander, 2001,"An Amino Acid Code for Protein Folding", *Proceedings of the National Academy of Science USA,* Vol. 98, pp. 105-112.

Schindler, T., M. Herrler, M. A. Marahiel and F. X. Schmid, 1995, "Extremely Rapid Protein-Folding in the Absence of Intermediates", *Nature Structural Biology* Vol. 2, pp. 663-673.

Schulz, G. E. and R. H. Schirmer, 1979, *Principles of Protein Structure*, Springer, New York.

Shakhnovich, E. I., 1997, "Theoretical Studies of Protein Folding Thermodynamics and Kinetics", Current Opinion in Structural Biology, Vol. 7, pp. 29-40.

Shea, J., J. N Onuchic, and C. L Brooks, 2000, "Energetic frustration and the nature of the transition state in protein folding", *Journal of Chemical Physics,* Vol, 113, pp.7663-7671.

Socci, N. D., J. N. Onuchic and P. G. Wolynes, 1998, "Protein Folding Mechanisms and the Multidimensional Folding Funnel", *Proteins,* Vol. 32, pp. 136-158.

Takada, S., 1999, "Go-ing for the Prediction of Folding Mechanisms", *Proceedings of the National Academy of Science USA*, Vol.96, pp.11698-11700.

Ternstorm, T., U. Mayor, M. Akke and M. Oliveberg, 1999, "From Snapshot to Movie: Φ Analysis of Protein Folding Transition States Taken One Step Further", *Proceedings of the National Academy of Science USA*, Vol. 96, pp. 14854-14859.

Thirumalai, D. and D. K. Klimov, 1999, "Deciphering the Timescales and Mechanisms of Protein Folding Using Minimal Off-lattice Model", *Current Opinion in Structural Biology*, Vol. 9, pp. 197-297.

Tiana, G. and R. A. Broglia, 2001, "Statistical Analysis of Native Contact Formation in the Folding of Designed Model Proteins", *Journal of Chemical Physics*, Vol. 114, pp. 2503-2510.

Ueda Y., Taketomi H. and N. Go, 1975, "Studies on Protein Folding Unfolding and Fluctuations by Computer Simulations", *International Journal of Peptide Research,* Vol. 7, pp. 445-459.

Veitshans, T., D. K. Klimov, and D. Thirumalai, 1996, "Protein Folding Kinetics: Timescales, Pathways and Energy Landscapes in Terms of Sequence-Dependent Proteins", *Folding and Design,* Vol. 2, pp.1-22.

Vendruscolo, M., E. Paci,.C. M. Dobson and M. Karplus, 2001, "Three Key Residues Form a Critical Network in a Protein Transition State", *Nature*, Vol. 409, pp. 641-644.

Voet, D. and J. G. Voet, 1995, *Biochemistry*, John Wiley, New York.

Ye, Y. J., D. R. Ripoll and H. A. Scheraga, 1999, "Kinetics of Cooperative Protein Folding Involving Two Separate Conformational Families", *Computational and Theoretical Polymer Science*, Vol. 9, pp. 359-370.

Zhou, Y. and M. Karplus, 1999, "Interpreting the Folding Kinetics of Helical Proteins", *Nature*, Vol. 402, pp. 400-403.

Zwanzig, R., 1997, "Simple Model of Protein Folding Kinetics", *Proceedings of the National Academy of Science USA,* Vol. 92, pp. 9801-9804.

Zwanzig, R., 1997, "Two-state Models of Protein Folding Kinetics", *Proceedings of the National Academy of Science USA,* Vol. 94, pp. 148-150.

Boğaziçi University Library
Reference Section
CIRCULATE