Aus dem Max-Planck-Institut für Kolloid- und Grenzflächenforschung

Transition states and loop-closure principles in protein folding

Habilitationsschrift

zur Erlangung des akademischen Grades
Doctor rerum naturalium habilitatus
(Dr. rer. nat. habil.)
in der Wissenschaftsdisziplin Theoretische Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät der Universität
Potsdam

von
Dr. Thomas Weikl
geboren am 1. 4. 1970 in Passau

Potsdam, im Juli 2007

This work is licensed under a Creative Commons License:
Attribution - Noncommercial - Share Alike 2.0 Germany
To view a copy of this license visit
http://creativecommons.org/licenses/by-nc-sa/2.0/de/deed.en

Online published at the
Institutional Repository of the Potsdam University:
http://opus.kobv.de/ubp/volltexte/2008/2697/
urn:nbn:de:kobv:517-opus-26975
[http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-26975]

Contents

1 **Introduction to protein folding** **5**
  1.1 Protein structures 5
  1.2 Folding kinetics 6
  1.3 Mutational analysis of the folding kinetics 10
  1.4 Traditional interpretation of Φ-values 11
  1.5 Overview 13

2 **Transition states** **16**
  2.1 A model for small β-sheet proteins 16
  2.2 The transition states of the FBP and PIN WW domains 20
  2.3 Modeling mutational data for α-helices 28
  2.4 The helices of protein A and CI2 31
  2.5 Summary 38

3 **Loop-closure principles** **41**
  3.1 Topology and loop closure 41
  3.2 Contact maps, contact clusters, and topology 43
  3.3 Folding rates and topological measures 45
  3.4 Effective contact order and folding routes 53
  3.5 Kinetic impact and average Φ-values 60
  3.6 Summary 69

4 **Folding cooperativity** **71**
  4.1 Contact clusters and energy landscapes 71
  4.2 Cooperativity in two-state protein folding kinetics 75
  4.3 Parallel and sequential unfolding events in MD simulations 79
  4.4 Substructural cooperativity 86
  4.5 Summary 89

Publications used in this work **91**

Bibliography **92**

Acknowledgements **113**

1 Introduction to protein folding

1.1 Protein structures

Proteins are biomolecules that participate in all cellular processes of living organisms. Some proteins have structural or mechanical function, such as the protein collagen, which provides the structural support of our connective tissues, or the proteins that form the cellular cytoskeleton. Other proteins catalyze biochemical reactions, transport or store electrons, ions, and molecules, perform mechanical work in our muscles, transmit information within or between cells, act as antibodies in immune responses, or control the expression of genes and, thus, the generation of other proteins. Proteins achieve this functional versatility by folding into different, unique three-dimensional structures, which distinguishes them from other large classes of biomolecules such as nucleic acids, polysaccharides or lipids [1].

Proteins are polymers that are built up from twenty different standard types of amino acids. Each amino acid consists of a central carbon atom, called C_α, to which an amino group NH_2, a carboxyl group COOH, a hydrogen atom, and a side chain are attached. In a protein chain, the amino acids are covalently connected by peptide bonds. A peptide bond is formed when the carboxyl group of one amino acid reacts with the amino group of another amino acid, under release of a water molecule. Each type of amino acid has a characteristic side chain. In standard classifications, the twenty different side chains are grouped into hydrophobic side chains, polar side chains, and charged side chains [2].

The amino acid sequence of a protein chain is also called the *primary structure* of a protein. The sequence determines into which three-dimensional structure a protein folds. The three-dimensional, folded structure of a protein has characteristic *secondary structural elements*, α-helices and β-strands. The interactions of secondary elements in the folded structure are denoted as *tertiary interactions* or *tertiary structure* of the protein [2]. The folded structure of the protein CI2 shown in fig. 1.1, for example, contains a single α-helix and four β-strands as secondary structural elements. The β-strands form a four-stranded β-sheet,

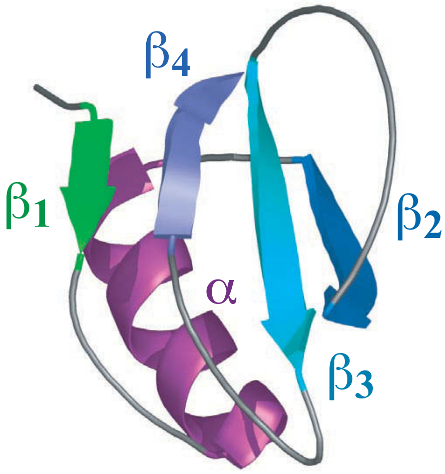

Figure 1.1: The structure of the protein CI2 consists of an α-helix packed against a four-stranded β-sheet [3]. In this 'cartoon representation' [2], only the protein backbone is shown, with schematic illustrations of strands and helix, while the amino acid sidechains are omitted. CI2 is a two-state protein that folds from the denatured state to the native state without experimentally detectable intermediate states (see section 1.2).

which is packed against the α-helix. To date, the three-dimensional structures of close to 40000 proteins have been determined by X-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy and have been deposited to the Protein Data Bank (PDB) of the Research Collaboratory for Structural Bioinformatics (RCSB).

1.2 Folding kinetics

How proteins fold into their native, three-dimensional structure remains an intriguing question [4]. Given the vast number of unfolded protein conformations, Cyrus Levinthal argued in 1968 [5, 6] that proteins are guided to their native structure by a sequence of folding intermediates. In the following decades, experimentalists focused on detecting and characterizing metastable intermediates with a variety of methods [7]. While such folding intermediates continue to be of considerable interest [8, 9], the view that proteins have to fold in sequential pathways from intermediate to intermediate, now known as 'old view' [10, 11], changed in the '90s when statistical-mechanical models demonstrated that fast and efficient folding can also be achieved on funnel energy landscapes that are smoothly biased towards the native state and do not exhibit metastable intermediates [12, 13]. The paradigmatic proteins of this 'new view' are two-state proteins, first discovered in 1991 [14]. Two-state proteins fold from the denatured state to the native state without experimentally detectable intermediate states. Since then, many small single-domain proteins have been shown to fold in two-state kinetics [15–17].

A characteristic signature of two-state folding is the single-exponential relaxation of an ensemble of proteins into equilibrium, see fig. 1.2(a). In

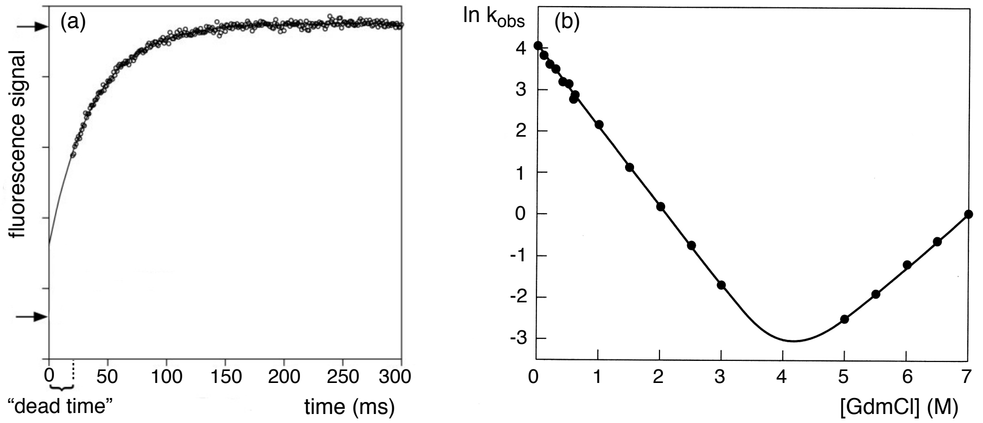

Figure 1.2: (a) Typical time-dependent fluorescence signal from a rapid mixing experiment (adapted from [18]). The protein solution initially contains a high concentration of chemical denaturant. At time $t = 0$, the denaturant is diluted by rapid mixing. The arrows indicate the fluorescence signal in the 'old' and 'new' equilibrium. – (b) "Chevron plot" for the two-state protein CI2 shown in fig. 1.1 (adapted from [16]). The logarithm of the observed relaxation rate k_{obs} from rapid mixing experiments is plotted as a function of the guanidinium chloride concentration after mixing. In the 'left arm' of the plot (low denaturant concentration), the relaxation rate $k_{\text{obs}} = k_f + k_u$ is dominated by the folding rate k_f, and in the 'right arm' (high denaturant concentration) by the unfolding rate k_u. The linear slope of the two 'arms' results from a linear dependence of $\ln k_f$ and $\ln k_u$ on the denaturant concentration, see eqs. (1.3) and (1.4).

the denatured state, the proteins can adopt a large number of unfolded conformations, such as the conformation shown in fig. 1.3(b) below. The native state of a small protein, in contrast, essentially consists of a single folded conformation. In rapid mixing experiments, the protein solution initially contains, e.g., a high concentration of the chemical denaturants urea or guanidinium chloride, which stabilizes the denatured state of the protein. At time $t = 0$, the denaturant is diluted by rapid mixing, and the protein solution starts to relax into its new equilibrium. The fluorescence signal is emitted from aromatic amino acids and changes during folding because of the different chemical environment of these amino acids in the denatured and native state. The arrows in fig. 1.2(a) indicate the fluorescence signal in the 'old' and 'new' equilibrium. The characteristic 'dead times' of rapid mixing experiments are in the millisecond range. Submillisecond folding processes can be observed, e.g., in temperature-jump experiments in which the initial equilibrium is perturbed by a short laser pulse that increases the temperature of the solution by several degrees Celsius [18].

The two-state relaxation process can be described by the equation

$$\frac{dP_N(t)}{dt} = k_f P_D(t) - k_u P_N(t) \tag{1.1}$$

Here, $P_N(t)$ and $P_D(t) = 1 - P_N(t)$ are the time-dependent probabilities for the native state N and denatured state D, and k_f and k_u are the folding and unfolding rate. The solution of eq. (1.1) is

$$P_N(t) = a + b \exp[-(k_f + k_u)t] \tag{1.2}$$

with $a = k_f/(k_f + k_u)$ and $b = P_N(t = 0) - a$, which represents a single-exponential relaxation with rate $k_f + k_u$. Thus, the relaxation rate k_{obs} that is observed in experiments is the sum of the folding rate k_f and the unfolding rate k_u.

The folding and unfolding rates k_f and k_u of two-state proteins can be determined by measuring the relaxation rate k_{obs} as a function of the denaturant concentration [den]. For two-state proteins, $k_{\mathrm{obs}}([\mathrm{den}])$ exhibits a characteristic V-shape with two linear 'arms' at low and high denaturant concentration (see fig. 1.2(b)). In the 'left arm' of the plot, the relaxation rate $k_{\mathrm{obs}} = k_f + k_u$ is dominated by the folding rate k_f, and in the 'right arm' by the unfolding rate k_u. Because of its characteristic shape, the plot shown in fig. 1.2(b) has been termed 'chevron plot'.

The two linear arms can be understood by assuming that the logarithms of the folding and unfolding rate depend linearly on the denaturant concentration (see, e.g., [16]):

$$\ln k_f([\mathrm{den}]) = \ln k_f(0) - m_f[\mathrm{den}] \tag{1.3}$$
$$\ln k_u([\mathrm{den}]) = \ln k_u(0) + m_u[\mathrm{den}] \tag{1.4}$$

Here, $\ln k_f(0)$ and $\ln k_u(0)$ are the folding and unfolding rates at zero denaturant concentration, and m_f and m_u are proportionality constants. At small denaturant concentrations, the folding equilibrium with constant $K \equiv k_f/k_u$ is shifted towards the native state. The folding rate k_f then is much larger than the unfolding rate k_f, and we have $\ln k_{\mathrm{obs}} = \ln(k_f + k_u) \simeq \ln k_f$ with $\ln k_f$ given in eq. (1.3). The slope of the 'left arm' in the chevron plot of fig. 1.2(b) thus is $-m_f$. At high denaturant concentrations, in contrast, the folding equilibrium is shifted towards the denatured state, and we have $\ln k_{\mathrm{obs}} \simeq \ln k_u$ with $\ln k_u$ given in eq. (1.4). The slope of the 'right arm' in the chevron plot at high denaturant concentration therefore is m_u. The rates $\ln k_f(0)$ and $\ln k_u(0)$ can be obtained by extrapolating both arms to [den]$= 0$. The protein stability, i.e. the

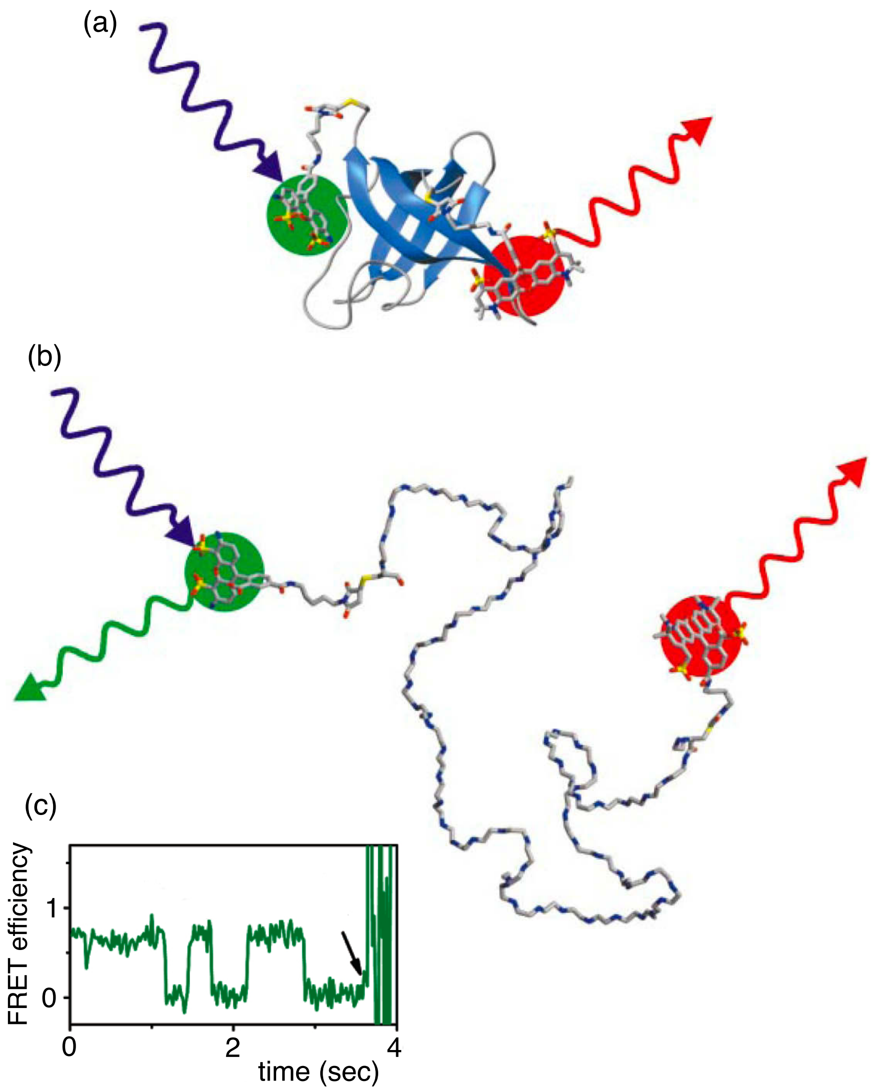

Figure 1.3: (a) and (b) Schematic illustration of a fluorescence resonance energy transfer (FRET) single-molecule experiment on two-state protein folding (adapted from [19]). Two fluorescent dyes are attached to the chain ends of the cold-shock protein. The green donor dye absorbs blue light and either emits green light or transfers the excitation energy to the acceptor dye, which emits red light. The transfer efficiency depends on the distance between the dyes and, thus, on the protein state. In the folded state (a) of the protein, the transfer efficiency is high since the separation of the two dyes is relatively small. In the unfolded state (b), the average separation of the dyes is large, and the transfer efficiency is small. – (c) The time-dependent FRET signal for a single cold-shock protein reveals sudden jumps between the native and the denatured state (adapted from [20]). The arrow indicates the bleaching of one of the dyes.

difference $G_{\text{N-D}} = G_{\text{N}} - G_{\text{D}}$ between the free energy G_{N} of the native state and the free energy G_{D} of the denatured state, follows from [16]

$$G_{\text{N-D}} = -RT \ln K = -RT \ln(k_f/k_u) \tag{1.5}$$

Two-state folding of the cold-shock protein from the bacterium *Thermotoga maritima* has also been investigated in single-molecule experiments [19, 20]. In these experiments, fluorescent donor and acceptor dyes are attached to both ends of the protein (see fig. 1.3). The fluorescence transfer efficiency between the dyes depends on the protein state, since the distance between the dyes is relatively small in the native state and large in the denatured state. The folding and unfolding transitions between the two states are reflected by jumps in transfer efficiency.

1.3 Mutational analysis of the folding kinetics

Two-state processes are often described in classical transition-state theory. For two-state proteins, the folding rate then is assumed to have the form (see, e.g., [16])

$$k_f = k_o \exp[-G_{\text{T-D}}/RT] \tag{1.6}$$

where $G_{\text{T-D}}$ is the free-energy difference between the transition state T and the denatured state D (see fig. 1.4(a)), and k_o is a prefactor that depends on the conformational diffusion coefficient of the protein. Similarly, the unfolding rate k_u is proportional to $\exp[-G_{\text{T-N}}/RT]$, where $G_{\text{T-N}}$ is the free-energy difference between the transition state T and the native state N. From a statistical-mechanical perspective, the transition state is thought to consist of a large number of extremely short-lived transition-state conformations. Each of these transition-state conformations is partially folded and will either complete the folding process or will unfold again, with equal probability [21–23].

Since transition-state conformations are highly instable, they cannot be observed directly. Instead, the folding kinetics of many two-state proteins has been investigated via mutational analysis [24–43]. In a mutational analysis, a large number of mostly single-residue mutants of a protein is generated. For each mutant, the effect of the mutation on the folding dynamics is quantified by its Φ-value [16, 44]

$$\Phi = \frac{RT \ln(k_{\text{wt}}/k_{\text{mut}})}{\Delta G_{\text{N-D}}} \tag{1.7}$$

Here, k_{wt} is the folding rate for the wildtype protein, k_{mut} is the folding rate for the mutant protein, and $\Delta G_{\text{N-D}} = G_{\text{N'-D'}} - G_{\text{N-D}}$ is the change of

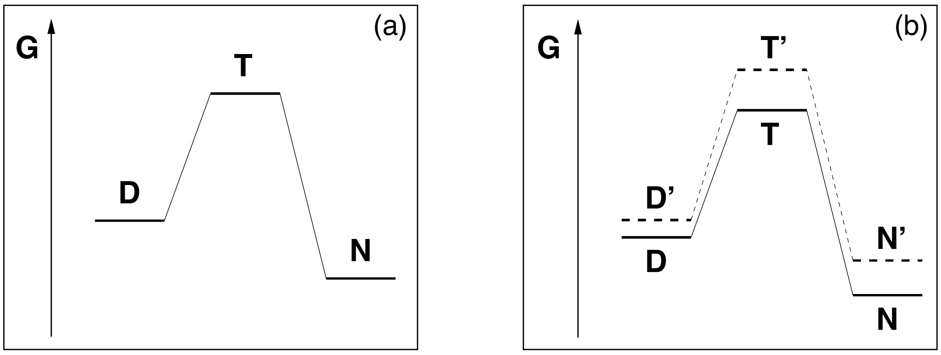

Figure 1.4: (a) In classical transition-state theory, the folding kinetics of a two-state protein is dominated by a transition state T between the denatured state D and the native state N. The folding rate depends on the difference $G_{\text{T-D}} = G_{\text{T}} - G_{\text{D}}$ between the free energy G_{T} of the transition state T and the free energy G_{D} of the denatured state D, see eq. (1.6). – (b) Mutations perturb the free energies of the denatured state, transition state, and native state.

the protein stability induced by the mutation. $G_{\text{N'-D'}}$ and $G_{\text{N-D}}$ denote the stabilities of the mutant and the wildtype, see fig. 1.4(b).

With eq. (1.6), Φ-values can be written in the form

$$\Phi = \frac{\Delta G_{\text{T-D}}}{\Delta G_{\text{N-D}}} \tag{1.8}$$

if one assumes that the pre-exponential factor k_o is not affected by the mutation [16]. Here, $\Delta G_{\text{T-D}} = G_{\text{T'-D'}} - G_{\text{T-D}}$ is the mutation-induced change of the free-energy barrier $G_{\text{T-D}}$, see fig. 1.4(b). The central question is if we can reconstruct the transition state from the observed Φ-values for a large number of mutants.

1.4 Traditional interpretation of Φ-values

In the traditional interpretation, a Φ-value of 1 is interpreted to indicate that the residue has a native-like structure in T (see fig. 1.5(a)), since the mutation shifts the free energy of the transition state T by the same amount as the free energy of the native state N. A Φ-value of 0 is interpreted to indicate that the residue is as unstructured in T as in the denatured state D (see fig. 1.5(b)), since the mutation does not shift the free-energy difference between these two states. Φ-values between 0 and 1 are typically taken to indicate partial native-like structure in

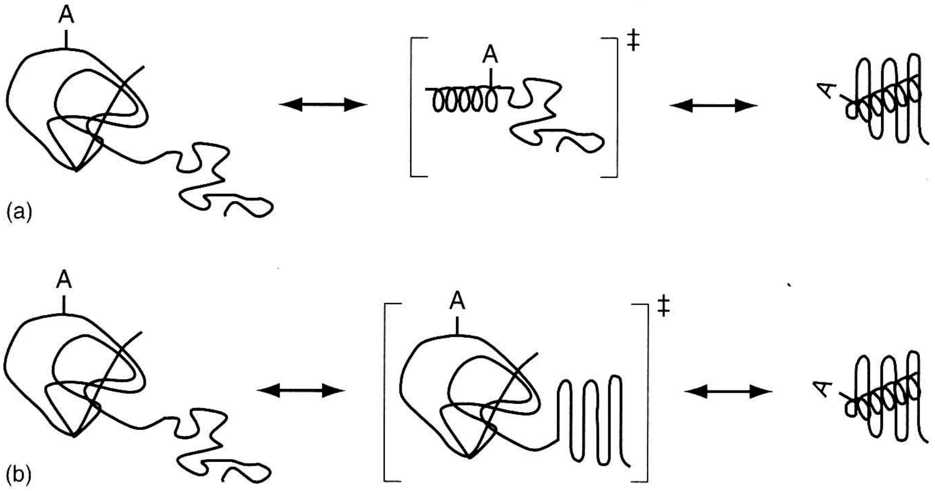

Figure 1.5: Schematic illustration of the traditional interpretation of Φ-values (from [16]). The denatured state of a protein is shown on the left side, the native state on the right side, and the transition state (indicated by $[\ldots]^{\ddagger}$) in the center of the illustrations. The native structure of the protein consists of an α-helix packed against a β-sheet. The mutated residue is residue A in the α-helix. In (a), the helix is fully formed in the transition state T, and residue A is assumed to have a fully native-like structural environment in T. The mutation then shifts the free energy of the transition state by the same amount as the free energy of the native state, which results in a Φ-value of 1 according to eq. (1.4). In (b), the helix is not yet formed in T, and the environment of residue A in T is identical with the environment of A in the denatured state D. The mutation then does not change the free-energy difference $G_{\text{T-D}}$ between the transition state T and the denatured state D (i.e., $\Delta G_{\text{T-D}} = 0$), which results in a Φ-value of 0 for the mutation.

T [16, 45]. In the traditional interpretation, a Φ-value thus is taken to indicate the *degree of structure formation of the mutated residue in the transition-state ensemble T*.

However, this traditional interpretation is often not consistent. First, some Φ-values are negative or larger than 1 [46, 47] and cannot be interpreted as a degree of structure formation. Second, Φ-values are sometimes significantly different for different mutations at a given chain position. The mutations E15D and E15N in the helix of the protein CI2, for example, have Φ-values of 0.22 ± 0.05 and 0.53 ± 0.05 [24], which differ by more than a factor 2 (see also table 2.4 in chapter 2). In the traditional interpretation, Φ-values for different mutations of the same residue are expected to be identical, since they just reflect the degree of structure formation of this residue in T. Third, Φ-values for neighboring residues within a given secondary structure often span a wide range of values. For example, the Φ-values for 12 different mutations in the CI2 helix with stability changes $\Delta G_N > 0.7$ kcal/mol range from -0.25 to 1.06 (see table 2.4). In the traditional interpretation, this means that some of the helical residues are unstructured in the transition state, while other residues, often direct neighbors, are highly structured. This contradicts the notion that secondary structures are cooperative. The formation of helices, for example, requires that several consecutive helical turns are structured, stabilizing each other.

Φ-values provide indirect information on the folding kinetics of a protein and, therefore, have attracted considerable theoretical interest. To understand the experimentally determined Φ-values for a protein, Molecular Dynamics (MD) simulations with atomistic models are often performed [48–65]. However, such simulations are computationally demanding and in general do not allow direct calculations of folding rates and Φ-values. Instead, the MD approaches typically rely on the assumption of the traditional interpretation that Φ-values reflect the degree of structure formation of residues in the transition state T. For example, Φ-values are often calculated from the fraction of contacts a residue forms in the transition state T, compared to the fraction of contacts in the native and the denatured states [48–61]. In an alternative approach, Daggett and coworkers compute an S-value [62], which is "a measure of the amount of structure at a given residue, defined by the amounts of secondary and tertiary structure at each residue" [63]. Exceptions to such structural assumptions are a recent MD study of an ultrafast mini-protein in which Φ-values are calculated from rates for the wildtype and mutants via eq. (1.7) [64], and the calculation of Φ-values from free-energy shifts of the transition-state ensemble using eq. (1.8) [65]. In statistical-mechanical models with simplified energetic interactions, in contrast, folding rates and stabilities for wildtype and mutants can be easily calculated [66–72]. However, the lack of atomistic detail in these models appeared to made make it difficult to reproduce detailed mutational data.

1.5 Overview

In chapter 2 of this thesis, we present models that lead to a novel interpretation of Φ-values from mutational analyses of the folding kinetics. The central assumption of these models is that structural elements such as α-helices and β-hairpins form cooperatively. The structural elements then are either fully formed or not formed in partially folded conformations, in particular in transition-state conformations. A further central aspect of these models is that mutation-induced free-energy changes are split into different components. For mutations in protein helices, the free-energy changes are split into secondary and tertiary free-energy components. In the case of three-stranded β-sheet proteins, which consist of just two β-hairpins, mutations can affect the free-energy contributions of hairpin 1, hairpin 2, or the small hydrophobic core of the proteins. The structural parameters in our models are the degrees to which the structural elements of a protein are formed in the folding transition state. These structural parameters are obtained from fitting to experimental Φ-values. The models can capture negative Φ-values and Φ-values larger than 1, which have been difficult to understand in the traditional interpretation presented in section 1.4. In addition, the models explain how different mutations at a given site can lead to different Φ-values.

Chapter 3 is devoted to loop-closure aspects of the protein folding kinetics. In 1998, Plaxco et al. [73] reported the remarkable observation that the folding rates of two-state proteins correlate with a simple measure of native-state topology, the relative contact order (CO). The CO of a contact (i, j) between two residues i and j simply is the sequence separation $|i - j|$, and the relative contact order is defined as the average CO of all contacts in the native structure, divided by the chain length of the protein. The CO of a contact thus is the length of the loop that has to be closed to form the contact *in the fully unfolded state of the protein chain.* Plaxco et al. found that proteins with small relative CO fold faster than proteins with large relative CO, which seems plausible from loop-closure principles, since small loops close faster than large loops.

The central result of chapter 3 is the extension of this loop-closure principle from folding rates to folding routes. The graph-theoretical concept that enables this extension is the concept of effective contact order (ECO). The ECO is the length of the loop that has to closed to form a contact *in a partially folded chain conformation.* The ECO concept takes into account the contacts that are already present in a partially folded conformation 'short-circuit' the protein chain, which decreases the loop lengths for contacts that are formed subsequently. In contrast to COs, the ECOs depend on the sequence in which contacts are formed and, thus, enable the prediction of folding routes from loop-closure principles. In addition, the ECO concept will be applied in chapter 3 to estimate the folding rates of proteins with covalent crosslinks.

Finally, chapter 4 is focused on cooperativity aspects of protein folding. First, we will investigate the folding cooperativity of two-state proteins in a statistical-mechanical model that includes ECO-based loop-closure dependencies between contacts and structural elements. Second, we will analyze the substructural cooperativity of proteins by quantifying the correlations between contacts in Molecular Dynamics unfolding simulations with an atomistic model. The correlation analysis reveals high correlations predominantly between contacts of the same structural element.

2 Transition states

2.1 A model for small β-sheet proteins

In this section, we present a simple model for the folding kinetics of three-stranded, antiparallel β-sheet proteins. The β-sheet of these proteins consists of just two hairpins, $\beta_1\beta_2$ and $\beta_2\beta_3$, which share the central strand β_2 (see fig. 2.1). Important representatives of this class of proteins are WW domains, named after two conserved tryptophan residues, which are represented by the letter 'W' in the single-letter code for amino acids. Because of their small size and abundance as protein domains, WW domains are important model systems for understanding β-sheet folding and stability. The design principles [74, 75] and folding kinetics [35, 39, 43, 76–83] of WW domains and other three-stranded β-sheet proteins have been studied extensively. The central result of this section is a general formula for Φ-values of three-standed proteins, eq. (2.11). In the next section, this formula will be applied to detailed mutational data for the PIN WW domain and the FBP WW domain, which provides structural information on the transition state of these proteins.

The central assumption of the model is that each of the hairpins is either fully formed or not formed in partially folded states of the protein. The model has then just four states: the denatured state D in which none of the hairpins is formed, a partially folded state in which only hairpin 1 is formed, a partially folded state in which only hairpin 2 is formed, and the native state with both hairpins formed. The energy landscape can be characterized by three free-energy differences: The free-energy difference G_N of the native state and the free-energy differences G_1 and G_2 of the partially folded states with respect to the denatured state (see fig. 2.2).

The folding kinetics is described by the master equation

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} = \sum_{m \neq n} \left[w_{nm} P_m(t) - w_{mn} P_n(t) \right], \tag{2.1}$$

for the time evolution of the probability $P_n(t)$ that the protein is in state n at time t. Here, w_{nm} is the transition rate from state m to n, defined by

$$w_{nm} = \frac{1}{t_o} \left(1 + e^{G_n - G_m} \right)^{-1} \tag{2.2}$$

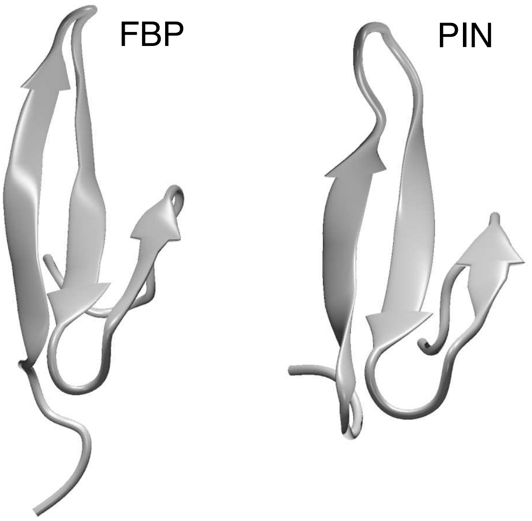

Figure 2.1: The native structures of the FBP [84] and the PIN WW domain [85] consist of two β-hairpins, which form a three-stranded β-sheet. The structural representations have been generated with the programs VMD [86] and Raster3D [87].

provided the states n and m are connected via a transition step in which only a single hairpin folds or unfolds. For other transitions, i.e. for the direct transition from the denatured state to the native state, and vice versa, the transition rates are zero. Here, t_o is a reference time scale. The transition rates defined above obey detailed balance $w_{nm}P_m^e = w_{mn}P_n^e$ where $P_n^e \sim \exp[-G_n/(RT)]$ is the equilibrium weight for the state n. Detailed balance ensures that the system ultimately reaches thermal equilibrium [88].

The master equation (2.1) can be written in the matrix form

$$\frac{\mathrm{d}\boldsymbol{P}(t)}{\mathrm{d}t} = -\boldsymbol{W}\boldsymbol{P}(t) \tag{2.3}$$

The elements of the vector $\boldsymbol{P}(t)$ are the probabilities $P_n(t)$ that the protein is in state n at time t, and the matrix elements of $\boldsymbol{W}$ are given by

$$W_{nm} = -w_{nm} \quad \text{for} \quad n \neq m; \qquad W_{nn} = \sum_{m \neq n} w_{mn}. \tag{2.4}$$

For the model with four states considered here, the matrix $\boldsymbol{W}$ is given by

$$\frac{1}{t_o}\begin{pmatrix} \frac{1}{1+e^{g_1}}+\frac{1}{1+e^{g_2}} & -\frac{1}{1+e^{-g_1}} & -\frac{1}{1+e^{-g_2}} & 0 \\ -\frac{1}{1+e^{g_1}} & \frac{1}{1+e^{-g_1}}+\frac{1}{1+e^{g_N-g_1}} & 0 & -\frac{1}{1+e^{g_1-g_N}} \\ -\frac{1}{1+e^{g_2}} & 0 & \frac{1}{1+e^{-g_2}}+\frac{1}{1+e^{g_N-g_2}} & -\frac{1}{1+e^{g_2-g_N}} \\ 0 & -\frac{1}{1+e^{g_N-g_1}} & -\frac{1}{1+e^{g_N-g_2}} & \frac{1}{1+e^{g_1-g_N}}+\frac{1}{1+e^{g_2-g_N}} \end{pmatrix}$$

To simplify the notation, we have used here dimensionless free-energy differences $g_i \equiv G_i/RT$ (i = 1, 2, or N) of the partially folded states 1 and 2 and the native state N with respect to the denatured state.

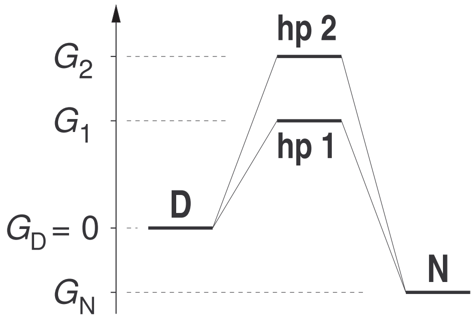

Figure 2.2: Simple energy landscape of the four-state model for WW domains. The four states are the denatured state D, the native state N, and two partially folded states hp 1 and hp 2 in which one of the two hairpins is formed. Here, G_N is the free-energy difference between the native state N and the denatured state D, which has the 'reference free energy' $G_D = 0$, and G_1 and G_2 are the free-energy differences between the partially folded states and the denatured state.

The general solution $\boldsymbol{P}(t)$ of the master equation can be expressed in terms of the eigenvalues λ and eigenvectors $\boldsymbol{Y}_\lambda$ of the matrix $\boldsymbol{W}$:

$$\boldsymbol{P}(t) = \sum_\lambda c_\lambda \boldsymbol{Y}_\lambda \exp[-\lambda t] \tag{2.5}$$

The prefactors c_λ in this general solution depend on the initial conditions at time $t = 0$. For the 4×4 matrix above, the 4 eigenvalues are given by $\lambda = 0$, $1 - q$, $1 + q$, and 2, in units of $1/t_o$, with

$$q \equiv \frac{1 - e^{g_N - g_1 - g_2}}{\sqrt{(1 + e^{-g_1})(1 + e^{-g_2})(1 + e^{g_N - g_1})(1 + e^{g_N - g_2})}} \tag{2.6}$$

Since we have $-1 < q < 1$, the three nonzero eigenvalues are positive and describe the relaxation to the equilibrium state of the model, see eq. (2.5). The equilibrium probability distribution is $c_o \boldsymbol{Y}_o$ where $\boldsymbol{Y}_o$ is the eigenvector with eigenvalue 0.

This model exhibits two-state folding kinetics under two conditions. First, the native state has to be stable, i.e. the free energy g_N of the native state must be significantly smaller than the free energies of the other three states. Second, the free-energy differences g_1 and g_2 between the partially folded states and the denatured state have be to significantly larger than RT. The partially folded states then constitute the transition-state ensemble. Under these two conditions, the three Boltzmann weights $e^{g_N - g_1 - g_2}$, $e^{g_N - g_1}$, and $e^{g_N - g_2}$ in eq. (2.6) are much smaller than 1, and also much smaller than e^{-g_1} and e^{-g_2}, which leads to

$$q \simeq \frac{1}{\sqrt{(1 + e^{-g_1})(1 + e^{-g_2})}} \tag{2.7}$$

For large barrier energies g_1 and g_2, we have $e^{-g_1} \ll 1$ and $e^{-g_2} \ll 1$, and therefore $(1+e^{-g_1})(1+e^{-g_2}) \simeq (1+e^{-g_1}+e^{-g_2})$. If we now use the expansion $(1+x)^{-1/2} \simeq 1-x/2$ with $x = e^{-g_1}+e^{-g_2} \ll 1$, the smallest nonzero relaxation rate, or folding rate, $k \equiv 1-q$ is given by

$$k(G_1,G_2) \simeq \frac{1}{2}\left(e^{-g_1}+e^{-g_2}\right) = \frac{1}{2}\left(e^{-G_1/RT}+e^{-G_2/RT}\right) \tag{2.8}$$

in units of $1/t_o$. The folding rate k simply is the sum of the rates for the two possible folding routes on which either hairpin 1 or hairpin 2 forms first. The factor $\frac{1}{2}$ in the equation above arises because a molecule, after reaching one of the barrier states 1 or 2, either proceeds to N or returns to D, with almost equal probability. The folding rate k is much smaller than the other two relaxation rates $1+q$ and 2, which reflect an initial, fast 'burst phase'.

Mutations correspond to perturbations of the free-energy landscape. A mutation therefore can be characterized by the free-energy changes ΔG_1, ΔG_2, and ΔG_N. The folding rate of the mutant then is $k_{\text{mut}} \equiv k(G_1+\Delta G_1, G_2+\Delta G_2)$. For small perturbations ΔG_1 and ΔG_2, a Taylor expansion of $\ln k_{wt} \equiv \ln k$ to first order leads to

$$\ln k_{\text{mut}} - \ln k_{\text{wt}} \simeq \frac{\partial \ln k}{\partial G_1}\Delta G_1 + \frac{\partial \ln k}{\partial G_2}\Delta G_2 = -\frac{1}{RT}\left(\chi_1 \Delta G_1 + \chi_2 \Delta G_2\right) \tag{2.9}$$

with

$$\chi_1 \equiv \frac{e^{-G_1/RT}}{e^{-G_1/RT}+e^{-G_2/RT}} \quad \text{and} \quad \chi_2 \equiv \frac{e^{-G_2/RT}}{e^{-G_1/RT}+e^{-G_2/RT}} \tag{2.10}$$

The two parameters χ_1 and χ_2 quantify the extent to which the partially folded state 1 and the partially folded state 2 are populated in the transition-state ensemble. From the Φ-value definition (1.7) and eq. (2.9), we obtain the general form [89]

$$\boxed{\Phi = \frac{\chi_1 \Delta G_1 + \chi_2 \Delta G_2}{\Delta G_N}} \tag{2.11}$$

of Φ-values for mutations in three-stranded β-sheet proteins. Different Φ-values for different mutations arise from characteristic 'free-energy signatures' ΔG_1, ΔG_2, and ΔG_N of the mutations. The structural parameters χ_1 and χ_2 of the model, in contrast, are independent of the mutation and characterize the degree to which hairpin 1 and hairpin 2 are structured in the transition-state ensemble.

2.2 The transition states of the FBP and PIN WW domains

In this section, we test eq. (2.11) derived in the previous section and determine the structural parameters χ_1 and χ_2 for the FBP WW domain and the PIN WW domain by fitting to experimental data. We first consider the FBP WW domain. Petrovich et al. [43] have performed an extensive mutational analysis of the folding kinetics. The Φ-values and stability changes ΔG_N for the considered mutations are summarized in table 2.1, together with an assessment which structural elements are affected by the mutations. This assessment is based on the contact matrix of the FBP WW domain shown in fig. 2.3. A black dot at position (i, j) of this matrix indicates that the two amino acids i and j are in contact, i.e. that the distance between any of their non-hydrogen atoms is smaller than the cutoff distance 4 Å. Since the contact matrix is symmetric, only one half is represented in fig. 2.3. The two contact clusters in the matrix correspond to hairpin 1 and hairpin 2 of the FBP WW domain. The remaining contacts largely correspond to contacts of hydrophobic amino acids, the small hydrophobic core of the protein. About half of the mutations performed by Petrovich et al. affect only either hairpin 1 or hairpin 2. The mutation E7A of amino acid 7, for example, affects the contacts $(7, 22)$, $(7, 23)$, and $(7, 24)$, which are all located in hairpin 1 (see contact map in fig. 2.3). The remaining mutations also affect the hydrophobic core, or both hairpins. The mutation Y21A, for example, affects the contacts $(8, 21)$ and $(9, 21)$ in hairpin 1, and the contacts $(21, 26)$, $(21, 27)$, and $(21, 28)$ in hairpin 2.

To test our model, we first consider all mutations that affect only one of the hairpins. The model predicts that all mutations that affect only hairpin 1 should have the same Φ-value χ_1, and all mutations that affect only hairpin 2 the same Φ-value χ_2. This is a direct consequence of eq. (2.11). For mutations that affect only hairpin 1, for example, we have $\Delta G_2 = 0$ since the mutations don't shift the stability of hairpin 2, and $\Delta G_N = \Delta G_1$ since they also don't affect the hydrophobic core. Eq. (2.11) then results in $\Phi = \chi_1$ for these mutations. The Φ-values for the 10 mutations that only affect hairpin 1 are plotted in fig. 2.4. Except for one outlier, all Φ-values are centered around the value 0.8, mostly within experimental errors. The mean value of these nine Φ-values (dashed line in fig. 2.4) leads to the estimate $\chi_1 = 0.81 \pm 0.06$. The error here is estimated as error of the sample mean. The standard deviation of the Φ-values from the the mean value is 0.18. The four Φ-values for mutations that affect only hairpin 2 range from 0.08 to 0.39 (see table 2.1), with mean value $\chi_2 = 0.30 \pm 0.08$ and standard deviation

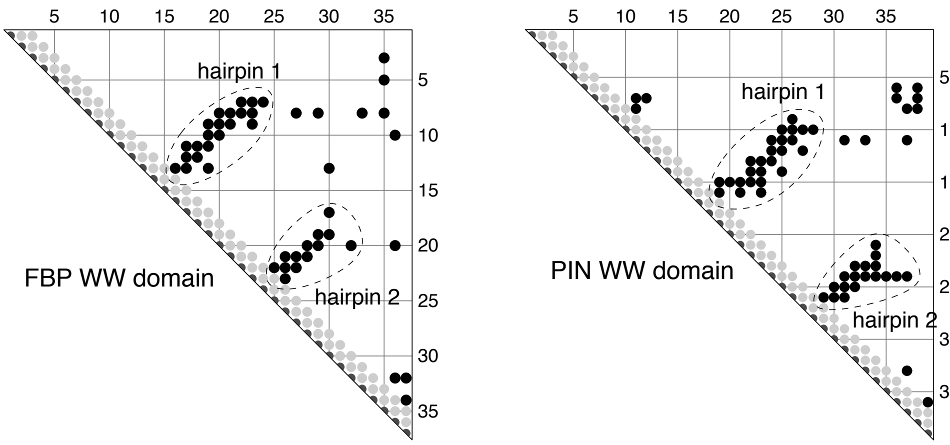

Figure 2.3: Contact matrices of the FBP and PIN WW domains. A black dot at position (i, j) of a matrix indicates that the residues i and j are in contact. Two residues are defined here to be in contact if the distance between any of their non-hydrogen atoms is smaller than the cutoff distance 4 Å. The hairpins 1 and 2 of the WW domains correspond to clusters of contacts.

0.16. For both sets of mutations, we thus obtain good agreement with the model. In addition, the sum of the above estimated values for the parameters χ_1 and χ_2 is close to 1, within the error bounds, which is an additional consistency requirement of the model. The two parameters χ_1 and χ_2 are the fractions to which the two transition-state conformations with either hairpin 1 or hairpin 2 formed are populated. These fractions sum up to 1 since the protein has to take one of the two possible routes (see fig. 2.2).

To include other mutations in the model, we have to estimate the impact of these mutations on the stability of the different structural elements they affect (hairpin 1, hairpin 2, or the hydrophobic core). We use FOLD-X here, a force field with entropic terms that has been optimized for the prediction of mutation-induced stability changes [90, 91]. We calculate the mutation-induced stability changes ΔG_N for the whole protein, and the stability changes ΔG_1 and ΔG_2 of hairpin 1 and 2, depending on whether the considered mutation affects these hairpins. To calculate ΔG_1 and ΔG_2, we simply 'cut out' these hairpins from the PDB structure and estimate the stability of the wildtype and mutant hairpins with FOLD-X (see caption of table 2.2 for details). The resulting data are summarized in table 2.2. The calculated stability changes ΔG_N can be directly compared to the experimentally measured stability changes

Table 2.1: Mutational data for the FBP WW domain

mutation	Φ_{exp}	$\Delta G_{N,\text{exp}}$	affected struct. elements
E7A	0.67 ± 0.21	0.52 ± 0.16	hairpin 1
W8F	0.24 ± 0.03	1.65 ± 0.16	hairpin 1, hydrophobic core
T9A	-0.09 ± 0.04	0.93 ± 0.09	hairpin 1
T9G	0.94 ± 0.20	0.50 ± 0.10	hairpin 1
Y11A	0.55 ± 0.10	0.63 ± 0.11	hairpin 1
T13A	-0.03 ± 0.07	0.81 ± 0.17	hairpin 1, hydrophobic core
T13G	-0.32 ± 0.25	0.58 ± 0.22	hairpin 1, hydrophobic core
A14G	0.69 ± 0.28	0.50 ± 0.22	hairpin 1
D15A	0.82 ± 0.16	0.42 ± 0.09	hairpin 1
D15G	0.77 ± 0.17	0.39 ± 0.09	hairpin 1
G16A	1.17 ± 0.22	1.33 ± 0.27	hairpin 1
T18A	0.93 ± 0.27	0.54 ± 0.17	hairpin 1
T18G	0.73 ± 0.05	1.14 ± 0.09	hairpin 1
Y19A	0.11 ± 0.05	0.67 ± 0.13	hairpin 1 and 2
Y20F	0.05 ± 0.16	0.68 ± 0.18	hairpin 1 and 2, hydroph. core
Y21A	0.28 ± 0.02	1.70 ± 0.10	hairpin 1 and 2
R24A	0.29 ± 0.09	0.78 ± 0.17	hairpin 1 and 2
T25A	0.39 ± 0.04	2.51 ± 0.18	hairpin 2
T25S	0.27 ± 0.03	1.08 ± 0.09	hairpin 2
L26A	0.08 ± 0.08	0.56 ± 0.12	hairpin 2
L26G	0.45 ± 0.04	-1.29 ± 0.10	hairpin 2
E27A	0.12 ± 0.04	1.02 ± 0.13	hairpin 2, hydrophobic core
T29G	0.09 ± 0.02	1.89 ± 0.11	hairpin 2, hydrophobic core
W30A	0.19 ± 0.06	0.76 ± 0.14	hairpin 2, hydrophobic core
L36A	-0.30 ± 0.16	0.91 ± 0.14	hairpin 2, hydrophobic core
L36V	-0.13 ± 0.09	0.53 ± 0.14	hairpin 2, hydrophobic core

Experimental Φ-values and stability changes $\Delta G_{N,\text{exp}}$ are from Petrovich et al. [43]. The information on the structural elements affected by the mutations is derived from the contact map shown in fig. 2.3 (see text).

$\Delta G_{N,\text{exp}}$. We include here only mutations in the model for which the FOLD-X predicted stability changes ΔG_N do not differ by more than a factor 2 from the experimental stability changes $\Delta G_{N,\text{exp}}$. For other mutations, the force-field calculations are unreliable. In table 2.2, the calculated stability changes for these mutations are shown in brackets.

The mutations in table 2.2 affect two of the structural elements: The mutations W8F and T13A affect hairpin 1 and the hydrophobic core. For

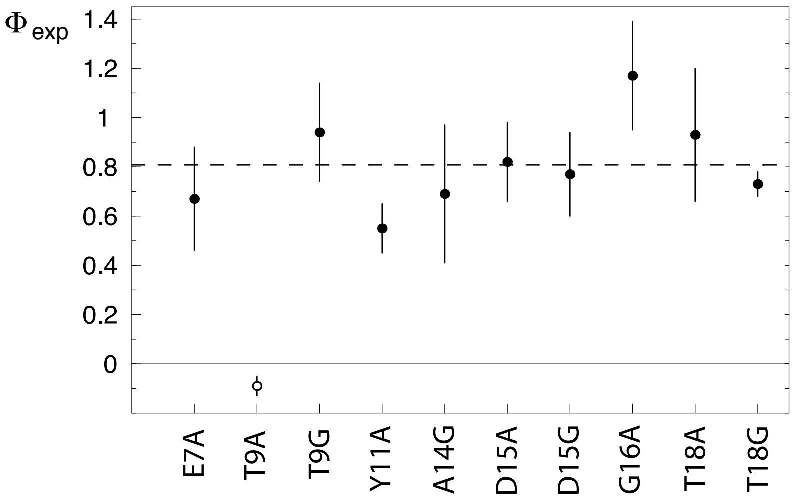

Figure 2.4: Φ-values for mutations that only affect haipin 1 of the FBP WW domain (see table 2.1). Except for one outlier (open circle for mutation T9A), the Φ-values are centered around the mean value 0.81 ± 0.06, with deviations mostly within the estimated experimental errors [43].

these mutations, we have $\Delta G_2 = 0$, and $\Phi = \chi_1 \Delta G_1/\Delta G_N$ according to eq. (2.11). The mutation Y21A affects both hairpins, hence $\Phi = (\chi_1 \Delta G_1 + \chi_2 \Delta G_2)/(\Delta G_1 + \Delta G_2)$. Finally, the mutations T29G, W30A, and L36V affect hairpin 2 and the hydrophobic core. Therefore, we have $\Delta G_1 = 0$ for these mutations, and $\Phi = \chi_2 \Delta G_2/\Delta G_N$.

Let us now consider the set of 20 mutations that consists of these 6 mutations that affect two structural elements and the 14 mutations that affect either only hairpin 1 or only hairpin 2. Our model has two parameters, χ_1 and χ_2. However, since $\chi_1 + \chi_2 = 1$, there is only one independent parameter. We determine this parameter from a least-square fit between the theoretical Φ-value formula given in eq. (2.11) and the experimental Φ-values and obtain the values $\chi_1 = 0.77 \pm 0.05$ and $\chi_2 = 0.23 \pm 0.05$, see fig. 2.5.

In the model, the magnitude of a Φ-value depends on which structural elements are affected, and on the mutation-induced free-energy changes of these elements. The mutation E7A of the FBP WW domain, for example, has a relatively large Φ-value since this mutation only affects hairpin 1, which is structured in the dominant substate 1 of the transition-state ensemble, whereas the mutation W8F has a relatively small Φ-value since the mutation mainly affects the free energy of the small hydrophobic core, which is not yet formed in the transition state. The model also

Table 2.2: Experimental and calculated stability changes for mutations of the FBP WW domain that affect several structural elements

mutation	$\Delta G_{N,\mathrm{exp}}$	ΔG_N	ΔG_1	ΔG_2
W8F	1.65 ± 0.16	2.39	0.21	–
T13A	0.81 ± 0.17	0.69	0.22	–
T13G	0.58 ± 0.22	(1.28)	(0.56)	–
Y19A	0.67 ± 0.13	(2.65)	(1.60)	(1.01)
Y20F	0.68 ± 0.18	(-0.76)	(0.31)	(-0.45)
Y21A	1.70 ± 0.10	2.58	0.56	1.42
R24A	0.78 ± 0.17	(-0.23)	(-0.31)	(-0.38)
E27A	1.02 ± 0.13	(0.17)	–	(0.17)
T29G	1.89 ± 0.11	1.47	–	1.14
W30A	0.76 ± 0.14	1.32	–	0.53
L36A	0.91 ± 0.14	0.47	–	-0.30
L36V	0.53 ± 0.14	(0.23)	–	(-0.34)

Experimental data for the stability changes $\Delta G_{N,\mathrm{exp}}$ are from Petrovich et al. [43]. The stability changes ΔG_N, ΔG_1, and ΔG_2 for the whole protein and hairpin 1 or 2, respectively, have been calculated with the program FOLD-X [90, 91]. For mutations to alanine (A) or glycine (G) and the muation W8F, native structures for the mutant proteins have been generated by truncation of atoms. For the mutations Y20F and L36V, mutant structures were generated with the program WHAT IF [92]. The wildtype structure used in the calculations is model 1 of the PDB structure 1E0L [84]. To calculate ΔG_1 and ΔG_2, substructures consisting of the residues 1 to 24 and 15 to 37 of the PDB structure have been used. The FOLD-X calculations have been performed at the ionic strength 150 mM and temperature 283 K of the experiments [43]. Numbers in brackets indicate that the calculated stability changes are not reliable since ΔG_N differs by more than a factor 2 from $\Delta G_{N,\mathrm{exp}}$.

reproduces the negative Φ-value of the mutation L36A, which results from different signs of the mutation-induced free-energy changes ΔG_1 and ΔG_N in table 2.2. According to the free-energy calculations with FOLD-X, the mutation stabilizes hairpin 1 ($\Delta G_1 < 0$), but has an overall destabilizing effect ($\Delta G_N > 0$) since it destabilizes the hydrophobic core.

Mutational analyses of the PIN WW domain's folding kinetics have been performed by Jäger et al. [35] and Deechongkit et al. [39]. While Jäger et al. have considered standard single-site amino-acid replacements, Deechongkit et al. synthesized amid-to-ester mutants that specifically

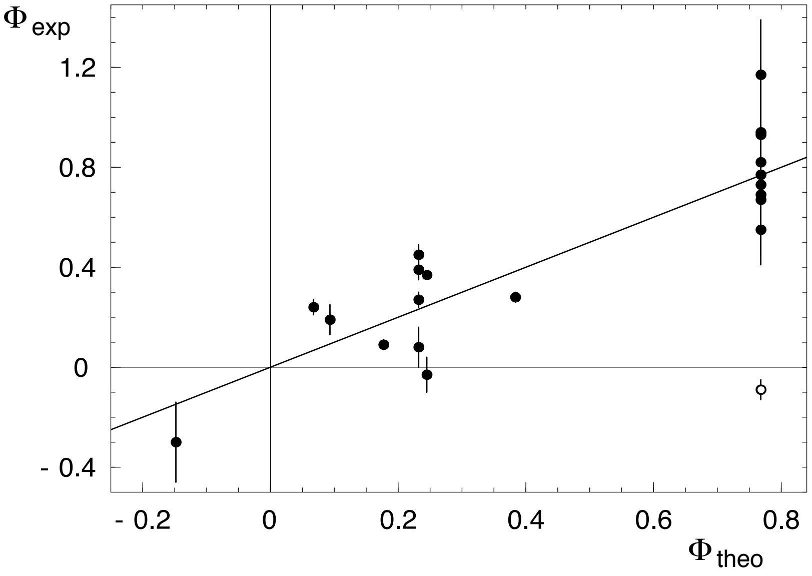

Figure 2.5: Experimental versus theoretical Φ-values obtained from a least-square fit of eq. (2.11) with the single fit parameter χ_1. From this fit, we obtain the values $\chi_1 = 0.77 \pm 0.05$ and $\chi_2 = 1 - \chi_1 = 0.23 \pm 0.05$ for the fractions of the two transition-state conformations in which either hairpin 1 or hairpin 2 are formed. The Pearson correlation coefficient between theoretical and experimental Φ-values is $r = 0.90$ if the outlier data point for mutation T9A (open circle) is not considered, and $r = 0.77$ if the outlier is included.

perturb backbone H-bonds. The experimental Φ-values and stability changes $\Delta G_{N,\mathrm{exp}}$ for these mutations are summarized in table 2.3. The synthetic amino acids in the mutations of Deechongkit et al. are denoted by lowercase greek letters (last six lines in table 2.3). Since these mutations perturb the backbone H-bonds, they only affect either hairpin 1 or hairpin 2, which is indicated in the last column in table 2.3. For the mutations considered by Jäger et al., the affected structural elements are again assessed based on the contact map shown in fig. 2.3. We consider here only mutations with stability changes $\Delta G_{N,\mathrm{exp}} > 0.8$ kcal/mol. Φ-values of mutations that cause significantly smaller stability changes are often considered as unreliable [40,45,93] (see also discussion on page 27).

Seven mutations in table 2.3 affect only hairpin 1 of the PIN WW domain. The mean value of the Φ-values for these mutations leads to the estimate $\chi_1 = 0.69 \pm 0.05$. The standard deviation of the Φ-values from the mean is 0.12, which is comparable to the experimental errors. The four Φ-values of the mutations that affect only hairpin 2 have the mean

Table 2.3: Mutational data for the PIN WW domain

mutation	Φ_{exp}	$\Delta G_{N,\text{exp}}$	affected structural elements
L7A	0.18 ± 0.07	2.06	hydrophobic core
R14F	0.68 ± 0.11	1.29	hairpin 1
M15A	0.63 ± 0.14	0.90	hairpin 1
Y23L	0.64 ± 0.08	1.51	hairpin 1 and 2
Y24F	0.52 ± 0.14	0.87	hairpin 1 and 2
F25L	0.49 ± 0.08	1.69	hairpin 1 and 2
N26D	0.33 ± 0.05	2.13	hairpin 1 and 2
T29D	0.30 ± 0.07	1.77	hairpin 2
A31G	0.44 ± 0.06	1.88	hairpin 2, hydrophobic core
W34A	0.36 ± 0.13	1.12	hairpin 2
K13κ	0.50 ± 0.05	1.00	hairpin 1
S16σ	0.70 ± 0.05	1.39	hairpin 1
R17ρ	0.78 ± 0.11	0.74	hairpin 1
S19σ	0.83 ± 0.04	2.03	hairpin 1
H27η	0.28 ± 0.03	1.77	hairpin 2
S32σ	0.51 ± 0.03	1.77	hairpin 2

Experimental Φ-values and stability changes $\Delta G_{N,\text{exp}}$ for the mutations L7A to W34A are from Jäger et al. [35], and for the amid-to-ester mutants K13κ to S32σ from Deechongkit et al. [39]. Here, only mutations with stability change $\Delta G_{N,\text{exp}} > 0.8$ kcal/mol are considered. The structural elements affected by the mutations are assessed from the contact map shown in fig. 2.3. These structural elements are the hairpin 1 (hp 1), hairpin 2 (hp 2), and the hydrophobic core (hc) of the protein (see text).

value $\chi_2 = 0.36 \pm 0.05$ and the standard deviation 0.10. In agreement with our model, these estimates for χ_1 and χ_2 again add up to 1, within the statistical errors. In an alternative approach, the values of χ_1 and χ_2 can be obtained from a least-square fit between theoretical and experimental Φ-values (see fig. 2.6). From the fit, we obtain $\chi_1 = 0.67 \pm 0.05$ and $\chi_2 = 1 - \chi_1 = 0.33 \pm 0.05$, and a Pearson correlation coefficient of 0.85 between theoretical and experimental Φ-values.

We do not include mutations that affect more than one structural element here since the stability changes estimated with FOLD-X appear to be unreliable. For four of the five mutants, the calculated stability changes ΔG_N differ by significantly more than a factor 2 from the experimental values $\Delta G_{N,\text{exp}}$ (data not shown). The stabilities for the PIN WW domain mutants may be more difficult to calculate since they in-

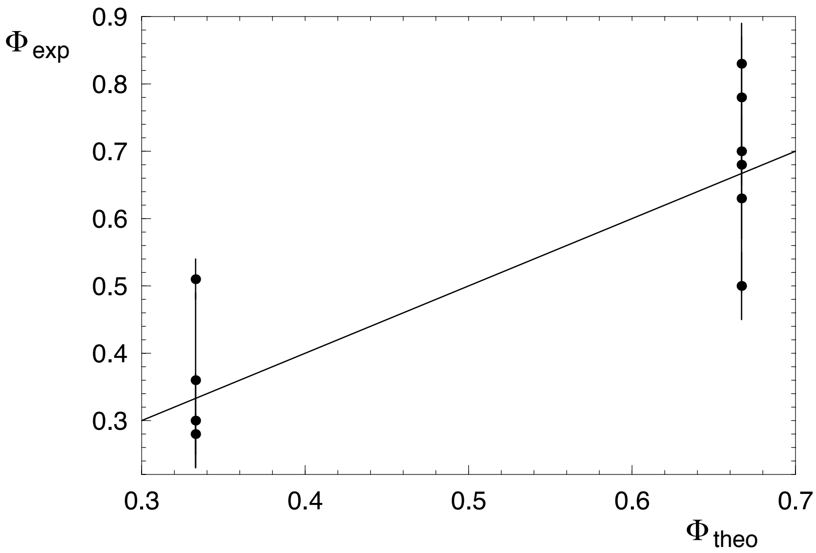

Figure 2.6: Experimental versus theoretical Φ-values obtained from a least-square fit of eq. (2.11), which results in the values $\chi_1 = 0.67 \pm 0.05$ and $\chi_2 = 1 - \chi_1 = 0.33 \pm 0.05$ for the fractions of the two transition-state conformations. The Pearson correlation coefficient between theoretical and experimental Φ-values is $r = 0.85$.

volve a larger range of amino acids, compared to the FBP WW mutants that mostly involve changes to the small amino acids Alanine or Glycine, which can be modeled via simple truncation of sidechains prior to the FOLD-X calculations.

The deviations between experimental and theoretical Φ-values are within reasonable errors. It has been recently suggested that experimental errors for Φ-values may be underestimated since it is usually assumed that the errors in the measured free-energy changes of the transition state and the folded state are independent, which is not the case [94]. In case of the PIN WW domain, we have only considered mutations with stability changes $\Delta G_N > 0.8$ kcal/mol. For mutations that induce significantly smaller stability changes, experimental errors in ΔG_N may lead to large errors in Φ-values since ΔG_N constitutes the denominator of the Φ-value defined in eq. (1.7). However, the large Φ-values up to 1.8 for three mutations with small stability changes in the turn of hairpin 1 of the PIN WW domain [35], which have not been considered, may also result from structural changes in the denatured or native state, which is beyond the simple model presented here.

We have modeled Φ-values from extensive mutational analyses of two WW domains based on the central assumption that the transition-state ensemble of these proteins consists of two substates in which either hairpin 1 or hairpin 2 are formed. The structural information obtained from the mutational data by fitting a single model parameter is that the transition-state ensemble of the FBP WW domain consists to roughly $\frac{3}{4}$ of substate 1 with hairpin 1 formed, and to $\frac{1}{4}$ of substate 2 with hairpin 2 formed. The transition-state ensemble of the PIN WW domain consists to roughly $\frac{2}{3}$ of substate 1, and to $\frac{1}{3}$ of substate 2, according to the model.

2.3 Modeling mutational data for α-helices

In this section, we present a simple model for the formation of α-helices during protein folding. The central result of this section is a general formula, eq. (2.18), for Φ-values of mutations in α-helices of proteins. In the next section, this formula will be applied to detailed mutational data for several α-helices.

The model has two main ingredients. First, the central assumption is that helices are either fully formed or not formed in partially folded conformations, in particular in transition-state conformations. The transition state is described as an ensemble of M different conformations (see fig. 2.7). Each transition-state conformation is directly connected to the native state N and to the denatured state D. The model thus has M parallel folding and unfolding routes.

Second, mutation-induced free-energy changes are split into two components. The overall stability change ΔG_N is split into the change in intrinsic helix stability ΔG_α, and the change in tertiary free energy ΔG_t caused by the mutation:

$$\Delta G_N = \Delta G_\alpha + \Delta G_t \tag{2.12}$$

The intrinsic helix stability G_α is the stability of the 'isolated' helix, i.e. the free-energy difference between the folded and the unfolded state of the helix, in the absence of any tertiary interactions with other structural elements. Similarly, we decompose each ΔG_m, the mutation-induced free-energy change for the transition-state conformation m, into two terms:

$$\Delta G_m = s_m \Delta G_\alpha + t_m \Delta G_t \tag{2.13}$$

Here, G_m is the free-energy difference between transition-state conformation m and the denatured state. Because we assume cooperative formation of the helix, s_m is either 0 or 1, depending on whether the helix is formed or not in the transition-state conformation m. The coefficient

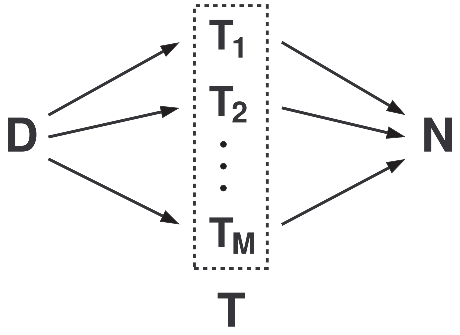

Figure 2.7: In our model, the transition-state ensemble T consists of M transition-state conformations T_1, T_2, …, T_M. The arrows indicate the folding direction from the denatured state D to the native state N via the transition-state conformations.

t_m is between 0 and 1 and represents the degree of tertiary structure formation in conformation m.

We assume that the free-energy barrier for each transition-state conformation is significantly larger than the thermal energy, i.e. that $G_m/RT \gg 1$ [19, 95]. The rate of folding along each route m is then proportional to $\exp[-G_m/RT]$, and the total folding rate as the sum over all the parallel routes is

$$k = c\sum_{m=1}^{M} e^{-G_m/RT} \tag{2.14}$$

where c is a constant prefactor. For $M = 2$, eq. (2.14) with $c = 0.5$ agrees with eq. (2.8) derived in section 2.1. We also assume that the protein is stable, i.e. that G_N, the free-energy difference between the native and the denatured state, is negative.

The folding rate for the mutant protein i is $k_{\rm mut} = k\big(G_1 + \Delta G_1, G_2 + \Delta G_2, \ldots, G_M + \Delta G_M\big)$ with k given in eq. (2.14). The folding rate of the wildtype is $k_{\rm wt} = k(G_1, G_2, \ldots, G_M)$. We assume here that the mutations do not affect the prefactor c in eq. (2.14). For small values $|\Delta G_m|$ of the mutation-induced free-energy changes, a Taylor expansion of $\ln k_{\rm mut}$ leads to

$$\ln k_{\rm mut} - \ln k_{\rm wt} \simeq \sum_{m=1}^{M} \frac{\partial \ln k_{\rm wt}}{\partial G_m}\Delta G_m = -\frac{1}{RT}\frac{\sum_m \Delta G_m e^{-G_m/RT}}{\sum_m e^{-G_m/RT}} \tag{2.15}$$

With the decomposition of the ΔG_m's in eq. (2.13), we obtain

$$\ln k_{\rm mut} - \ln k_{\rm wt} \simeq -\frac{1}{RT}\left(\chi_\alpha \Delta G_\alpha + \chi_t \Delta G_t\right) \tag{2.16}$$

with the two terms

$$\chi_\alpha \equiv \frac{\sum_m s_m e^{-G_m/RT}}{\sum_m e^{-G_m/RT}} \quad \text{and} \quad \chi_t \equiv \frac{\sum_m s_t e^{-G_m/RT}}{\sum_m e^{-G_m/RT}}\ . \tag{2.17}$$

The term χ_α represents the Boltzmann-weighted average of the secondary structure parameter s_m for the transition-state ensemble T. χ_α ranges from 0 to 1 and indicates the average degree of structure formation for the helix in T. The value $\chi_\alpha = 1$ indicates that the helix is formed in all transition-state conformations m, and $\chi_\alpha = 0$ indicates that the helix is formed in none of the transition-state conformations. Values of χ_α between 0 and 1 indicate that the helix is formed in some of the transition-state conformation, and not formed in others. The term χ_t represents the Boltzmann-weighted average of the tertiary structure parameter t_m in T, and also ranges from 0 to 1.

From eq. (2.16) and the definition in eq. (1.7), we then obtain the general form [96]

$$\boxed{\Phi = \frac{\chi_\alpha \Delta G_\alpha + \chi_t \Delta G_t}{\Delta G_N} = \chi_t + (\chi_\alpha - \chi_t)\,\frac{\Delta G_\alpha}{\Delta G_N}} \tag{2.18}$$

of Φ-values for mutations in helices. The second expression simply results from replacing ΔG_t by $\Delta G_N - \Delta G_\alpha$, see eq. (2.12). The two parameters χ_α and χ_t of our model are 'collective' structural parameters for all mutations in the helix. Different Φ-values simply result from different free-energetic 'signatures' ΔG_α and ΔG_N of the mutations. In particular, eq. (2.18) captures that different mutations of the same residue can lead to different Φ-values, and that Φ-values can be 'nonclassical', i.e. < 0 or > 1. Since the two structural parameters χ_α and χ_t range between 0 and 1, a nonclassical Φ-value implies that the changes ΔG_α and ΔG_t in secondary and tertiary free energy caused by the mutation have opposite signs.

To apply our model, we first estimate ΔG_α, the change in helical stability, for each mutation in a particular helix, using standard helix propensity methods (see next section). We then plot all experimental values for Φ versus $\Delta G_\alpha/\Delta G_N$, and obtain the two structural parameters χ_α and χ_t from a linear fit of eq. (2.18). In principle, the two structural parameters can be extracted if Φ-values and stability changes for at least two mutations in a helix are available. However, to test our model, and to obtain reliable values for χ_α and χ_t, we focus here on helices for which more than 10 Φ-values have been determined. The modeling quality then can be assessed from the standard deviation of the data points from the regression line, and from the Pearson correlation coefficients between Φ and $\Delta G_\alpha/\Delta G_N$. Our model can be applied to all mutations for a helix, or to a subset of mutations that affect only the tertiary interactions with one other structural element.

In principle, the parameter χ_t for the tertiary interactions can also be seen to depend on the residue position. To derive eq. (2.18), we don't have to assume that the tertiary parameters t_m for the m transition-state conformations are independent of the residue position and/or mutation.

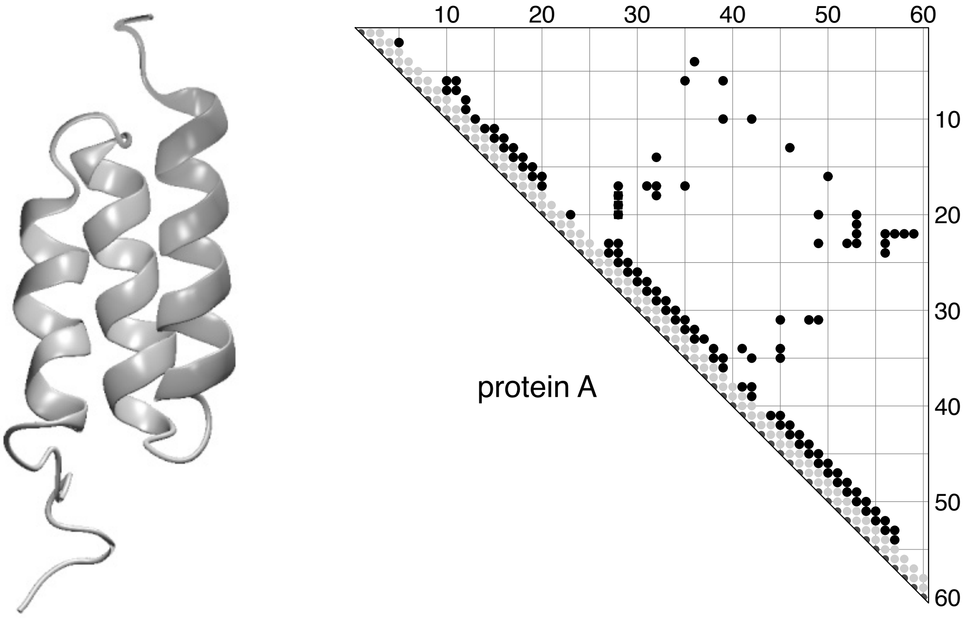

Figure 2.8: (Left) The structure of protein A consists of three helices. – (Right) Contact matrix of protein A. A black dot at position (i, j) of the matrix indicates that the two non-neighboring residues i and j are in contact in the native structure (protein data bank file 1SS1, model 1). Two residues here are defined to be in contact is the distance between any of their non-hydrogen atoms is smaller than the cutoff distance 4 Å. Protein A is an α-helical protein with three helices. Helix 1 consists of the residues 10 to 19, helix 2 of the residues 25 to 37, and helix 3 of the residues 42 to 56.

However, we focus here on the simplest version of our model and show that a consistent structural interpretation of experimental Φ-values in a helix can be obtained with just two structural parameters χ_α and χ_t for the whole helix, which implies a cooperativity of secondary as well as tertiary interactions.

2.4 The helices of protein A and CI2

More than 20 two-state proteins with α/β [24–27, 31, 33, 34, 36, 41, 42, 97, 98], α-helical [28, 38, 99, 100], or all-β structures [29, 30, 32, 35, 37, 40, 101, 102] have been investigated by mutational analysis in the past few years. Mutational data are also available for several proteins that fold via intermediates [103–105] or apparent intermediates [106]. We focus

Table 2.4: Mutational data for the helix of the protein CI2

mutation	Φ	ΔG_N	$\Delta G_\alpha^{\text{AGADIR}}$	$\Delta G_\alpha^{\text{prop}}$
S12G	0.29	0.8	0.28	–
S12A	0.43	0.89	0.14	–
E15D	0.22	0.74	0.13	0.29
E15N	0.53	1.07	0.57	0.25
A16G	1.06	1.09	0.82	1.0
K17G	0.38	2.32	0.80	0.74
K18G	0.7	0.99	0.75	0.74
I20V	0.4	1.3	0.14	0.2
L21A	0.25	1.33	-0.01	-0.21
L21G	0.35	1.38	0.26	0.79
D23A	-0.25	0.96	-0.41	–
K24G	0.1	3.19	0.12	–

Experimental Φ-values and stability changes ΔG_N are from Itzhaki et al. [24]. The change in intrinsic helix stability $\Delta G_\alpha^{\text{AGADIR}}$ is calculated with AGADIR [107–109], see Merlo et al. [110]. The change in intrinsic helix stability $\Delta G_\alpha^{\text{prop}}$ is calculated from the helix-propensity scale of Pace and Scholtz [111]. The helix propensities of the residues are (in kcal/mol): Ala (A) 0, Leu (L) 0.21, Arg (R) 0.21, Met (M) 0.24, Lys (K) 0.26, Gln (Q) 0.39, Glu (E) 0.40, Ile (I) 0.41, Trp (W) 0.49, Ser (S) 0.50, Tyr (Y) 0.53, Phe (F) 0.54, Val (V) 0.61, His (H) 0.61, Asn (N) 0.65, Thr (T) 0.66, Cys (C) 0.68, Asp (D) 0.69, and Gly (G) 1. For the terminal residues 12, 13, 23, and 24 of the helix, the propensity scale is not applicable. We only consider mutations with $\Delta G_N > 0.7$ kcal/mol.

here on the well-characterized α-helices of two-state proteins for which at least 10 Φ-values are available: the helices 2 and 3 from the protein A, and the helix of CI2. Protein A is an α-helical protein with three helices (see fig. 2.8), whereas CI2 is an α/β-protein with a single α-helix packed against a β-sheet (see fig. 1.1).

Our analysis of experimental Φ-values requires an estimate of the mutation-induced changes ΔG_α of the intrinsic helix stability. In the case of the CI2 helix, we estimate ΔG_α both with the program AGADIR [107–109] and from a helix-propensity scale [111], see table 2.4. The change in intrinsic helix stability ΔG_α can be estimated from the helical content predicted by AGADIR via $\Delta G_\alpha = RT \ln(P_\alpha^{\text{wt}}/P_\alpha^{\text{mut}})$. Here, P_α^{wt} is the helical content of the wildtype helix, and P_α^{mut} the helical content of the mutant. The program AGADIR is based on helix/coil transition

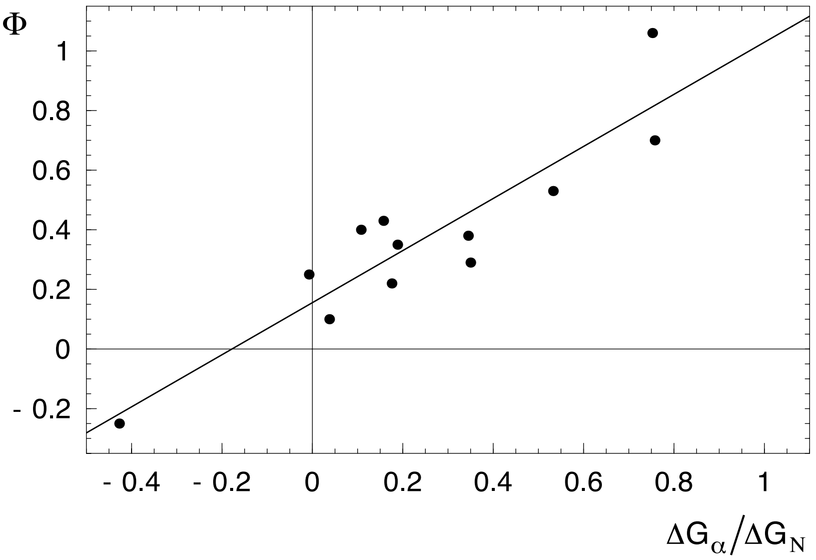

Figure 2.9: Analysis of Φ-values for mutations in the helix of the protein CI2. The change in intrinsic helix stability ΔG_α for the 12 mutations has been calculated with AGADIR (see table 2.4). We only consider mutations with experimentally measured stability changes $\Delta G_N > 0.7$ kcal/mol. The Pearson correlation coefficient of the 12 data points is 0.91. From the regression line $\Phi = 0.16 + 0.87 \Delta G_\alpha / \Delta G_N$, we obtain the structural parameters $\chi_\alpha = 1.03 \pm 0.05$ and $\chi_t = 0.16 \pm 0.05$. The structural parameter χ_α close to 1 indicates that the helix is fully formed in the transition state. The parameter χ_t indicates that tertiary interactions are on average present in the transition state to a degree around 16 %.

theory, with parameters fitted to data from Circular Dichroism (CD) spectroscopy. In table 2.4, the values for ΔG_α obtained from AGADIR are compared to values from a helix-propensity scale [111]. Helix propensities of the amino acids are typically given as free-energy differences with respect to Alanine. These free-energy differences represent averages of experimentally measured changes in helix stability induced by mutations. We use the propensity scale of Pace and Scholtz [111], which has been obtained from experimental data on 11 different helical systems. For example, the value $\Delta G_\alpha = 0.29$ kcal/mol for the mutant E15D in the CI2 helix is simply the difference between the helix propensity 0.69 kcal/mol for the amino acid D (Aspartic acid) and the propensity 0.40 kcal/mol for amino acid E (Glutamic acid). The helix-propensity scale can be applied for residues at 'inner' positions' of a helix, not for residues at the termini or 'caps' of the helix. The N-terminal residues of the CI2 helix are the

Table 2.5: Mutational data for helix 2 of protein A

mutation	Φ	ΔG_N	ΔG_α	tertiary contacts
A27G	1.0	1.0	1.0	–
A28G	0.6	2.2	1.0	helix 1
A29G	1.1	1.0	1.0	–
F31A	0.3	3.9	-0.54	helices 1 and 3
F31G	0.5	4.7	0.46	helices 1 and 3
I32V	0.6	1.2	0.2	helix 1
I32A	0.5	1.9	-0.41	helix 1
I32G	0.6	3.4	0.59	helix 1
A33G	1.1	0.9	1.0	–
A34G	0.7	1.2	1.0	–
L35A	0.4	2.4	-0.21	helices 1 and 3
L35G	0.5	4.1	0.79	helices 1 and 3

Experimental Φ-values and stability changes ΔG_N are from Sato et al. [99]. The change in intrinsic helix stability ΔG_α is calculated from the helix-propensity scale of Pace and Scholtz [111]. The information whether tertiary contacts with helix 1 and 3 are affected by the mutations is taken from the contact matrix of protein A shown in fig. 2.8. We only consider Φ-values for single-residue mutations with the wildtype sequence as reference state at those sites where multiple mutations have been performed. For example, we consider the Φ-values for the mutations I32V, I32A, and I32G in helix 2 of protein A, but not the Φ-values for V32A and A32G also given by Sato et al. [99]. However, we include the Φ-values for the Ala-Gly scanning mutants at the residue positions 27, 28, 29, 33, and 34 given in table 1 of Sato et al. [99].

residues 12 and 13, the C-terminal residues are the residues 23 and 24. For the 8 mutations at 'inner positions' of the CI2 helix, the values for ΔG_α from AGADIR and from the helix-propensity scale correlate with a Pearson correlation coefficient of 0.77. For the protein A helices, the helicities predicted by AGADIR are significantly smaller than the helicities around 5 % for the CI2 helix. Estimates for ΔG_α based on AGADIR therefore are not reliable for these helices. The values of ΔG_α shown in the tables 2.5 and 2.6 are calculated from helix propensities.

Three helices in protein A constitute its three structural elements. Based on the contact map of protein A shown in fig. 2.8, the mutations in helix 2 of protein A can be divided into three groups: 'purely secondary' mutations that don't affect tertiary contacts; mutations that

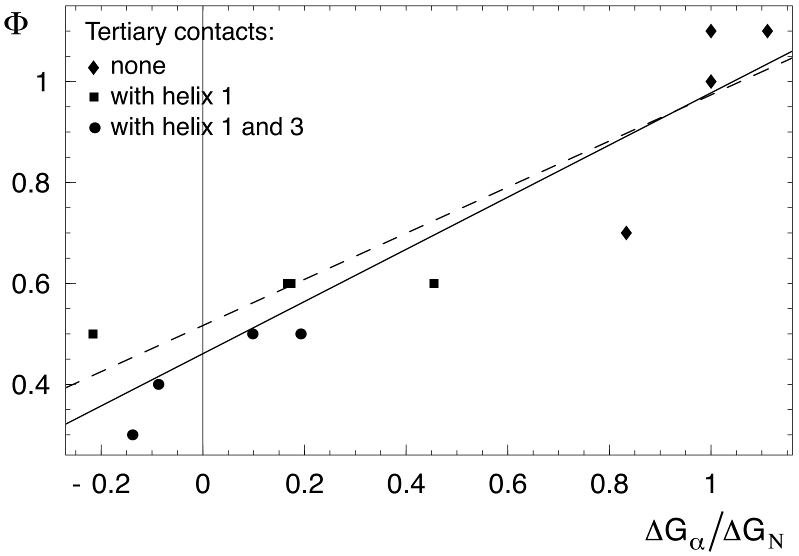

Figure 2.10: Analysis of Φ-values for helix 2 of protein A. The solid line represents the regression line $\Phi = 0.46 + 0.52\,\Delta G_\alpha/\Delta G_N$ for all points. The correlation coefficient of the data points is 0.93. The dashed line is the regression line $\Phi = 0.52 + 0.46\,\Delta G_\alpha/\Delta G_N$ of the 8 data points for mutations of residues that have either no tertiary interactions or tertiary interactions with helix 1 (see also table 2.5). The correlation coefficient of these data points is 0.90. From the regression lines and eq. (2.18), we obtain the structural parameters χ_α and χ_t shown in table 2.7. The values of χ_α close to 1 indicate that the helix is fully formed in the transition state, and the values of χ_t close to 0.5 indicate that tertiary interactions are present to a degree of about 50 %.

affect only tertiary contacts with helix 1; and mutations that affect tertiary contacts both with helix 1 and 3. If only the first two groups of mutations are considered in our analysis, χ_t represents the average degree of structure formation with helix 1. If all groups and, thus, all mutations are considered, χ_t is the average degree of structure formation with the helices 1 and 3. In the case of helix 3, we distinguish between mutations that affect either tertiary contacts with helix 1 or helix 2, or none of the tertiary interactions, see table 2.6. In the case of CI2, we do not distinguish between different tertiary contacts. One reason is that there are at least three other structural elements to consider, the three strand pairings $\beta_2\beta_3$, $\beta_3\beta_4$, and $\beta_1\beta_4$ of the four-stranded β-sheet that is packed against the CI2 helix [110]. Another reason is that the degree χ_t of tertiary structure formation in the transition state is small for this helix.

Table 2.6: Mutational data for helix 3 of protein A

mutation	Φ	ΔG_N	ΔG_α	tertiary contacts
A44G	-0.1	1.3	1.0	–
L45A	0.6	1.5	-0.21	helix 2
L45G	0.3	4.4	0.79	helix 2
L46A	0.2	1.9	-0.21	helix 1
L46G	0.3	4.0	0.79	helix 1
A47G	0.2	1.5	1.0	–
A48G	0.0	1.8	1.0	helix 2
A49G	0.2	3.6	1.0	helix 2
A51G	0.1	1.2	1.0	–
L52A	0.3	1.3	-0.21	helix 2
L52G	0.1	3.8	0.79	helix 2
A54G	0.0	1.4	1.0	–

Experimental Φ-values and stability changes ΔG_N are from Sato et al. [99]. The change in intrinsic helix stability ΔG_α is calculated from helix propensities [111]. The information on tertiary contacts is taken from fig. 2.8.

The structural parameters χ_α and χ_t obtained from our analyses shown in the figs. 2.9 to 2.11 are summarized in table 2.7. We estimate the overall errors of χ_α and χ_t, which result from experimental errors in Φ and ΔG_N and from modeling errors, as ± 0.05 for the CI2 helix and helix 2 of protein A, and as ± 0.1 for helix 3 of protein A. The χ_α values for the CI2 helix and the helix 2 of protein A are close to 1. This indicates that the helices are fully formed in the transition-state ensemble. In contrast, χ_α for helix 3 of protein A is close to 0, indicating that the helix is not formed in the transition state. Our χ_t values indicate that the degree of tertiary structure formation in the transition state is around 16 % for the CI2 helix, around 50 % for helix 2 of protein A, and around 30 % for helix 3 of protein A.

To assess the quality of our modeling, we consider two quantities: the correlation coefficient r, and the estimated standard deviation SD of the data points from the regression line. High correlation coefficients up to 0.9 and larger indicate a high quality of modeling. However, it's important to note that the correlation coefficient can only be used to assess the modeling quality in the cases where the structural parameters χ_α and χ_t are sufficiently different from each other. The case $\chi_\alpha = \chi_t$ corresponds to a regression line with slope 0, and hence a correlation

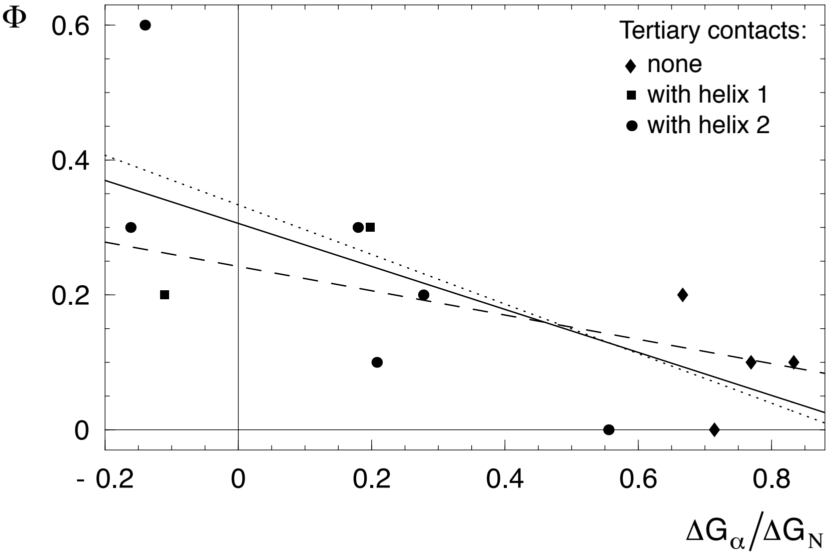

Figure 2.11: Analysis of Φ-values for mutations in helix 3 of protein A. The solid line represents the regression line $\Phi = 0.31 - 0.38\Delta G_\alpha/\Delta G_N$ of all data points; the dashed line is the regression line $\Phi = 0.24 - 0.25\Delta G_\alpha/\Delta G_N$ of the data points for mutations that affect the tertiary interactions with helix 1 (or no tertiary interactions); and the dotted line is the regression line $\Phi = 0.34 - 0.43\Delta G_\alpha/\Delta G_N$ of data points for mutations that affect tertiary interactions interactions with helix 2 (or no tertiary interactions). The absolute values of the correlation coefficient for these three data sets are $|r| = 0.75$, 0.65, and 0.79, respectively (see table 2.7).

coefficient of 0, irrespective of how well the data are represented by this line.

We only consider here mutations with stability changes $\Delta G_N > 0.7$ kcal/mol. Because of experimental errors, Φ-values for mutations with smaller stability changes are generally considered as unreliable [40,45,93]. In Ref. [110], we considered all the published mutations for the CI2 helix, including those for which ΔG_N is significantly smaller than 0.7 kcal/mol. The correlation coefficient 0.91 obtained here for the subset of mutations with $\Delta G_N > 0.7$ kcal/mol is larger than the correlation coefficient 0.85 for all mutations.

In our model, nonclassical Φ-values < 0 or > 1 can arise if $\Delta G_\alpha/\Delta G_N$ is < 0 or > 1. Since $\Delta G_N = \Delta G_\alpha + \Delta G_t$, this implies that ΔG_α and ΔG_t have opposite signs. Our model reproduces the clearly negative Φ-value for the mutation D23A in the CI2 helix. The mutation stabilizes the helix (i.e. $\Delta G_\alpha < 0$), but destabilizes tertiary interactions ($\Delta G_t > 0$).

Table 2.7: Structural parameters, standard deviations, and correlation coefficients

helix	tertiary contacts	χ_α	χ_t	SD	$\lvert r\rvert$
CI2 helix	all	1.03	0.16	0.14	0.91
helix 2 of protein A	all	0.98	0.46	0.10	0.93
	with helix 1	0.98	0.52	0.12	0.90
helix 3 of protein A	all	-0.07	0.31	0.13	0.75
	with helix 1	-0.01	0.24	0.13	0.65
	with helix 2	-0.09	0.34	0.13	0.79

The structural parameters χ_α and χ_t, estimated standard deviations SD of the data points from the regression lines, and absolute values of the correlation coefficient r obtained in our model. The second column of the table indicates whether we consider all mutations for a helix, or only mutations affecting tertiary interactions with one structural element. The structural parameter χ_t then either indicates the overall degree of tertiary structure formation in the transition state, or the degree of tertiary structure formation with the given structural element. In both cases, we have included the 'purely secondary' mutations that do not affect tertiary interactions. The structural elements of protein A are defined in fig. 2.8. The standard deviation SD is estimated as $\mathrm{SD} = \sqrt{\left(\sum_{i=1}^{M} d_i^2\right)/(M-2)}$ where d_i is the vertical deviation of data point i from the regression line, and M is the number of data points. We estimate the errors in the structural parameters χ_α and χ_t, which result from experimental and modeling errors, as ± 0.05 for the CI2 helix and helix 2 of protein A, and as ± 0.1 for helix 3 of protein A.

2.5 Summary

In this chapter, we have presented simple statistical-mechanical models for the interpretation of mutational data on the folding kinetics of proteins. The models have two main features. First, the models are based on the assumption that substructural elements such as β-hairpins and α-helices form cooperatively. These substructural elements are either fully formed are not formed in partially folded states of the models. Second, mutation-induced free-energy changes are split into different components. Free-energy changes for mutations in protein helices, for example, are split into secondary and tertiary components, see section 2.3.

In section 2.1, we have considered a statistical-mechanical model for the folding kinetics of WW domains and other three-stranded β-sheet proteins. The characteristic substructural elements of these proteins are two β-hairpins. The assumption of cooperative hairpin formation implies that the model has two transition-state conformations in which either of the hairpins is formed (see fig. 2.2). In agreement with this model, a transition state that is "characterized by the presence of one of the two native hairpins formed while the rest of the peptide is mainly unstructured" [112] has also been obtained from Molecular Dynamics simulations of the folding and unfolding of a three-stranded mini-protein.

In our model, Φ-values for mutations of WW domains have the general form (see eq. 2.11)

$$\Phi = \frac{\Delta G_T}{\Delta G_N} = \frac{\chi_1 \Delta G_1 + \chi_2 \Delta G_2}{\Delta G_N}$$

Here, χ_1 is the probability, or fraction, of the transition-state conformation in which hairpin 1 is formed, and $\chi_2 = 1 - \chi_1$ is the probability of the transition-state conformation with hairpin 2 formed. The mutation-induced changes of the free-energy difference between the two transition-state conformations and the denatured state are denoted by ΔG_1 and ΔG_2.

To test eq. (2.11) in section 2.2, we have first considered those mutations of the FBP and PIN WW domains that affect only one of the hairpins, which are identified from the contact maps shown in fig. 2.3. Our model predicts that all mutations that affect only hairpin 1 have the same Φ-value χ_1, since we have $\Delta G_1 = \Delta G_N$ and $\Delta G_2 = 0$ for these mutations. Correspondly, mutations that affect only hairpin 2 have the same Φ-value χ_2. We find that the experimental Φ-values for mutations that affect only hairpin 1 and hairpin 2 follow this rule, within reasonable errors (see table 2.1, table 2.3, and fig. 2.4). A second test of our model is that the values for χ_1 and χ_2 determined from these mutations add up to 1, which is the case, within reasonable errors (see section 2.2).

Third, to apply eq. (2.11) to mutations that affect both hairpins, or one of the hairpins and the small hydrophobic core of the FBP WW domain, we have estimated ΔG_1 and ΔG_2 using the molecular modeling programs WHAT IF [92] and FOLD-X [90, 91]. We obtain good agreement with the experimental data by fitting a single parameter (see fig. 2.5). In particular, the model reproduces the negative Φ-value for the mutation L36A of the FBP WW domain. According to the model, this mutation stabilizes hairpin 2 ($\Delta G_2 < 0$), but destabilizes the hydrophobic core to a larger extent ($\Delta G_N > 0$), which leads to a negative Φ-value according to eq. (2.11) since ΔG_1 equals 0 for this mutation (see table 2.2).

In section 2.3, we have presented a simple model for the formation of helices during protein folding. In this model, the helix formation in the transition-state ensemble T is described by two structural parameters: χ_α, the degree of secondary structure formation of the helix in T, and χ_t, the degree of tertiary structure formation. The mutation-induced free-energy changes are split into two components. The overall stability change ΔG_N is split into the change in intrinsic helix stability ΔG_α, and the change in tertiary free energy ΔG_t caused by the mutation. Similarly, ΔG_T, the change of the free-energy difference between the transition state and the denatured state, is split into a change $\chi_\alpha \Delta G_\alpha$ in secondary free energy, and a change $\chi_t \Delta G_t$ in tertiary free energy. The Φ-values for the mutations in the helix then have the general form (see eq. 2.18)

$$\Phi = \frac{\chi_\alpha \Delta G_\alpha + \chi_t \Delta G_t}{\Delta G_N} = \chi_t + (\chi_\alpha - \chi_t) \frac{\Delta G_\alpha}{\Delta G_N}$$

The second expression results from replacing ΔG_t by $\Delta G_N - \Delta G_\alpha$.

In section 2.4, eq. (2.18) has been applied to interpret mutational data for the CI2 helix and the helices 2 and 3 of protein A, which requires estimates of ΔG_α with standard helix-propensity scales. The structural parameters χ_α and χ_t for each of the helices then result from fitting eq. (2.18) to the experimentally measured Φ-values and stability changes ΔG_N. Different Φ-values for different mutations in a helix result from characteristic free-energetic 'signatures' ΔG_α and ΔG_N of the mutations. In particular, eq. (2.18) captures that different mutations of the same residue can lead to different Φ-values, and that Φ-values can be 'nonclassical', i.e. < 0 or > 1. Nonclassical Φ-values can arise if the changes ΔG_α and ΔG_t in secondary and tertiary free energy caused by the mutation have opposite signs. The model reproduces the negative Φ-value for the mutation D23A in the CI2 helix, which stabilizes the helix, but destabilizes tertiary interactions (see table 2.4).

3 Loop-closure principles

3.1 Topology and loop closure

The topic of this chapter is the relation between the folding kinetics of proteins and their three-dimensional, native structures. Since their discovery in 1991 [14], two-state proteins have been in the focus of experimental studies [15–17,113]. These proteins fold from the denatured state to the native state without experimentally detectable intermediate states. The size of most two-state proteins is rather similar, roughly between 60 and 120 residues, with a few smaller or larger exceptions [15,17,113,114]. Nonetheless, their folding rates range over six orders of magnitude: the fastest proteins fold on a microsecond [115, 116] and, if designed for speed, sub-microsecond time scale [117,118], whereas slow two-state proteins fold on a time scale of seconds [26]. In 1998, Plaxco, Simons, and Baker [73] discovered that these folding rates correlate with a simple measure of the structural 'topology', the relative contact order (CO). The relative CO is the average 'localness' or sequence separation $|i - j|$ of all contacts between amino acids i and j in the native structure, divided by the chain length. Proteins with many local contacts and, hence, small relative CO, tend to fold faster than proteins with many nonlocal contacts and large relative CO. The discovery of Plaxco et al. pointed towards a 'surprising simplicity' [119] in protein folding kinetics. The folding kinetics problem, i.e. the problem of predicting folding rates and routes from native structures, appeared to be considerably simpler than the structure problem, the prediction of native structures from sequences, which requires detailed atomistic models [120].

The physical principle that underlies the correlation between folding rates and relative CO seems to be loop closure [121]. Contacts with small CO can be formed by closing a small loop, which is fast and requires a small amount of loop-closure entropy, compared to closing a large loop [122,123]. It seems plausible that protein structures with many local contacts form faster than proteins with more complex structures involving many nonlocal contacts, provided that the loop-closure entropies, or chain entropies, dominate over sequence-dependent interaction energies in the folding process. The strength of the correlation between folding rates and relative CO or related structural measures discussed in section

3.3 indicates such a dominance of topological or loop-closure aspects, at least for a majority of proteins. Depending on the considered set of two-state proteins, the absolute values $|r|$ of the Pearson correlation coefficients between folding rates and relative CO of two-state proteins vary between 0.75 and 0.9 (see section 3.3). The squares of these correlation coefficients range roughly from 0.6 to 0.8, which indicates that between 60 % to 80 % of the observed variations in the folding rates can be traced back to simple aspects of the overall structure or topology, rather than sequence-specific energetic aspects.

Several experimental observations support the importance of protein topology and loop closure. First, insertion of small loops into turns of the protein structure slows down folding [122, 124–126]. Second, inserting covalent crosslinks into the protein chain speeds up the folding process [126–130]. The crosslinks interconnect the chain and increase the localness of some of the contacts in the protein structure. Third, single-residue mutations that locally perturb energetic interactions typically have a ‘less than tenfold effect’ [119] on the folding rate, which appears small compared to the variations in folding rate observed for two-state proteins. For few single-residue mutants, larger changes in the folding rate have been observed [115, 131]. Also, homologous proteins of the same size, which have the same structure but can differ considerably in sequence, have folding rates that differ typically by less than one or two orders of magnitude [15, 132, 133], which appears, again, small compared to the six orders of magnitude observed for two-state proteins.

Can we also predict routes from loop-closure principles? The CO or sequence separation of a contact is the length of the loop that has to be closed to form the contact, provided that no other contacts have been formed prior that ‘short-circuit’ the chain. In other words, the CO measures loop lengths for the fully unfolded state of the protein chain. But during folding, other contacts may have been formed prior to a specific contact between residues i and j. The actual length of the loop that has be closed to form this contact in the partially folded state of the protein chain can be estimated via the graph-theoretical concept of effective contact order (ECO) [134, 135]. The ECO is the length of the shortest path between the two residues i and j that are brought in contact, see fig. 3.1. The steps on this path either are bonds between neighboring residues in the chain, or contacts between residues that have been formed prior, such as contact C_1 between the resides k and l in fig. 3.1. In contrast to COs, the ECOs thus are route-dependent: they depend on the sequence in which contacts are formed. On the minimum-ECO routes discussed in section 3.4, proteins fold, or ‘zip up’, in sequences of events that involve only closures of small loops, which minimizes the entropic loop-closure

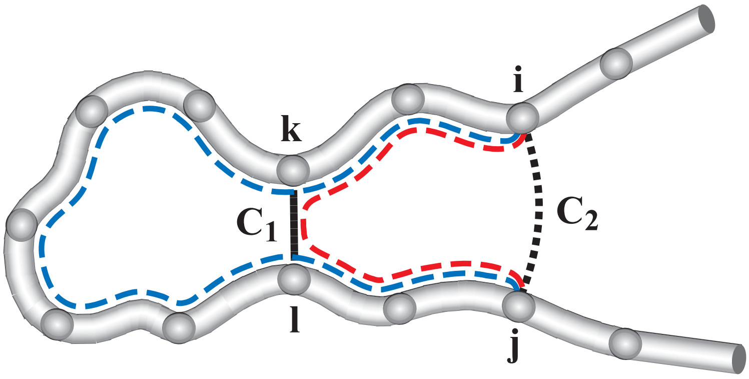

Figure 3.1: Loop lengths in partially folded conformations of a protein chain can be estimated via the graph-theoretical concept of effective contact order (ECO) [134, 135]. The ECO of the contact C_2 is the length of the shortest path between the two residues i and j forming the contact. The 'steps' along this shortest path either are covalent bonds between adjacent residues, or noncovalent contacts formed previously in the folding process such as the contact C_1. In this example, the ECO for the contact C_2 is 5, since the shortest path (shown in red) involves two steps from i to k, one step for the contact C_1 between k and l, and two steps from l to j. The contact order (CO), in contrast, is the sequence separation $|i - j|$ between the two residues, the number of residues along the blue path between i and j. In this example, the CO of the contact C_2 is 10.

barriers during folding. The minimum-ECO routes help to understand the shape of Φ-value distributions from mutational analyses of the folding kinetics, see section 3.5.

3.2 Contact maps, contact clusters, and topology

To capture the concept of native-state topology more precisely, it is helpful to consider native contact maps. Contact maps are two-dimensional representations of three-dimensional protein structures. The native contact map of a protein is a matrix in which element (i, j) indicates whether the residues i and j are in contact in the native structure. To some extent, the native contact map depends on the contact definition. In the map of fig. 3.2(a), two residues are defined to be in contact if the distance between their backbone C_α atoms is smaller than 7 Å, and in fig. 3.2(b),

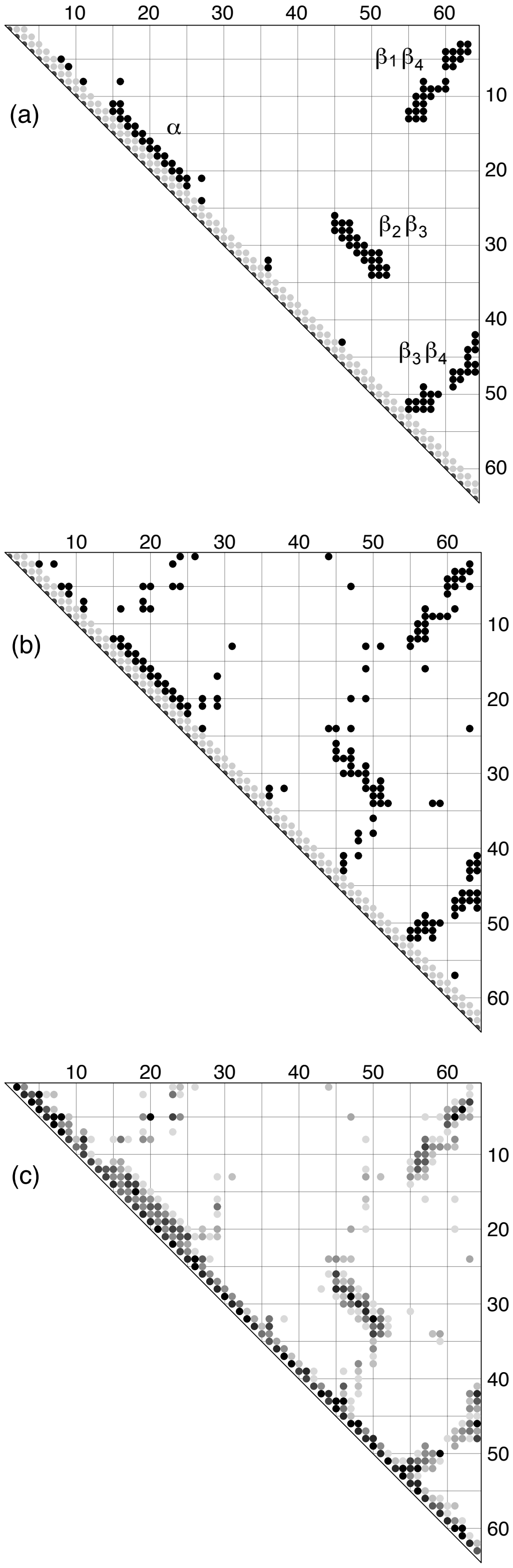

Figure 3.2: Native contact maps of the protein CI2 shown in fig. 1.1, for different contact definitions: (a) A black dot at position (i, j) indicates that the C_α atoms of the residues i and j are within the cutoff distance 7 Å. The four large clusters of contacts represent the structural elements of CI2, i.e. the α-helix and the three β-strand pairings $\beta_2\beta_3$, $\beta_3\beta_4$, and $\beta_1\beta_4$. – (b) Black dots (i, j) indicate that at least two non-hydrogen atoms of the residues i and j are within cutoff distance 4.5 Å. As above, contacts of neighboring or next-nearest neighboring residues in the chain (with $|i - j| \leq 2$) are not taken into account. – (c) The gray scale of the dots indicates the numbers of non-hydrogen-atom pairs of two residues i and j that are within the cutoff distance 6 Å. Black dots represent residues for which more than 40 different non-hydrogen-atom pairs are in contact, lighter gray colors represent residues with fewer non-hydrogen-atom contacts.

if the distance between any of their non-hydrogen atoms is smaller than 4.5 Å. In the C_α contact map of fig. 3.2(a), the contacts are arranged in clusters that correspond to the characteristic structural elements of CI2. These clusters are also present in the non-hydrogen-atom contact map of fig. 3.2(b). In addition, the non-hydrogen-atom contact map contains more ‘isolated’ contacts that mostly correspond to interactions of large sidechains, which are not represented in the backbone-centric C_α contact map. A third type of contact map is shown in fig. 3.2(c). The different gray tones in this map indicate the numbers of contacting non-hydrogen atom pairs of two residues. This contact map is the basis for the calculation of the relative CO.

The contact maps of fig. 3.2 indicate the chain positions i and j of contacting amino acids, but not which of the twenty different types amino acids are located at these positions. In other words, the contact maps do not contain sequence information, they just contain information on the structure. This structural information is rather detailed. Single-residue mutations can lead to deletion or addition of contacts, and homologous proteins of the same size can differ in many native contacts. Nonetheless, single-residue mutants and homologous proteins have the same overall structure. To capture the overall structure or ‘structural topology’ of a protein, it is helpful to take a more coarse-grained view of contact maps and to focus on contact clusters, e.g. on the clusters in the C_α contact map of fig. 3.2(a). The size of contact clusters may vary between wildtype and mutants of a protein, or between homologous proteins of similar size. But the overall location of these clusters in the contact map in general stays the same. The contact clusters thus capture the overall structural topology of a protein.

3.3 Folding rates and topological measures

Simple measures of native-state topology are characteristic, average properties of contact maps. The relative CO defined by Plaxco et al. [73] is the average CO of all contacts between non-hydrogen atoms of the contact map shown in fig. 3.2(c), divided by the chain length N. The CO of the contacting atoms simply is the sequence separation $|i - j|$ of the two residues $i \neq j$ in which the atoms are located. Depending on the data set, the obtained correlations between relative CO and the folding rates of two-state proteins vary between 0.75 and 0.92, see table 3.1 [121]. Proteins with many local contacts between residues that are close in the chain sequence have a small relative CO and fold faster than proteins with many nonlocal contacts and large relative CO. The typically fast-folding

α-helical proteins have small relative COs since their contact maps contain many local, intra-helical contacts between residues i and $i+3$ or $i+4$. Proteins with β-sheets, in contrast, have larger relative COs and, on average, fold slower. But also within the classes of α-helical and β-sheet containing proteins, significant correlations between folding rates and relative CO can be observed [136].

Related topological measures that correlate with the folding rates of two-state proteins are the 'long-range order' [137], the 'total contact distance' [138], and the number Q_D of nonlocal contacts with CO $>$ 12 [139, 140] (see table 3.1). The long-range order is the number of contacts with CO > 12, divided by the chain length, and the total contact distance is the sum over the COs of all contacts, divided by the chain length squared. The topomer-search model of Makarov et al. [139, 140] predicts that the number Q_D of nonlocal contacts is proportional to $\log k_f/Q_D$ where k_f is the folding rate [139, 140]. The diffusive search for a topomer [139–142], i.e. for the "set of unfolded conformations that share a common, global topology with the native state" [140], has been suggested as a physical principle that underlies the correlation between relative CO and folding rates [139, 140, 143]. Recent extensive simulations with an off-lattice model indicate, however, that an unbiased diffusive search process for a native topomer "would take an impossibly long average time to complete" [144].

Can topological measures capture the increase in folding rate that is caused by the insertion of covalent chain crosslinks [126–129]? Inserting crosslinks such as disulfide bonds into the protein chain decreases the localness of some of the native contacts, since the crosslinks 'short-circuit' the chain. The natural extension of the CO or localness of a contact in a crosslinked chain is the *minimum* number of covalently connected residues between the two residues in contact. This minimum number is the ECO of the contact in the crosslinked but otherwise unfolded chain, and the relative ECO is the natural extension of the relative CO for a crosslinked chain. The relative ECO of a protein structure is defined as [148]

$$\text{rel. ECO} = \frac{1}{MN}\sum_{i=1}^{M}\text{ECO}(i) \tag{3.1}$$

The sum is taken over all contacts i between non-hydrogen atoms of different residues, with total number M, and N is the chain length, the total number of residues. The ECO of contact i here is the minimum number of covalently connected residues between the residues in contact. As Plaxco et al. [73, 145], we define two non-hydrogen atoms to be in contact if their distance is less than 6 Å, see also fig. 3.2(c).

Table 3.1: Correlation coefficients $|r|$ between folding rates of two-state proteins and simple topological measures

Authors	Ref.	size of protein set	rel. CO	abs. CO	LRO	TCD	Q_D	rel. logCO
Plaxco et al.	[145], [140]	24	0.92				0.88	
Gromiha & Selvaraj	[137]	23	0.79(a)		0.78			
Zhou & Zhou	[138]	28	(0.74)(b)		0.81	0.88		
Micheletti	[146]	29	0.75	0.70				
Ivankov et al.	[147]	30	0.75	0.51				
Kamagata et al.	[114]	18	0.84	0.78			0.88	
Dixit & Weikl	[148]	26	0.92	0.69	0.84(c)	0.90(c)	0.82(c)	0.90

Absolute values $|r|$ of the Pearson coefficient for the correlations between folding rates of several sets of two-state proteins and relative contact order (rel. CO) [73], absolute contact order (abs. CO) [17], long-range order (LRO) [137], total contact distance (TCR) [138], and relative logCO [148]. In case of the number of nonlocal contacts Q_D [139], the given coefficients report the correlations between Q_D and $\log k_f/Q_D$ where k_f are the folding rates.

(a) calculated from table 1 of Ref. [137]

(b) the value is given in brackets since a slightly different definition for the relative CO has been used

(c) the values have been calculated for the protein structures given in Ref. [148]. The protein set is the set of Grantcharova et al. [17], which extends the set of Plaxco et al. [145] by two proteins. The folding rates, relative COs, and relative logCOs for all proteins of this set are shown in table 3.2.

Table 3.2: Two-state proteins without crosslinks

protein	PDB file	$\log(k_f)$[a]	rel. CO (%)	rel. logCO (%)	length
Cyt b_{562}	256B	5.30	7.5	24.7	106
myoglobin	1BZP	4.83	8.0	25.1	153
λ-repressor	1LMB3	4.78	9.4	26.0	80
PSBD	2PDD[b]	4.20	11.0	24.9	41
Cyt c	1HRC[c]	3.80	11.2	29.6	104
Im9	1IMQ	3.16	12.1	29.7	86
ACBP	2ABD	2.85	14.3	32.0	86
Villin 14T	2VIK	3.25	12.3	33.5	126
N-term L9	1DIV[d]	2.87	12.7	29.6	56
Ubiquitin	1UBQ	3.19	15.1	33.2	76
CI2	2CI2[e]	1.75	15.7	32.2	64
U1A	1URNA	2.53	16.9	34.7	96
Ada2h	1AYE[f]	2.88	16.7	33.0	80
Protein G	1PGB	2.46	17.3	34.4	56
Protein L	1HZ6A[g]	1.78	16.1	33.8	62
FKBP	1FKB	0.60	17.7	37.3	107
HPr	1POH	1.17	17.6	34.6	85
MerP	1AFI	0.26	18.9	36.7	72
mAcP	1APS	−0.64	21.7	40.0	98
CspB	1CSP	2.84	16.4	35.7	67
TNfn3	1TEN[h]	0.46	17.4	37.6	89
TI I27	1TIT	1.51	17.8	36.4	89
Fyn SH3	1SHF	1.97	18.3	36.7	59
Twitchin	1WIT	0.18	20.3	40.9	93
PsaE	1PSF	0.51	17.0	34.5	69
Sso7d	1BNZA	3.02	12.2	30.8	64

[a] Experimental values for the folding rates k_f are from table 1 of Grantcharova et al. [17].

[b] Residues 3 to 43.

[c] Residues 1 to 104.

[d] Residues 1 to 56.

[e] Residues 20 to 83.

[f] Residues 4A to 85A.

[g] Residues 1 to 62.

[h] Residues 803 to 891.

For NMR structures with multiple models, the values for the rel. CO and rel. logCO are averages over all models. Alternate locations for atoms in PDB files have been discarded to avoid double or triple counting of corresponding contacts.

Table 3.3: Two-state proteins with crosslinks

protein	Ref.	PDB file	length	crosslink(s)	log(k_f)[a]	rel. ECO	rel. CO	rel. logECO	rel. logCO
circularized CI2	[127]	2CI2[b]	64	3/63	2.64	9.4	15.7	27.8	32.2
src SH3 with SS 35/50	[126]	1SRL	56	35/50	3.15	14.8	19.6	32.4	36.6
src SH3 with SS 1/25	[126]	1SRL	56	1/25	1.99	15.7	19.6	34.3	36.6
circularized src SH3	[126]	1SRL	56	1/56	2.39	11.8	19.6	31.4	36.6
wild-type tendamistat	[128]	2AIT	74	11/27, 45/73	1.82	14.3	21.1	36.5	41.4
tendamistat w/o SS 11/27	[128]	2AIT	74	45/73	0.88	16.3	21.1	38.4	41.4
tendamistat w/o SS 45/73	[128]	2AIT	74	11/27	0.32	19.2	21.1	39.8	41.4
circularized c-Crk SH3	[129]	1CKB[c]	57[d]	1/56[d]	0.86	13.5	20.3	34.5	38.7

[a] Folding rates k_f in pure water and units of $1/s$. For circularized CI2 and the crosslinked mutants of the the src SH3 domain, folding rates in water have been extrapolated using the m_f-values given in the references.

[b] Residues 20 to 83.

[c] Residues 135 to 190.

[d] The N- and C-termini are crosslinked via an inserted glycine residue. To calculate the rel. ECO and rel. logECO for the circularized chain, we simply assume that this glycine residue makes 25 non-hydrogen atom contacts with the nearest neighbor residues 1 and 56, and 10 contacts with the next-nearest neighbor residues 2 and 55 (these are typical numbers for glycine residues), but makes no contacts with other residues. We also assume that the residues 1 and 56 have 10 contacts in the circularized chain.

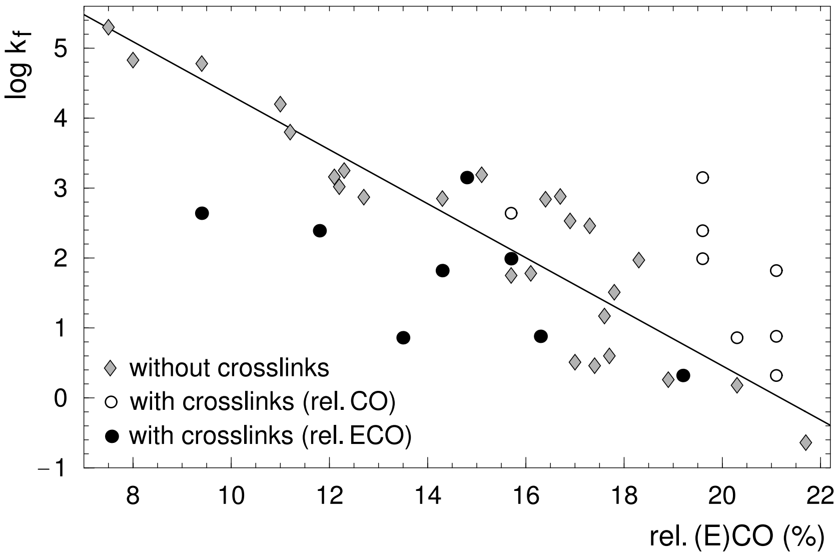

Figure 3.3: Relative CO of 26 two-state proteins without crosslinks (gray diamonds), relative CO of 8 two-state proteins with crosslinks (open circles), and relative ECO of these 8 proteins (filled circles) plotted against the decadic logarithm of their folding rates k_f. The regression line for the 26 proteins without crosslinks is given by $\log k_f = 8.18 - 0.386 \times (\text{rel. CO})$ and provides a topology-based estimator for the folding rates of such proteins. The location of the majority of filled circles clearly below the regression line indicates that the relative ECO, the natural extension of relative CO to proteins with crosslinks, tends to overestimate the folding rates of these proteins. The proteins are listed in the tables 3.2 and 3.3.

For proteins without crosslinks, the relative ECO of the protein structure is identical with the relative CO. Grantcharova et al. [17] have considered a set of 26 proteins without crosslinks, extending a previous set of Plaxco et al. [145] by two proteins. In fig. 3.3, the relative CO of these 26 proteins is plotted against the decadic logarithm of their folding rates (gray diamonds), together with the relative CO (open circles) and the relative ECO (filled circles) of 8 two-state proteins with crosslinks. For the 26 proteins without crosslinks, the Pearson correlation coefficient between folding rate and relative CO is 0.92. The line in fig. 3.3 represents the regression line for this proteins. The position of the open circles above this regression line indicates that the relative CO of the 8 proteins with crosslinks underestimates the folding rates of these proteins. This is not unexpected, since the relative CO does not capture crosslinks, which speed up the folding process. The standard deviation of the open circles in vertical direction from the regression line is 1.42, which is significantly larger than the standard deviation of 0.61 for the 26 proteins without crosslinks. On the other hand, the relative ECO overestimates the folding rate of the proteins with crosslinks. The majority of the filled circles is located clearly below the regression line for the proteins without crosslinks, and the standard deviation of the 8 points from the regression line is 1.23. Despite the small number of data points, this deviation for the relative ECO provides a relatively clear, negative answer, since it could only be 'compensated' in a much larger data set. For example, suppose we hypothetically add 8 'good' data points with the same standard deviation 0.61 as the 26 proteins without crosslinks to the 8 'poor' data points for the crosslinked proteins with standard deviation 1.23. The resulting set of 16 data points still has a standard deviation of $\sqrt{(1.23^2 + 0.61^2)/2} = 0.97$, which is significantly larger than the deviation 0.61 for the proteins without crosslinks.

In fig. 3.4, we consider the relative logECO, a novel measure of native-state topology and chain connectivity, defined as [148]

$$\text{rel. logECO} = \frac{1}{M \log N} \sum_{i=1}^{M} \log\left[\text{ECO}(i)\right] \tag{3.2}$$

Here, N is again the chain length, and M the number of contacts. For the 26 proteins without crosslinks, the relative logECO is identical with the relative logCO $= \sum_{i=1}^{M} \log\left[\text{CO}(i)\right]/(M \log N)$. The relative logCO correlates with the foldings rates of these 26 proteins with a Pearson coefficient of 0.90, which is only slightly smaller than the correlation coefficient 0.92 for the relative CO. In addition, the relative logECO captures the folding rates of the 8 proteins with crosslinks. The standard deviation of the filled circles from the regression line of the 26 proteins without crosslinks is 0.70 and, thus, comparable to the standard deviation 0.67 for these 26 proteins. The relative logECO therefore provides a simple estimator for the folding rates of two-state proteins both with and without crosslinks.

The CO of a contact is an estimate for the length of the loop that has to be closed to form this contact in an unfolded protein chain without crosslinks. For large loops, the logarithm of the loop length is proportional to the loop closure-entropy for forming this contact in the unfolded state [68,123,149–152]. The logarithm of the CO thus can be interpreted as a loop-closure entropy. The relative logCO is the average over the logarithm of the COs for all native contacts, multiplied by a prefactor $1/\log(N)$ where N is the chain length. To interpret this prefactor, it is important to note that the average over the logarithm of the COs clearly overestimates the folding barrier. The reason is that the loop-closure cost

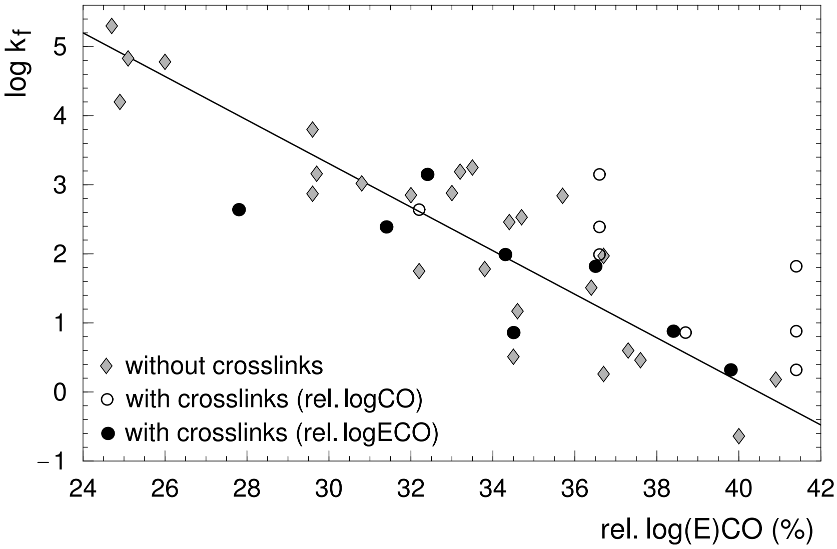

Figure 3.4: Relative logCO of 26 two-state proteins without crosslinks (gray diamonds), relative logCO of 8 two-state proteins with crosslinks (open circles), and relative logECO of these 8 proteins (filled circles) plotted against the decadic logarithm of their folding rates. The regression line for the 26 proteins without crosslinks is given by $\log k_f = 12.77 - 0.315 \times$ (rel. logCO). The standard deviation in vertical direction from the regression line is 0.70 for the filled circles, which is only slightly larger than the standard deviation 0.67 for the gray diamonds. This indicates that the relative logECO provides a simple, topology-based estimator for the folding rates of proteins both with and without crosslinks. In the absence of crosslinks, the relative logECO is identical with the relative logCO.

for contacts formed late in the folding process can be reduced by contacts that have been formed earlier [153–155], see next section. This overestimate should increase with the chain length N. The prefactor $1/\log(N)$ therefore may be seen as a heuristic, chain-length dependent correction of this overestimate, and the relative logCO as a naive estimate of entropic loop-closure barriers for proteins without crosslinks.

Topological measures without chain-length dependent prefactors exhibit weaker correlations with the folding rates of two-state proteins. In the case of the relative CO, the prefactor is $1/N$. The related topological measure without this prefactor has been termed absolute CO [17, 147]. For the 26 proteins without crosslinks considered here, the correlation coefficient between absolute CO and the folding rates is 0.69, significantly smaller than the correlation coefficient 0.92 for the relative CO. The correlation coefficient for the absolute logCO $= \sum_{i=1}^{M} \log\left[\mathrm{CO}(i)\right]/M$ is 0.80. This correlation coefficient is significantly smaller than the cofficient 0.90 for the relative logCO.

Clearly, simply topological measure have limitations in reproducing or predicting folding rates. One of these limitations seems to be exemplified by the three src SH3 domain mutants with crosslinks listed in table 3.3. The mutant with crosslink between residues 35 and 50 has the largest folding rate among the mutants. But the relative ECO and logECO of this mutant are only slightly smaller than the corresponding values for the mutant with crosslink between residues 1 and 25, and larger than the values for the circularized mutant with crosslink between residues 1 and 56. The reason seems to be that the crosslink between residues 35 and 50 stabilizes the hairpin between the strands β_3 and β_4 of the src SH3 domain. Mutational analysis of the wildtype src SH3 domain indicates that this β-hairpin is a central structural element in the transition state for folding [30] (see also fig. 3.9 in section 3.5). This seems to explain why crosslinking the hairpin has a particularly strong impact on the folding rate. The effect of native-state topology and crosslinks on the kinetics thus can also depend on structural details of transition states or native states beyond the overall localness of contacts in these states.

3.4 Effective contact order and folding routes

The correlations between protein folding rates and simple topological measures inspired the development of statistical-mechanical models based on native-state topology. These models can be grouped into three classes. First, there are models that use explicit representations of the protein chain with Go-type energy potentials [52, 69, 156–165], named after the Japanese physicist Nobuhiro Go [166]. In these potentials, amino acids that are in contact in the native structure attract each other, while amino acids not in contact in the native structure repel each other, irrespective, at least to some extent, of the physical interactions between the amino acids. The second class of models assumes that amino acids can be in either of two states: native-like structured, or unstructured [66–68, 70, 71, 167–174]. Partially folded states then are described by sets of structured amino acids. These models are inspired by the Zimm-Bragg model for helix-coil transitions [175], which assumes that amino acids in helices can either be in a helix or coil state. In the third class of models, partially folded states are characterized by the subset of native contacts formed in these states [153–155, 176, 177]. Other approaches that do not directly fall in one these three classes are the diffusion-collision model of

Table 3.4: Loop length (ECO) for forming the strand pairing $\beta_1\beta_4$ of CI2

structural elements formed prior	minimum ECO for $\beta_1\beta_4$
$\alpha + \beta_2\beta_3 + \beta_3\beta_4$	7
$\alpha + \beta_2\beta_3$	12
$\beta_2\beta_3 + \beta_3\beta_4$	19
$\alpha + \beta_3\beta_4$	24
$\beta_2\beta_3$	23
α	31
$\beta_3\beta_4$	36
–	42

The given ECOs are minimum ECOs among all contacts of the cluster $\beta_1\beta_4$, also termed *cluster ECOs.* The contact clusters are defined as in Ref. [155].

Karplus and Weaver [178–180] as well as free-energy-functional [181,182] and perturbed-Gaussian-chain methods [72,183].

Folding routes can be predicted from native contact maps rather directly via the concept of effective contact order (ECO). The ECO is an estimate for the length of the loop that has to be closed to form a contact or contact cluster in a partially folded chain conformation (see fig. 3.1). The contact clusters in a native contact map represent the characteristic structural elements of a protein. Molecular Dynamics simulations indicate increased correlations between contacts of the same contact cluster (see section 4.4). In a coarse-grained view, individual folding routes can be described by the sequence in which contact clusters are formed. For the protein CI2, which has just four contact clusters, there are $4! = 24$ possible sequences in which the clusters can be formed. The length of the loop that has to be closed to form a contact cluster in general depends on the sequence in which the clusters are formed. For example, the contact cluster $\beta_1\beta_4$ in the contact map of fig. 3.2(a) represents contacts between the two terminal strands β_1 and β_4 of CI2. Forming this contact cluster from the fully unfolded state, i.e. *prior* to the other three clusters, requires to close a relatively large loop of length 42 (see table 3.4). However, forming $\beta_1\beta_4$ *after* the other three clusters α, $\beta_2\beta_3$, and $\beta_3\beta_4$ requires only to close a relatively small loop of length 7. The reason is that the contacts of the clusters α, $\beta_2\beta_3$, and $\beta_3\beta_4$ short-circuit the chain, which brings the two chain ends with the strands β_1 and β_4 into closer spatial proximity. On the minimum-ECO route defined below, i.e. the folding route that minimizes the loop lengths and, thus, the entropic

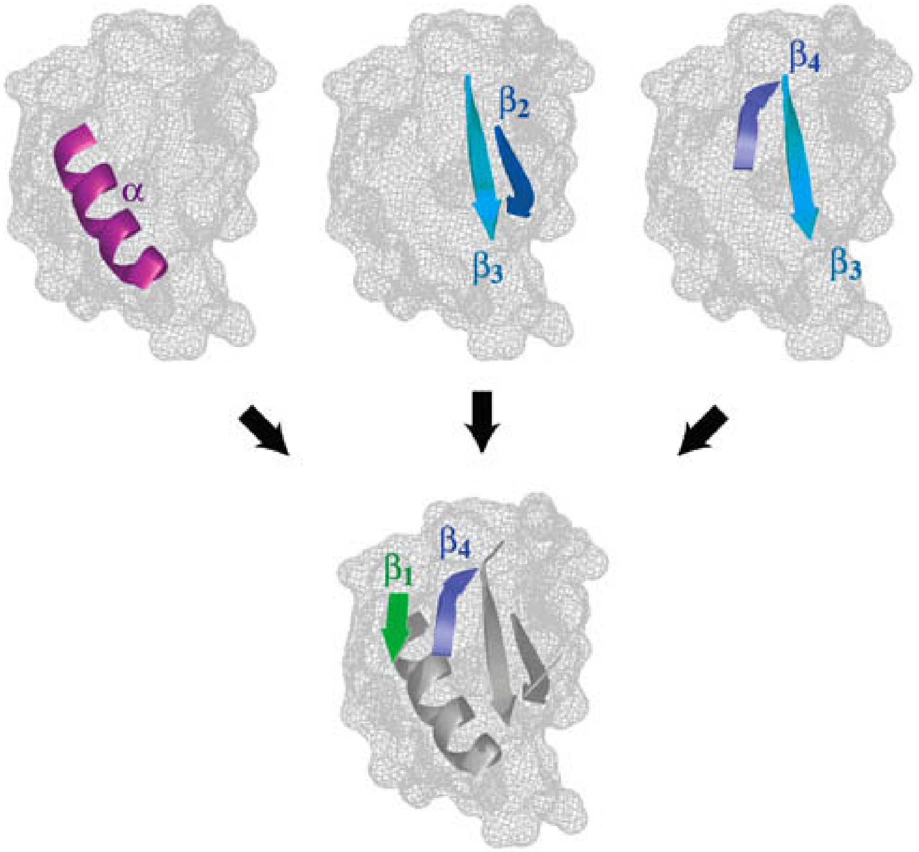

Figure 3.5: Minimum-ECO, or minimum-entropy-loss route of the protein CI2. Along this route, the strand pairing $\beta_1\beta_4$ is formed after the other three structural elements, the α-helix and the strand pairings $\beta_2\beta_3$ and $\beta_3\beta_4$. The route minimizes the length of the loop that has to be closed to bring the two terminal strands β_1 and β_4 into contact (see table 3.4).

loop-closure barriers, the cluster $\beta_1\beta_4$ is formed after the cluster α, $\beta_2\beta_3$, and $\beta_3\beta_4$ [153, 155]. On this route, α, $\beta_2\beta_3$, and $\beta_3\beta_4$ form in parallel since the ECOs of these three clusters do not depend on the sequence in which they are formed (see fig. 3.5).

To determine the minimum-ECO routes, we consider all possible sequences of cluster formation [155]. Sequences of cluster formation are called *folding sequences*. The formation of each contact cluster in a folding sequence requires to close a loop. The length of this loop is estimated as the minimum ECO among all cluster contacts, the *cluster ECO*. The cluster ECO thus is an estimate for the length of the shortest loop that has to be closed to form the cluster in a given partially folded conformation. Suppose we have a sequence of clusters $C_1C_2 \dots C_n$. Since no contacts have been formed prior to C_1, the ECO ℓ_1 of this cluster simply is the minimum CO among the cluster contacts. For the other clusters C_i in the folding sequence, the cluster ECO is the minimum ECO among the cluster contacts, given the contacts of the previously formed clusters $C_1, C_2, \dots, C_{i-1}$. This leads to a sequence of cluster ECOs, or loop lengths, $\ell_1, \ell_2, \dots, \ell_n$.

For each folding sequence $C_1C_2 \dots C_n$, the total loop-closure cost can be defined as $s = \sum_{i=1}^{n} f(\ell_i)$ where ℓ_i are the cluster ECOs along the sequence, and $f(\ell_i)$ is a weighting function which increases with the loop length ℓ_i. For simplicity, the linear weighting function $f(\ell_i) = \ell_i$ is used here [176]. This linear approximation for the free-energy cost of loop closure is not unreasonable since the range of relevant ECOs only spans roughly one order of magnitude, from 2 to 20 or 30 (see table 3.5). The total loop-closure cost then simply is the sum of ECOs $s = \sum_{i=1}^{n} \ell_i$ for all clusters along the sequence.

The *minimum-ECO sequences* to a given cluster C_n are simply defined as local minima of the loop-closure cost s in the space of all possible folding sequences to C_n. In this space, the neighbors of a given folding sequence $C_1C_2 \dots C_n$ are those sequences which are obtained either by deleting one or several of the clusters from $C_1C_2 \dots C_n$, or by adding one or several 'new' clusters somewhere in the sequence. In principle, two neighboring folding sequences can have the same local minimum value of s. In this case, the longer sequence among the two is selected as the minimum-ECO sequence.

Finally, all minimum-ECO sequences which consist of the same set of clusters are taken to represent the same *minimum-ECO route*. These sequences have the same loop-closure cost s and differ only by permutations from each other, which indicates parallel folding processes on the route. Suppose the ECO of the nonlocal cluster C_3 is only affected by the two local clusters C_1 and C_2. Since the ECOs of the local clusters C_1 and C_2 are independent of each other, the two sequences $C_1C_2C_3$ and $C_2C_1C_3$ then both are minimum-ECO sequences, representing the same minimum-ECO route. On this minimum-ECO route, the two local clusters C_1 and C_2 form in parallel, prior to the nonlocal cluster C_3.

Table 3.5 summarizes the loop-closure hierarchies on the minimum-ECO routes for 14 two-state proteins [155]. The native contact maps of the proteins are shown in the fig. 3.6, 3.7, and 3.8. These proteins (i) are small in the sense that they have less than 10 contact clusters, and (ii) are well-characterized in the sense that Φ-values for at least 10 residue positions are available. A 15th protein, ACBP, also considered in Ref. [155] is excluded here since recent experiments indicate that it

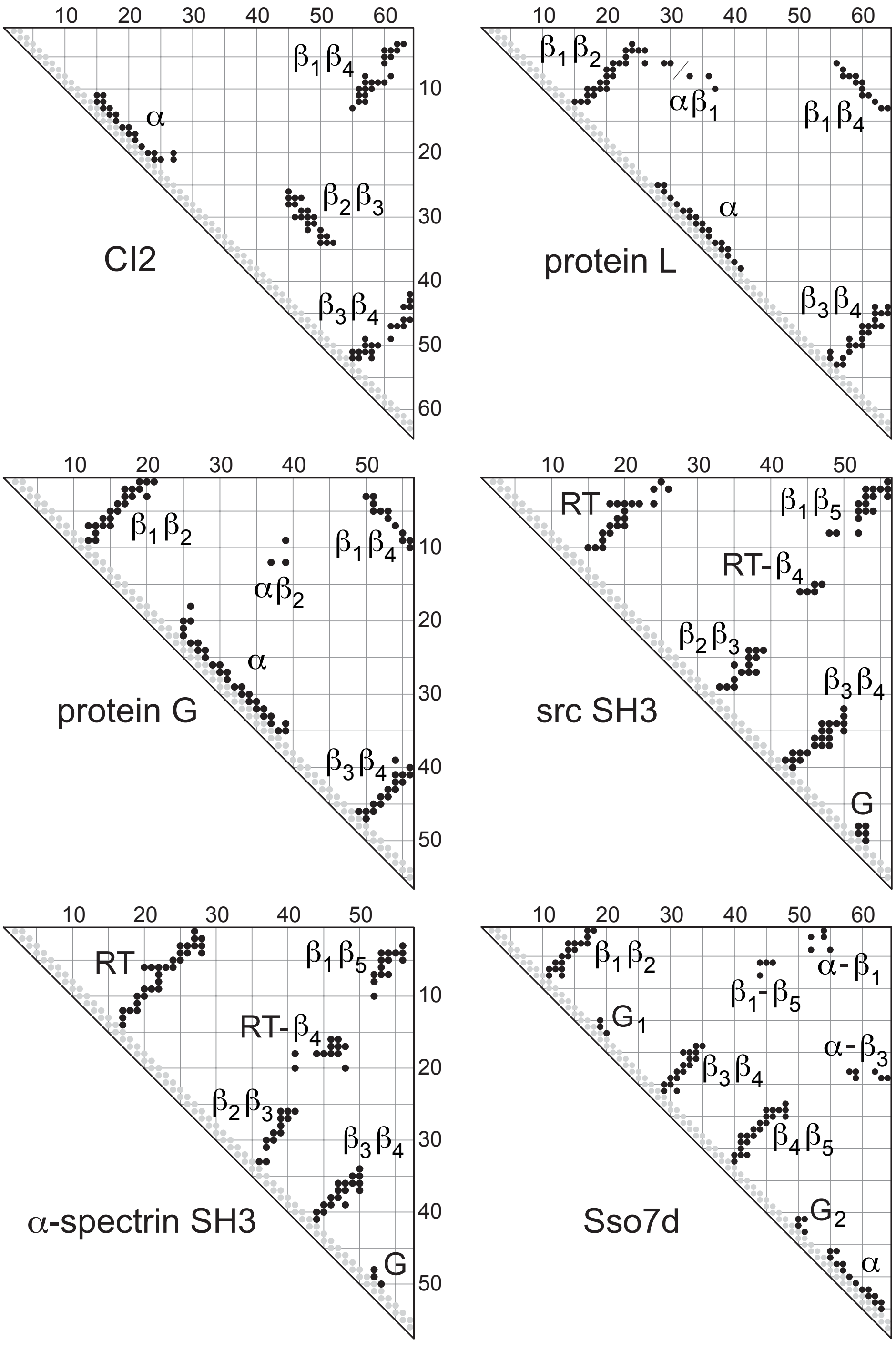

Figure 3.6: Contact maps of CI2 (protein data bank file 1COA), protein L (2PTL, residues 15 to 78), protein G (1PGB), the src SH3 domain, (1SRL), α-spectrin SH3 domain (1SHG), and Sso7d (1BNZ). Two residues are taken to be in contact if the distance between their C_α or C_β atoms is less than 6 Å. The contact clusters are defined as in Ref. [155].

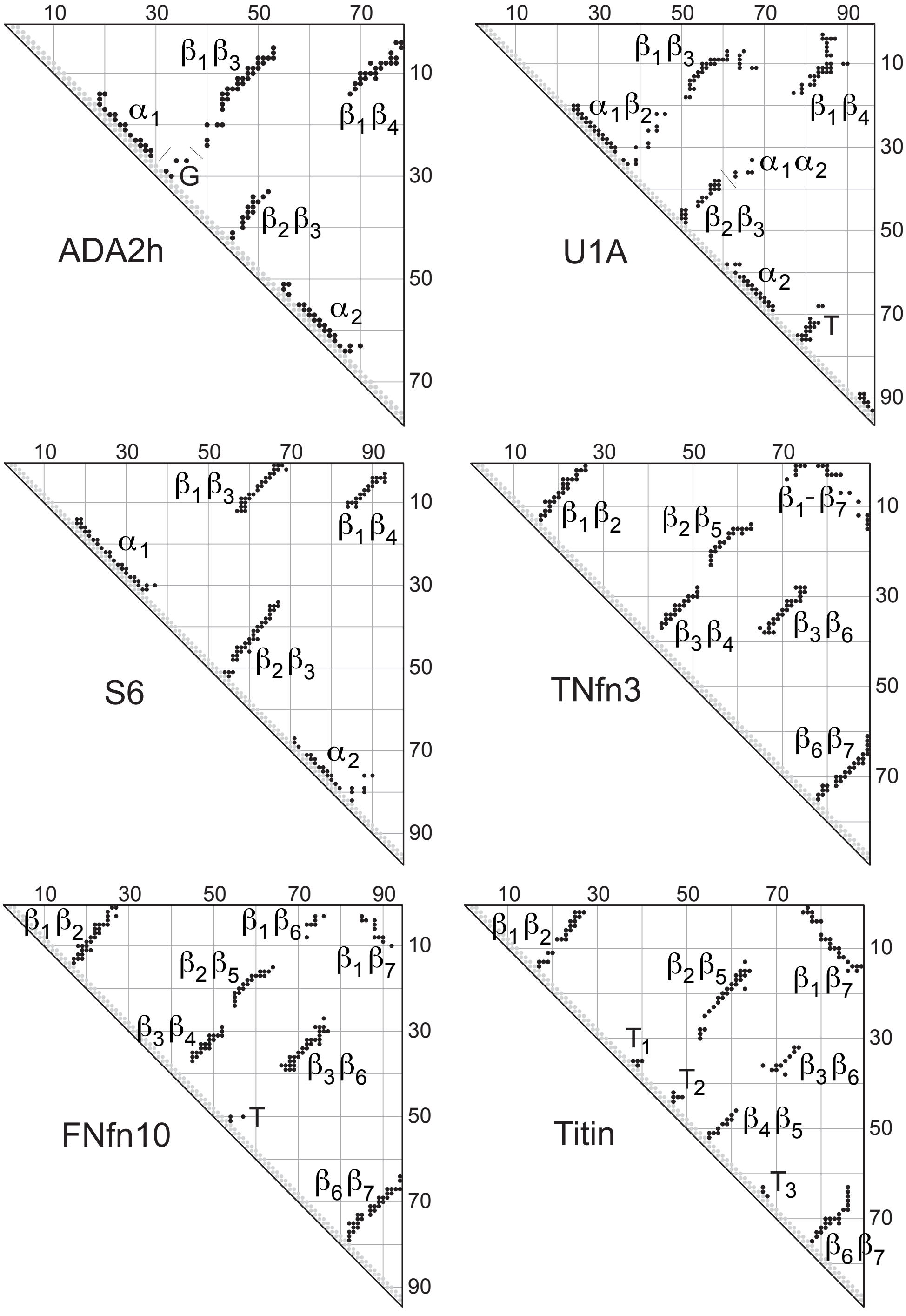

Figure 3.7: Contact maps of the proteins ADA2h (protein data bank file 1AYE), U1A (1URN, chain A), S6 (1RIS), TNfn3 (1TEN), FNfn10 (1FNF, residues 1416 to 1509), and Titin (1TIT). As in fig. 3.6, two residues are taken to be in contact if the distance between their C_α or C_β atoms is less than 6 Å. The contact clusters are defined as in Ref. [155].

Table 3.5: Loop-closure events on minimum-ECO routes

protein	nonlocal cluster	clusters formed prior	ECO for nonlocal cluster	loop-closure cost
CI2	$\beta_2\beta_3$	—	16	16
	$\beta_1\beta_4$	α_1, $\beta_3\beta_4$, $\beta_2\beta_3$,	7	27
protein L	$\alpha\beta_1$	$\beta_1\beta_2$	6	9
	$\beta_1\beta_4$	$\beta_1\beta_2$, α, $\beta_3\beta_4$, $\alpha\beta_1$	9	22
protein G	$\alpha\beta_2$	α	10	13
	$\beta_1\beta_4$	$\beta_1\beta_2$, α, $\beta_3\beta_4$	9	18
		or: α, $\beta_3\beta_4$, $\alpha\beta_2$	3	19
src SH3	RT-β_4	$\beta_2\beta_3$, $\beta_3\beta_4$	10	17
	$\beta_1\beta_5$	RT, $\beta_2\beta_3$, $\beta_3\beta_4$	5	17
α-spSH3	RT-β_4	$\beta_2\beta_3$	7	10
	$\beta_1\beta_5$	RT, $\beta_2\beta_3$, $\beta_3\beta_4$, G	5	17
		or: RT, $\beta_2\beta_3$, G, RT-β_4	3	17
Sso7d	α-β_3	$\beta_3\beta_4$, $\beta_4\beta_5$, α	7	16
	β_1-β_5	$\beta_1\beta_2$, $\beta_3\beta_4$, $\beta_4\beta_5$	9	18
	α-β_1	$\beta_1\beta_2$, $\beta_3\beta_4$, $\beta_4\beta_5$	12	21
ADA2h	$\beta_1\beta_3$	G	8	11
	$\beta_1\beta_4$	α_1, G, $\beta_2\beta_3$, α_2	9	21
		or: G, α_2, $\beta_1\beta_3$	7	21
U1A	$\alpha_1\alpha_2$	$\beta_2\beta_3$, α_2	3	9
	$\beta_1\beta_3$	$\alpha_1\beta_2$, $\beta_2\beta_3$	6	12
	$\beta_1\beta_4$	$\alpha_1\beta_2$, $\beta_2\beta_3$, T, $\beta_1\beta_3$	2	17
S6	$\beta_1\beta_3$	α_1, $\beta_2\beta_3$	14	20
	$\beta_1\beta_4$	α_1, $\beta_2\beta_3$, α_2	14	23
TNfn3	$\beta_2\beta_5$	$\beta_3\beta_4$	9	15
	β_1-β_7	$\beta_3\beta_4$, $\beta_6\beta_7$, $\beta_2\beta_5$	2	20
	$\beta_3\beta_6$	$\beta_1\beta_2$, $\beta_3\beta_4$, $\beta_6\beta_7$, $\beta_2\beta_5$, β_1-β_7	4	27
FNfn10	$\beta_2\beta_5$	$\beta_3\beta_4$	9	17
	$\beta_3\beta_6$	T	22	25
	$\beta_1\beta_6$	$\beta_1\beta_2$, T, $\beta_3\beta_6$	2	30
		or: $\beta_3\beta_4$, $\beta_6\beta_7$, $\beta_2\beta_5$, $\beta_1\beta_7$	2	31
	$\beta_1\beta_7$	$\beta_3\beta_4$, $\beta_6\beta_7$, $\beta_2\beta_5$	9	29
		or: $\beta_1\beta_2$, T, $\beta_6\beta_7$, $\beta_3\beta_6$, $\beta_1\beta_6$	2	34
Titin	$\beta_2\beta_5$	T_1, T_2	14	20
	$\beta_3\beta_6$	$\beta_4\beta_5$, T_3	14	20
	$\beta_1\beta_7$	T_1, T_2, $\beta_6\beta_7$, $\beta_2\beta_5$	2	25
		or: $\beta_1\beta_2$, $\beta_4\beta_5$, T_3, $\beta_3\beta_6$	8	31
CspB	$\beta_1\beta_4$	$\beta_1\beta_2$, $\beta_2\beta_3$, T	4	19
	$\beta_3\beta_5$	$\beta_4\beta_5$	14	17
L23	$\beta_1\beta_2$	α_1	3	6
	$\beta_3\beta_4$	–	12	12
	t-α_2	α_1, α_2, $\beta_1\beta_2$	5	14
	$\beta_2\beta_4$	α_2, $\beta_3\beta_4$	6	21

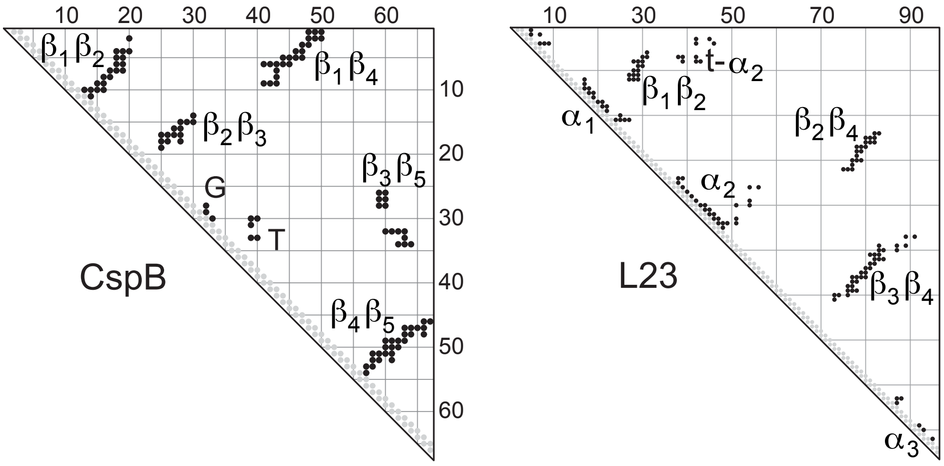

Figure 3.8: Contact maps of the two-state proteins CspB (protein data bank file 1CSP) and L23 (1N88). Two residues are taken to be in contact if the distance between their C_α or C_β atoms is less than 6 Å.

is not a two-state protein [184]. For each nonlocal cluster of a protein, all clusters formed prior on the minimum-ECO route are shown in table 3.5. Nonlocal clusters are clusters that do not contain contacts (i, j) with small CO $= |i - j| < 10$. For some nonlocal clusters, there are multiple minimum-ECO routes. These multiple routes correspond to different local minima of the loop-closure cost s in the space of folding sequences. However, local minima with a loop-closure cost s which is by 10 or more larger than the global minimum are neglected. These local minima represent folding routes with significantly larger entropic barriers.

3.5 Kinetic impact and average Φ-values

Several experimental methods provide information on folding routes. The characterization of metastable, partially folded states of non-two-state proteins gives direct information on folding intermediates, provided these metastable states or 'on-route' to the native state, and not 'off-route' traps. Structural information on these intermediates can be obtained with hydrogen-exchange or NMR methods [11, 185–190]. Two-state proteins do not exhibit experimentally detectable, metastable intermediates during folding. Instead, the folding kinetics of many two-state proteins has been investigated via mutational analysis [24–43], see section 1.3. In a mutational analysis, a large number of mostly single-residue mutants of a protein is generated. For each mutant, the effect of the mutation

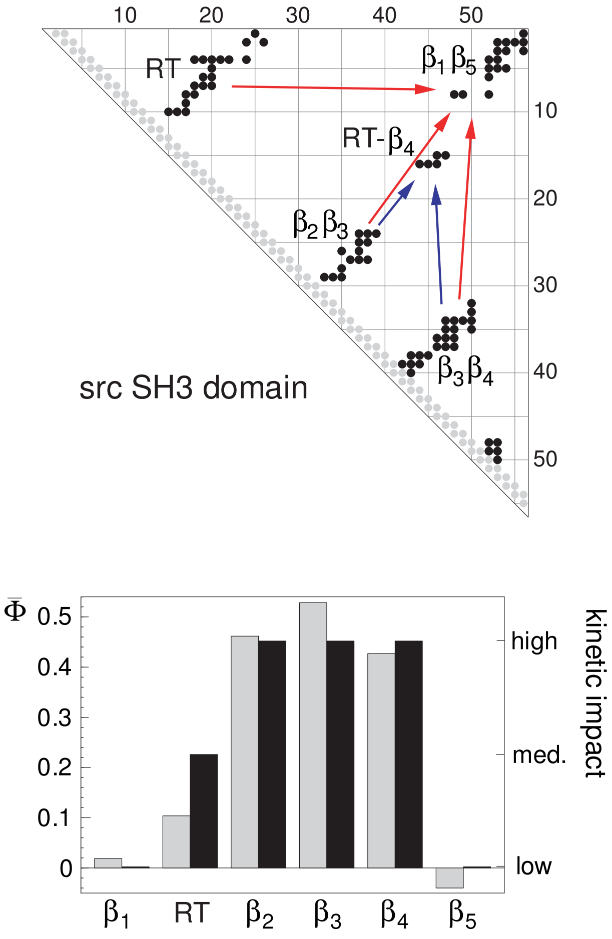

Figure 3.9: (Top) Minimum-ECO route of the src SH3 domain. The arrows indicate the sequences of events along this route. The red arrow pointing from the contact cluster RT to the cluster $\beta_1\beta_5$, for example, indicates that the RT loop is formed prior to the strand pairing $\beta_1\beta_5$. – (Bottom) Average experimental Φ-values [30] for the β-strands and the RT loop (grey bars) and kinetic impact estimated from the minimum-ECO route (black bars). The kinetic impact of the strands β_2, β_3, and β_4 is high since the clusters $\beta_2\beta_3$ and $\beta_3\beta_4$ are formed prior to both nonlocal clusters RT-β_4 and $\beta_1\beta_5$ [155]. The kinetic impact of RT is medium since the cluster is formed prior to only one of the nonlocal clusters, $\beta_1\beta_5$. The kinetic impact of β_1 and β_5 is low since the cluster $\beta_1\beta_5$ forms last, parallel to RT-β_4.

on the folding dynamics is quantified by its Φ-value [16, 44], defined in eq. (1.7).

Φ-values have been calculated in statistical-mechanical models that are based on native structures [52, 66–72, 158, 161, 163, 167] and from Molecular Dynamics unfolding simulations at elevated temperatures [49, 50, 63]. As discussed in chapter 2, the detailed modeling of Φ-values requires estimates for mutation-induced free-energy changes [89, 96, 110], which goes beyond simple topology-based modeling. However, on a more coarse-grained level, the level of average Φ-values for secondary structural elements, important aspects of Φ-value distributions are captured by native-state topology. An average Φ-value close to zero for a secondary structural element (i.e. a helix or a β-strand) indicates that mutations in the secondary element affect the folding rate only marginally, see eq. (1.7). In contrast, a large average Φ-value indicates that mutations have a strong impact on the folding rate. In a sense, the average Φ-values thus capture the 'kinetic impact' of secondary elements. The kinetic impact can also be estimated from minimum-ECO routes. The minimum-ECO route of the src SH3 domain is shown in fig. 3.9. Here, an arrow pointing from a contact cluster A to a cluster B in the contact map indicates that A is formed prior to B. On the minimum-ECO route, the contact cluster RT-β_4 forms after $\beta_2\beta_3$ and $\beta_3\beta_4$, and the cluster $\beta_1\beta_5$ after RT, $\beta_2\beta_3$ and $\beta_3\beta_4$. The clusters RT-β_4 and $\beta_1\beta_5$ are nonlocal clusters and form in parallel on this route. The other three large contact clusters, the RT loop, an irregular hairpin-like structure, and the two β-hairpins $\beta_2\beta_3$ and $\beta_3\beta_4$, are local clusters. Local clusters contain contacts with small CO and, thus, are located close to the diagonal of the contact map.

As discussed in detail below, the kinetic impact of a contact cluster can be estimated mainly from how often the cluster appears along the minimum-ECO route to other clusters [155]. The kinetic impact here is a semi-quantitative concept and can attain the values high, medium, or low. The kinetic impact of the clusters $\beta_2\beta_3$ and $\beta_3\beta_4$ of the src SH3 domain, for example, is high since these clusters appear on the route to both nonlocal clusters. Therefore, the kinetic impact of the three β-strands β_2, β_3, and β_4 is high. The kinetic impact of RT is medium since the cluster only appears on the route to $\beta_1\beta_5$. The kinetic impact of $\beta_1\beta_5$ and, thus, of the strands β_1 and β_5 is low since this cluster form last. The kinetic impact derived from the minimum-ECO route agrees with average Φ-values for the secondary elements, see fig. 3.9. The Φ-value distribution of the src SH3 domain is *polarized*, i.e. the average Φ-values are large for some of the secondary elements (the strands β_2, β_3, and β_4), and small for others (the strands β_1 and β_5).

More precisely, we have formulated three rules [155] to estimate the kinetic impact of contact clusters and secondary elements from the loop-closure hierarchies summarized in table 3.5. First, as mentioned above, it seems reasonable to assume that the kinetic impact of a cluster should be related to how often it appears on the minimum-ECO routes to other clusters. Suppose a local cluster appears on minimum-ECO routes to all non-local clusters. Mutations affecting the formation of this cluster then should strongly affect the overall folding kinetics. Hence, the cluster has a high kinetic impact. To quantify this notion, the *occurrence number* n of a cluster is defined as the number of times it appears on all routes to all (other) nonlocal clusters. In other words, n simply is the number

Table 3.6: Average Φ-values and kinetic impact of secondary structures

CI2	β_1	α	β_2	β_3	β_4		
$\bar{\Phi}_{\mathrm{exp}}$	0.23 (1)	0.32 (12)	0.15 (4)	0.32 (4)	0.03 (2)		
kin. impact	M	H	H	H	H		
protein L	β_1	β_2	α	β_3	β_4		
$\bar{\Phi}_{\mathrm{exp}}$	0.36 (9)	0.46 (7)	0.15 (16)	0.18 (4)	0.14 (7)		
kin. impact	H	H	M	M	M		
protein G	β_1	β_2	α	β_3	β_4		
$\bar{\Phi}_{\mathrm{exp}}$	0.36 (3)	−0.16 (4)	0.13 (9)	0.63 (2)	0.27 (4)		
kin. impact	M	M	H	H	H		
src SH3	β_1	RT	β_2	β_3	β_4	G	β_5
$\bar{\Phi}_{\mathrm{exp}}$	0.02 (4)	0.10 (8)	0.46 (3)	0.53 (6)	0.43 (6)	–	−0.04 (2)
kin. impact	L	M	H	H	H	L	L
α-spec SH3	β_1	RT	β_2	β_3	β_4	G	β_5
$\bar{\Phi}_{\mathrm{exp}}$	0.08 (2)	0.26 (3)	−0.20 (1)	0.66 (3)	0.60 (2)	0.53 (1)	0.16 (1)
kin. impact	L	H	H	H	M	H	L
Sso7d	β_1	β_2	G_1	β_3	β_4	β_5	α
$\bar{\Phi}_{\mathrm{exp}}$	−0.03 (2)	0.11 (2)	−0.03 (2)	0.96 (2)	0.27 (4)	0.19 (5)	0.41 (4)
kin. impact	H	H	L	H	H	H	H
ADA2h	β_1	α_1	G	β_2	β_3	α_2	β_4
$\bar{\Phi}_{\mathrm{exp}}$	0.42 (3)	0.26 (3)	–	0.06 (2)	0.29 (3)	0.49 (4)	0.14 (2)
kin. impact	M	M	H	M	M	H	M
U1A	β_1	α_1	β_2	β_3	α_2	β_4	
$\bar{\Phi}_{\mathrm{exp}}$ ($\beta = 0.5$)	0.23 (2)	0.38 (3)	0.73 (3)	–	0.00 (1)	0.00 (1)	
$\bar{\Phi}_{\mathrm{exp}}$ ($\beta = 0.7$)	0.43 (2)	0.63 (3)	0.98 (3)	–	0.50 (1)	0.23 (1)	
kin. impact	M	H	H	H	L	L	
S6	β_1	α_1	β_2	β_3	α_2	β_4	
$\bar{\Phi}_{\mathrm{exp}}$	0.34 (4)	0.25 (4)	0.24 (1)	0.31 (5)	0.28 (2)	0.14 (2)	
kin. impact	H	H	H	H	M	H	
TNfn3	β_1	β_2	β_3	β_4	β_5	β_6	β_7
$\bar{\Phi}_{\mathrm{exp}}$	0.12 (3)	0.27 (2)	0.36 (3)	0.55 (2)	0.47 (2)	0.42 (3)	0.11 (5)
kin. impact	M	H	H	H	H	H	H
FNfn10	β_1	β_2	β_3	β_4	β_5	β_6	β_7
$\bar{\Phi}_{\mathrm{exp}}$	0.3 (3)	−0.16 (2)	0.55 (3)	0.35 (2)	0.29 (2)	0.44 (4)	0.73 (1)
kin. impact	H	H	H	H	H	H	H
Titin	β_1	β_2	β_3	β_4	β_5	β_6	β_7
$\bar{\Phi}_{\mathrm{exp}}$	0.09 (3)	0.53 (4)	0.51 (2)	0.54 (2)	0.66 (3)	0.66 (3)	0.07 (3)
kin. impact	M	H	H	H	H	H	M
CspB	β_1	β_2	β_3	G	β_4	β_5	
$\bar{\Phi}_{\mathrm{exp}}$	0.64 (6)	0.27 (4)	0.75 (1)	−0.06 (2)	0.16 (2)	0.12 (2)	
kin. impact	H	H	H	L	H	H	
L23	β_1	α_1	β_2	α_2	β_3	β_4	α_3
$\bar{\Phi}_{\mathrm{exp}}$	0.08 (1)	0.03 (1)	0.20 (2)	0.34 (2)	0.10 (3)	0.29 (3)	0.02 (1)
kin. impact	L	L	M	H	H	H	L

Caption of table 3.6: The average Φ-values have been calculated from data published in the following articles: CI2 [24], protein L [33], protein G [34], src SH3 [30], α-spectrin SH3 [29], Sso7d [70], ADA2h [25], U1A [27], S6 [36], TNfn3 [32], FNfn10 [102], Titin [101], CspB [40], L23 [97]. The number in brackets behind an average Φ-value indicates the number of residues in the secondary element for which Φ-values have been measured. Averages taken from many Φ-values are more reliable. The kinetic impact of the secondary elements is derived from the results shown in table 3.5 and can attain the values low (L), medium (M), or high (H). Where possible, secondary structure classifications given in the protein data bank structure files of the proteins have been used to calculate average Φ-values, see Ref. [155] for details.

of times the cluster occurs in the third column of table 3.5. In terms of occurrence numbers, the first rule is:

(1) The kinetic impact of a cluster is high (H) if its occurrence number n on the minimum-ECO routes is larger than or equal to $\frac{2}{3}n_{\max}$. Here, n_{max} is the maximum value of n among all clusters of the protein. The impact of the cluster is medium (M) for $\frac{1}{3}n_{\max} \leq n < \frac{2}{3}n_{\max}$. The impact is low (L) for $n < \frac{1}{3}n_{\max}$.

Second, the kinetic impact of nonlocal clusters should also be affected by the cluster ECO. Suppose a nonlocal cluster has a high cluster ECO on all minimum-ECO routes. This means that forming the cluster always involves the closure of a relatively large loop. It seems reasonable to assume that the kinetic impact of the cluster then is high, since the contacts of these clusters have to balance a relatively high loop-closure entropy. In other words, the formation of the cluster and, hence, the overall folding kinetics should be highly sensitive to mutations affecting the cluster contacts. The second rule is:

(2) A nonlocal cluster has a high (H) kinetic impact if the ECO of this cluster is larger than 10 an all routes. The kinetic impact is medium (M) if the smallest cluster ECO has a value from 6 to 10, unless rule (1) specifies high impact.

According to the rules (1) and (2), the kinetic impact of a cluster thus is low if its occurrence number is small, and the cluster ECO is not larger than 5.

Finally, suppose a protein has two nonlocal clusters C_1 and C_2 which fold in parallel. This means that the cluster C_1 does not appear on the minimum-ECO routes to C_2, and vice versa. In general, the loop-closure cost for forming, e.g, C_1 can be significantly larger than the loop-closure cost for forming C_2. It seems reasonable that clusters appearing on the minimum-ECO routes to C_1 should then have a higher kinetic impact than clusters appearing only on minimum-ECO routes to C_2, since the entropic loop-closure barrier for forming C_1 is significantly larger. Therefore, the third rule is:

(3) If two nonlocal clusters C_1 and C_2 do not occur on minimum-ECO routes to other clusters and have minimum loop-closure costs s_1 and s_2 with $s_1 > s_2 + 5$, the cluster occurrences on the routes to C_2 are not taken into account in rule (1). In particular, clusters which appear only on routes to C_2 have a low kinetic impact, independent of their ECO.

The rules (1), (2), and (3) define the kinetic impact of clusters. The translation into kinetic impact of secondary elements (strands or helices) is straightforward. The kinetic impact of a secondary element is high (H) if it has contacts in a cluster with high kinetic impact, and low (L) if it only has contacts in clusters with low kinetic impact. The kinetic impact of a secondary element is medium (M) if it has contacts in clusters with medium kinetic impact, but no contacts in clusters with high kinetic impact. As an example, the high kinetic impact of the clusters α_i and $\beta_k\beta_l$ of a protein results in a high kinetic impact of the secondary elements α_i, β_k, and β_l. The relation between secondary elements and contact clusters is summarized in the cluster labels of the figs. 3.6, 3.7, and 3.8.

Table 3.6 shows average experimental Φ-values and kinetic impact for the strands and helices of the 14 proteins considered here. To illustrate the rules (1) and (2), let us consider again the minimum-ECO route of the src SH3 domain given in table 3.5 and illustrated in fig. 3.9. This protein has two nonlocal clusters, RT-β_4 and $\beta_1\beta_5$. The clusters $\beta_2\beta_3$ and $\beta_3\beta_4$ appear on the minimum-ECO routes to both nonlocal clusters and, hence, have the occurrence number 2. The cluster RT only appears on the route to $\beta_1\beta_5$ and, hence, has occurrence number 1. According to rule (1), the kinetic impact of $\beta_2\beta_3$ and $\beta_3\beta_4$ thus is high (H), and the kinetic impact of RT is medium (M). According to rule (2), the kinetic impact of the cluster RT-β_4 is medium since it has the cluster ECO 10. Finally, the kinetic impact of $\beta_1\beta_5$ is low (L) since it has a small cluster ECO of 5 and occurrence number 0. Therefore, the kinetic impact of the strands β_2, β_3, and β_4 is high, the kinetic impact of RT is medium, and the kinetic impact of β_1 and β_5 is low, in perfect agreement with the average Φ-values (see table 3.6 and fig. 3.9).

Rule (3) affects the proteins U1A and L23. In the case of U1A, the cluster $\alpha_1\alpha_2$ does not occur on the minimum-ECO routes to the two other nonlocal clusters $\beta_1\beta_3$ and $\beta_1\beta_4$ and has a significantly smaller

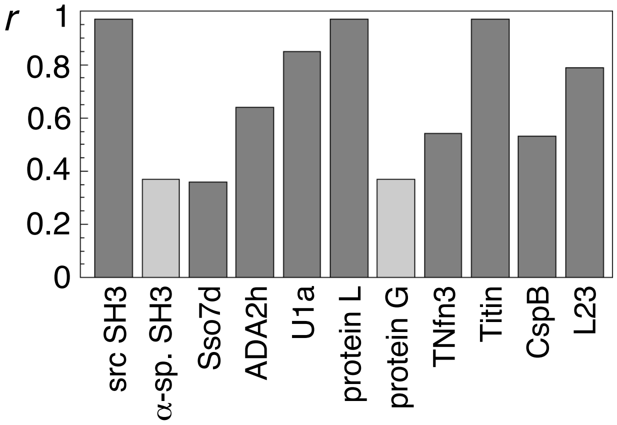

Figure 3.10: Correlation coefficients r for the comparison between average experimental Φ-values and kinetic impact of 11 proteins with polarized Φ-value distributions. The light grey bars represent the correlation coefficients of proteins with negative average Φ-values below -0.1 in one of the secondary elements. On average, the correlation coefficient is 0.67 for all 11 proteins, and 0.74 for the 9 proteins with positive Φ-values. For U1A, the correlation coefficients for the two Φ-value distributions at $\beta = 0.5$ and $\beta = 0.7$ (see table 3.6) are 0.91 and 0.79. Here, the average 0.85 of these two values is presented. To test the statistical significance of the observed correlations, one can compare the obtained correlation coefficient with those between the theoretical distribution and all possible random permutations of the experimental distribution for each of the proteins. The fraction p of random permutations of the experimental data which have an equally high or larger correlation coefficient with the theoretical distribution can be interpreted as probability to obtain the correlations shown in the figure, or larger ones, by chance. This probability is $p = 0.017$ for src SH3, $p = 0.20$ for α-spectrin SH3, $p = 0.10$ for Sso7d, $p = 0.17$ for ADA2h, $p = 0.033$ and 0.067 for U1A, $p = 0.033$ for protein L, $p = 0.30$ for protein G, $p = 0.29$ for TNfn3, $p = 0.026$ for Titin, $p = 0.17$ for CspB, and $p = 0.036$ for L23. Despite the relatively small number of data points (the proteins have between 4 and 7 secondary structural elements), the obtained correlations are statistically significant. The probability p for obtaining an average correlation coefficient of 0.67 or larger for all 12 proteins by chance is smaller than 10^{-6}.

loop-closure cost than $\beta_1\beta_4$. Therefore, α_2 has a low kinetic impact, since it only appears on the minimum-ECO route to $\alpha_1\alpha_2$. In the case of L23, the nonlocal clusters t-α_2 folds in parallel to $\beta_2\beta_4$, with significantly smaller loop-closure cost. As a consequence, the kinetic impact of β_1 and α_1 is low since these secondary elements are only involved in the folding of t-α_2.

The Φ-value distributions of two-state proteins are either polarized or diffuse. In a *diffuse* distribution, the average Φ-values for the secondary elements are of similar magnitude. A diffusive distribution of kinetic impact occurs, e.g., if all clusters are involved on an 'equal footing' in the formation of a single rate-limiting cluster. On the minimum-ECO route of CI2, for example, $\beta_1\beta_4$ forms after the other three clusters, which results in a diffuse distribution of kinetic impact, in agreement with the experimental Φ-value distribution. A polarized distribution, in contrast, occurs if some clusters have a central role on the minimum-ECO route, such as the clusters $\beta_2\beta_3$ and $\beta_3\beta_4$ of the src SH3 domain.

To quantify this notion, a Φ-value distribution here is defined as polarized if at least two average Φ-values are by more than a factor 2.5 smaller than the maximum value of the distribution. A Φ-value distribution is diffuse if this is not the case. In a diffuse distribution, all or all except one of the average Φ-values are larger than 40% of the maximum among these values. An analogous definition can also be applied to the distribution of kinetic impact derived from the minimum-ECO routes: The distribution is diffuse if all or all except one of the secondary elements have high kinetic impact.

According to this definition, three among the 14 proteins considered here have a diffuse Φ-value distribution. These proteins are CI2, S6, and FNfn10. In agreement with the experiments, the distribution of kinetic impact for the secondary structural elements of these proteins is also diffuse (see table 3.6). The remaining 11 proteins have polarized Φ-value distributions. Fig. 3.10 shows the Pearson correlation coefficient r between average Φ-values and kinetic impact for each of these proteins. The calculate the correlation coefficients, the values 0, 1, and 2 are assigned to the kinetic impact L, M, and H. Any other 'equidistant' values a, $a+b$, and $a+2b$ with $b>0$ for the kinetic impact L, M, and H result in the same correlation coefficients. Correlation coefficients are only given for proteins with polarized distributions since they do not reflect the quality of the modeling in the case of diffuse distributions with rather similar average Φ-values for the secondary elements. The correlation coefficient r can attain values in the range -1 to 1 where 1 means 'perfect' correlation (proportionality), 0 means no correlation, and negative values mean anticorrelation.

Two of the lowest correlation coefficients are obtained for the α-spectrin SH3 domain and protein G (see fig. 3.10). These proteins have clearly negative average Φ-values smaller than -0.1 in one of the secondary elements. Negative Φ-values can be captured be the models presented in chapter 2, but are beyond the simple topology-based modeling considered here. For the comparison with kinetic impact, the negative average Φ-values were simply taken to be zero. However, excluding the strand β_2

of the α-spectrin SH3 domain from the comparison leads to a correlation coefficient of 0.69 instead of 0.37. Thus, at least in case of the src SH3 domain, the relatively low correlation coefficient of fig. 3.10 can be traced back to the secondary element with negative Φ-value.

Two other proteins with relatively low correlation coefficients in fig. 3.10 are Sso7d and CspB. These proteins have in common that the nonlocal clusters fold in parallel on the minimum-ECO routes. In the case of CspB, the nonlocal clusters are $\beta_1\beta_4$ and $\beta_3\beta_5$. Since the total loop-closure cost of the two parallel folding processes leading to these clusters are similar (see table 3.5), the model takes them to be equally important for the kinetics. However, the experimental Φ-values seem to indicate that the folding process leading the $\beta_1\beta_4$ has a larger impact on the kinetics than the parallel process leading to $\beta_3\beta_5$. The strands β_1 to β_3 of the two clusters $\beta_1\beta_2$ and $\beta_2\beta_3$, which are formed prior to $\beta_1\beta_4$, have relatively large average Φ-values. In contrast, the strands of the cluster $\beta_4\beta_5$, which is formed prior to $\beta_3\beta_5$ on the parallel folding process, have significantly smaller average Φ-values. In the case of Sso7d, the three nonlocal clusters α-β_3, β_1-β_5 and α-β_1 fold in parallel, with comparable loop-closure cost. Here, the experimental Φ-values seem to indicate that the folding process leading to α-β_3 dominates the folding kinetics. According to the model, the clusters formed prior to α-β_3 are $\beta_3\beta_4$, $\beta_3\beta_5$, and α. The secondary elements of these clusters have medium are large average Φ-values, whereas the Φ-values of the remaining secondary structural elements β_1, β_2, and G_1 are significantly smaller. In both proteins, specific energetic interactions, which are not taken into account in the model, may be responsible for the dominance of one the parallel folding processes with similar entropic loop-closure barriers.

For the remaining majority of proteins, the model reproduces the polarized Φ-value distributions with relatively large correlation coefficients, which indicates that the shape of these Φ-value distributions can be traced back to native-state topology. In the model, the native-state topology is captured by the topology of the native contact maps, or more precisely, by the ECO-dependencies between the contact clusters. Interestingly, the model is able to reproduce the experimentally observed differences in the Φ-value distributions of protein L and G without sequence-specific information. These two proteins have very similar folds, but nonetheless small differences in their contacts maps (see fig. 3.6). Whereas protein L has a small tertiary $\alpha\beta_1$ cluster, protein G has a tertiary $\alpha\beta_2$ cluster. This results in different folding routes and different distributions of kinetic impact (see tables 3.5 and 3.6). In the case of protein L, the N-terminal hairpin $\beta_1\beta_2$ has higher kinetic impact than the C-terminal hairpin $\beta_3\beta_4$, in agreement with the average Φ-values. In the case of protein G, the kinetic impact and average Φ-values are larger for the C-terminal hairpin $\beta_3\beta_4$. Other groups have used sequence-specific interaction energies to reproduce these differences between protein L and G [72, 158, 161, 163].

3.6 Summary

In this chapter, we have focused on loop-closure aspects of the protein folding kinetics. In 1998, Plaxco and colleagues [73] made the remarkable observation that the folding rates of two-state proteins correlate with a simple measure of native-state topology, the relative CO (see section 3.1). Subsequently, comparable correlations have also been found for other simple measures of native-state topology (see table 3.1). The physical principle that underlies these correlations seems to be loop closure (see section 3.1).

In section 3.3, we have tested whether topological measures can be generalized to capture the effect of chain crosslinks on the folding rate. Crosslinks change the chain connectivity and therefore also the localness of some of the native contacts. These changes in localness can be taken into account by the graph-theoretical concept of effective contact order (ECO), see fig. 3.1. The relative ECO, however, the natural extension of the relative CO for proteins with crosslinks, appears to overestimate the changes in the folding rates caused by crosslinks (see fig. 3.3). But a closely related pair of measures, the relative logCO and relative logECO, captures the folding rates of two-state proteins with and without crosslinks (see fig. 3.4). The relative logCO is the average value for the logarithm of the CO of all contacts, divided by the logarithm of the chain length, and the relative logECO is the natural extension of this measure for crosslinked chains. The logarithm of the loop length of a contact is an estimate for the chain entropy loss caused by the loop closure [68, 123, 149–152]. The relative logCO and logECO therefore may be seen as naive measures of entropic folding barriers.

The graph-theoretical ECO concept leads rather directly to folding routes of proteins (see section 3.4). To predict routes, we have focused on the contact clusters in native contact maps. Contact clusters capture the overall topology of a protein structure (see section 3.2). In general, there are two scenarios for two contact clusters (or structural elements) A and B of a protein. In the first scenario, the ECOs (or loop lengths) of the contact clusters A and B do not depend on the sequence in which the clusters are formed. The two clusters then are predicted to form *parallel* to each other. In the second scenario, the ECO of one of the two clusters, e.g. cluster B, is significantly smaller if cluster A is formed prior to B. The clusters are then predicted to form *sequentially*, provided that the total loop-closure cost for cluster B along this route, which includes the loop-closure cost for cluster A, is smaller than on other routes (see section 3.4). An important point here is that the loop-closure dependencies between two contact clusters typically are strong in the second scenario, i.e. the differences in loop lengths are large if the sequence of events in which the clusters are formed is reversed. Therefore, simple estimates of loop-closure entropies [153, 176] or minimization of loop lengths as in section 3.4 are sufficient to derive the dominant minimum-ECO or minimum-entropy-loss routes. The minimum-ECO routes help to understand the distribution of average Φ-values for the secondary structural elements (α-helices and β-strands) of a protein (see section 3.5).

4 Folding cooperativity

4.1 Contact clusters and energy landscapes

The folding routes of section 3.4 are hierarchic in the sense that the formation of nonlocal structural elements typically requires the prior formation of other, more local structural elements. In this section and the following section, we will show that the hierarchic folding routes do not contradict cooperative two-state folding with a characteristic single-exponential relaxation kinetics (see section 1.2). In the model presented in this section, partially folded states are described by the contact clusters (or structural elements) formed in these states. An energy landscape is obtained by assigning free energies to each of these states, which include entropic terms that reflect the loop-closure dependencies between the contact clusters introduced in section 3.4. In the next section, we will show that single-exponential relaxation kinetics after an initial, fast 'burst phase' is obtained if the free energies for forming local structural elements such as α-helices and β-hairpins are positive. The formation of these local structural elements than constitutes a barrier on the free-energy landscape.

As in section 3.4, our model starts from the contact clusters in the native contact map of a protein. We assume that each contacts cluster is either fully formed or not formed, and neglect partial degrees of formation. Thus, for a protein with M clusters, there are 2^M possible states. Each of these states is characterized by a vector $n = \{n_1, n_2, \ldots, n_M\}$, where $n_i = 1$ indicates that cluster i is formed and $n_i = 0$ indicates that cluster i is not formed.

The free energy of a state n is given by [155]

$$G_n = \sum_{i=1}^{M} n_i \left[c \cdot \ell_i(n) + g_i \right] \tag{4.1}$$

Each cluster i that is formed ($n_i = 1$) contributes to the free energy G_n of the state n with two terms: A state-dependent free energy of loop closure $c \cdot \ell_i(n)$, and a free energy g_i for forming the cluster contacts. Here, c is a loop-closure parameter. The quantity $\ell_i = \ell_i(n)$ is the *cluster ECO* for cluster i. The cluster ECO is the length of the smallest loop that has to be closed in order to form the cluster. For a local cluster, the cluster ECO is the smallest CO among the contacts. For a nonlocal cluster, the cluster ECO depends on which other clusters are present in the state n.

In general, ECOs depend on the sequence in which contacts or contact clusters are formed. However, in order to apply the master equation formalism described below, we define here a cluster ECO that depends only on the state n, and not on the sequence of cluster formation. The free energy G_n defined in eq. (4.1) then is a state function, i.e. a function that does no depend on the route on which the state is attained. For this purpose, we use the following scheme: If only one nonlocal cluster is formed in a certain state, the cluster ECO is the smallest ECO among the cluster contacts, given all the local clusters formed in that state. If multiple nonlocal clusters are present in a state, we consider all the possible sequences along which these clusters can form, and select the sequence that has the smallest sum of ECOs. For instance, for a state with two nonlocal clusters C_i and C_j, there are two sequences: (1) $C_i \to C_j$, and (2) $C_j \to C_i$. The minimum ECOs for the clusters are determined sequentially: $\ell_i^{(1)}$ and $\ell_j^{(1)}$ along sequence (1), and $\ell_i^{(2)}$ and $\ell_j^{(2)}$ along sequence (2). If $\ell_i^{(1)} + \ell_j^{(1)}$ is smaller than $\ell_i^{(2)} + \ell_j^{(2)}$, the cluster ECOs ℓ_i and ℓ_j of the clusters i and j in the given state are taken to be $\ell_i^{(1)}$ and $\ell_j^{(1)}$. The cluster ECOs ℓ_i and ℓ_j are an estimate for the smallest loop lengths required to form the two clusters in the state.

In eq. (4.1), the free-energy cost of the loops is estimated by a simple linear approximation in the loop length. This is not unreasonable since the range of relevant ECOs only spans roughly one order of magnitude, from about $\ell = 3$ to $\ell = 30$ or 40. In general, determining the free energy of a chain molecule with multiple constraints or contacts is a complicated and unsolved problem. For the simpler problem of hairpin-like loop closures, several estimates have been given in the literature (see, e.g., [68, 150, 191]).

As in section 2.1, the folding dynamics of the model is described by the master equation

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} = \sum_{m \neq n} \left[w_{nm} P_m(t) - w_{mn} P_n(t) \right] \tag{4.2}$$

for the time evolution of the probability $P_n(t)$ that the protein is in state n at time t. Here, w_{nm} is the transition rate from state m to n. The master equation can be written in matrix form

$$\frac{\mathrm{d}\boldsymbol{P}(t)}{\mathrm{d}t} = -\boldsymbol{W}\boldsymbol{P}(t) \tag{4.3}$$

where $\boldsymbol{P}(t)$ is the vector with elements $P_n(t)$, and the matrix elements of $\boldsymbol{W}$ are given by

$$W_{nm} = -w_{nm} \quad \text{for} \quad n \neq m; \qquad W_{nn} = \sum_{m \neq n} w_{mn}. \tag{4.4}$$

The transition rates are defined as

$$w_{nm} = \frac{\delta_{|n-m|,1}}{t_o} \left[1 + \exp\left(\frac{G_n - G_m}{k_B T} \right) \right]^{-1} \tag{4.5}$$

where t_o is a reference time scale. The only transitions that are assigned to have nonzero rates w_{nm} are 'incremental' steps that change the state n by a single cluster unit. This is enforced by the term $\delta_{|n-m|,1}$ in eq. (4.5) where the Kronecker $\delta_{i,j}$ is one for $i = j$ and zero otherwise. The condition $|n - m| = 1$ is only satisfied by pairs of states $n = \{n_1, \ldots, n_M\}$ and $m = \{m_1, \ldots, m_M\}$ with $n_k \neq m_k$ for a single cluster k, and with $n_k = m_k$ for all other clusters. The transition rates (4.5) satisfy detailed balance, $w_{nm} P^e_m = w_{mn} P^e_n$ where $P^e_n \sim \exp[-G_n/(k_B T)]$ is the equilibrium weight for the state n. We have chosen here the 'Glauber dynamics' with $w_{nm} \sim (1 + \exp[(G_n - G_m)/(k_B T)])^{-1}$. Another standard choice satisfying detailed balance is the Metropolis dynamics, which should lead to equivalent results.

The detailed balance property of the transition rates implies that the eigenvalues of the matrix $\boldsymbol{W}$ are real. One of the eigenvalues is zero, corresponding to the equilibrium distribution, while all other eigenvalues are positive [88]. The solution to the master equation is given by

$$\boldsymbol{P}(t) = \sum_{\lambda} c_\lambda \boldsymbol{Y}_\lambda \exp[-\lambda t] \tag{4.6}$$

where $\boldsymbol{Y}_\lambda$ is the eigenvector corresponding to the eigenvalue λ, and the coefficients c_λ are determined by the initial condition $\boldsymbol{P}(t = 0)$. For $t \to \infty$, the probability distribution $\boldsymbol{P}(t)$ tends towards the equilibrium distribution $\boldsymbol{P}^e \sim \boldsymbol{Y}_0$ where $\boldsymbol{Y}_0$ is the eigenvector with eigenvalue $\lambda = 0$. The coefficients c_λ in eq. (4.6) depend on the initial condition at $t = 0$. As initial condition, we start from the fully unfolded state in which no clusters are formed, i.e. we choose a probability 1 for this state, and probabilities of 0 for all other states.

In principle, the model has $M + 1$ parameters for a protein with M clusters. These parameters are the loop-closure parameter c in eq. (4.1) and the free energy g_i of each cluster. But to reduce the number of parameters, we consider here only a simple version of the model in which all local clusters have the same free energy $g_i = g_l$, and all nonlocal clusters the free energy $g_i = g_{nl}$.

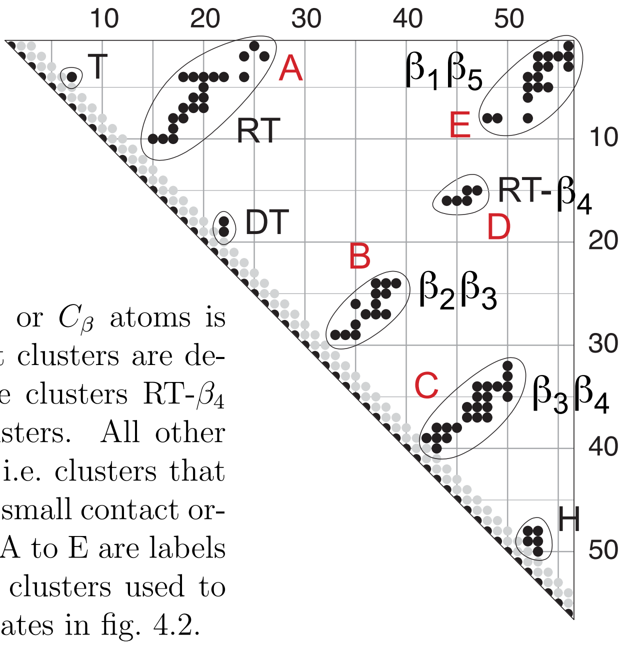

Figure 4.1: Native contact map of the src SH3 domain (protein data bank file 1SRL). Two residues here are defined to be in contact if the distance between their C_α or C_β atoms is less than 6 Å. The contact clusters are defined as in ref. [176]. The clusters RT-β_4 and $\beta_1\beta_5$ are nonlocal clusters. All other clusters are local clusters, i.e. clusters that include contacts (i, j) with small contact order $|i-j|$. The red letters A to E are labels for the five major contact clusters used to describe partially folded states in fig. 4.2.

In the next section, we will show that two-state folding kinetics in the model is obtained if g_l is nonnegative and g_{nl} negative. A nonnegative free energy g_l for local clusters is consistent with the experimental observation that local structures, such as helices or β-hairpins, are generally unstable in isolation. The rate-limiting barrier to folding in our model then turns out to be the formation of mostly local structures needed to reduce the ECOs of nonlocal clusters. The driving force for overcoming this barrier is the favorable, negative free energy g_{nl} of the nonlocal clusters, which stabilize the folded state.

The free-energy landscape of the src SH3 domain is shown in fig. 4.2, for the parameters $g_l = 0$ and $c = 0.5\ k_BT$, and $g_{nl} = -6.6\ k_BT$. Here, k_B is Boltzmann's constant, and T is the temperature. The value of g_{nl} is chosen so that the equilibrium probability that the two nonlocal clusters RT-β_4 and $\beta_1\beta_5$ are both folded ('native state') is 0.9. With these parameters, we obtain good agreement with average experimental Φ-values for secondary elements of the src SH3 domain (see next section). For clarity, we show in the figure only a reduced set of states based on the five major clusters RT, $\beta_2\beta_3$, $\beta_3\beta_4$, RT-β_4, and $\beta_1\beta_5$. The three small clusters T, DT, and H have negligible effects on the folding kinetics and on the Φ-values. Only states differing by the formation of a single cluster are kinetically connected. The uphill steps in this model either are steps in which a local cluster is formed, or steps involving high ECOs. The downhill steps are steps in which a nonlocal cluster is formed with a low ECO, or steps in which a local cluster significantly reduces the ECOs of

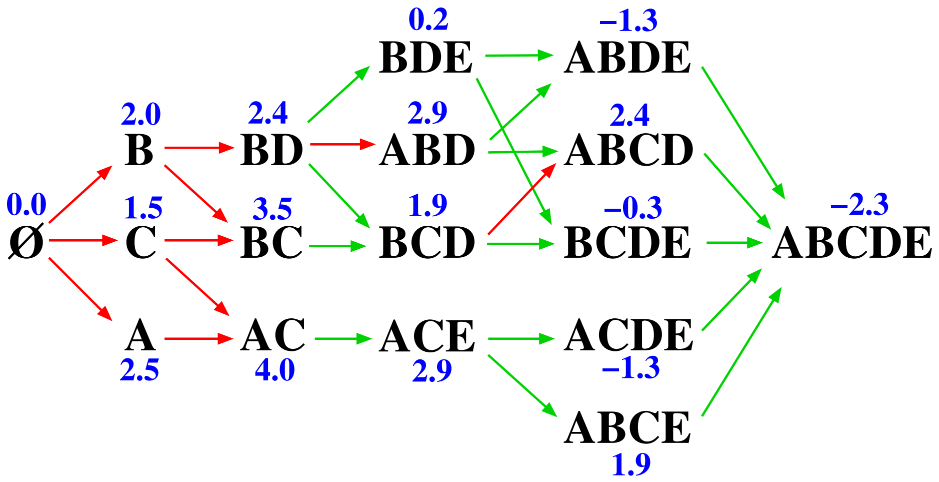

Figure 4.2: Energy landscape for the src SH3 domain as a function of the 5 major clusters (A) RT, (B) $\beta_2\beta_3$, (C) $\beta_3\beta_4$, (D) RT-β_4, and (E) $\beta_1\beta_5$. The contact clusters are defined in the contact map of fig. 4.1. Here, BD, for example, means that only clusters B and D are formed. The free energies given by eq. (4.1) are shown in blue (the units are k_BT). The loop-closure parameter of eq. (4.1) is $c = 0.5\ k_BT$, the free-energy parameter for the local clusters is $g_l = 0$, and the free-energy parameter for the nonlocal clusters RT-β_4 and $\beta_1\beta_5$ is $g_{nl} = -6.6\ k_BT$. Red arrows indicate uphill steps in folding direction, green arrows downhill steps. For clarity, states with free energies larger than 4 k_BT are neglected. The model parameters are given in the text.

previously formed nonlocal clusters.

The model predicts two main folding routes. Along the upper route (E) $\beta_1\beta_5$ folds after (D) RT-β_4; along the lower route, they form in the opposite order. Along these routes, the barriers (highest-free-energy states) are the states in which two clusters are formed: BD and BC for the upper route, and AC for the lower route.

4.2 Cooperativity in two-state protein folding kinetics

In this section, we consider the folding dynamics of the src SH3 domain in the model presented in the previous section. The model parameters throughout this section are the same parameters as in fig. 4.2. The signature of two-state folding is the single-exponential relaxation after an initial fast ‘burst phase’ (see section 1.2). In our model, two-state

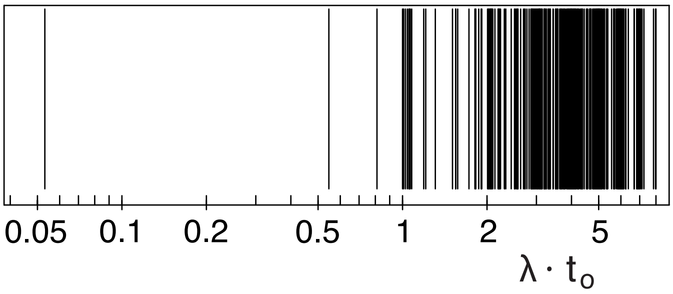

Figure 4.3: Eigenvalue spectrum for the src SH3 domain in units of $1/t_o$ where t_o is the reference time scale for the transition rates (4.5). The model parameters are the same as in fig. 4.2.

folding kinetics is obtained if the eigenvalue spectrum exhibits a gap, which is indeed the case for the src SH3 domain (see fig. 4.3). The slowest relaxation rate λ_1 is about one order of magnitude smaller than the other nonzero eigenvalues. At times $t \gtrsim 1/\lambda_1$, the probability distribution (4.6) is then well approximated by $\boldsymbol{P}(t) \simeq c_0\boldsymbol{Y}_0 + c_1\boldsymbol{Y}_1 \exp[-\lambda_1 t]$, i.e. by a single-exponential relaxation kinetics. Here, $\boldsymbol{Y}_0$ is the eigenvector with eigenvalue 0, which characterizes the equilibrium state, and $\boldsymbol{Y}_1$ is the eigenvector with eigenvalue λ_1. All other relaxation modes constitute a fast initial burst phase.

The time evolution of the folding process is shown in fig. 4.4. There are two time scales, t_o and $t_f \simeq 1/\lambda_1$. Here, t_o is the characteristic timescale of the burst phase in the model and t_f is the single-exponential folding time. At the earliest times $t < t_o$, single local clusters start to form: examples are the clusters A, B, and C of the src SH3 domain, see fig. 4.2. As shown in fig. 4.4, on this time scale, each cluster is only weakly populated, with a probability less than 10%. At intermediate times t with $t_o < t < t_f$, there is a crossover from the burst phase to the single-exponential folding process. During these intermediate times, cluster pairs (AC, BC, BD) begin to form. Fig. 4.2 shows that these pairwise clusters are the barrier events, i.e., they represent the conformational states of maximum free energy obtained during folding. Finally, on the longest time scale, $t \simeq t_f$, the pairwise and triplet clusters reach sufficiently high populations to assemble into multi-cluster complexes, proceeding downhill in free energy to the native structure.

What is the basis for the cooperativity of folding in our model, i.e. for the separation of time scales? First, the formation of local structures, or contact clusters in our model reduces the loop-closure entropies for the formation of the nonlocal structures. Second, only the nonlocal structures have favorable free energies $g_i = g_{nl} < 0$. The formation of the nonlocal

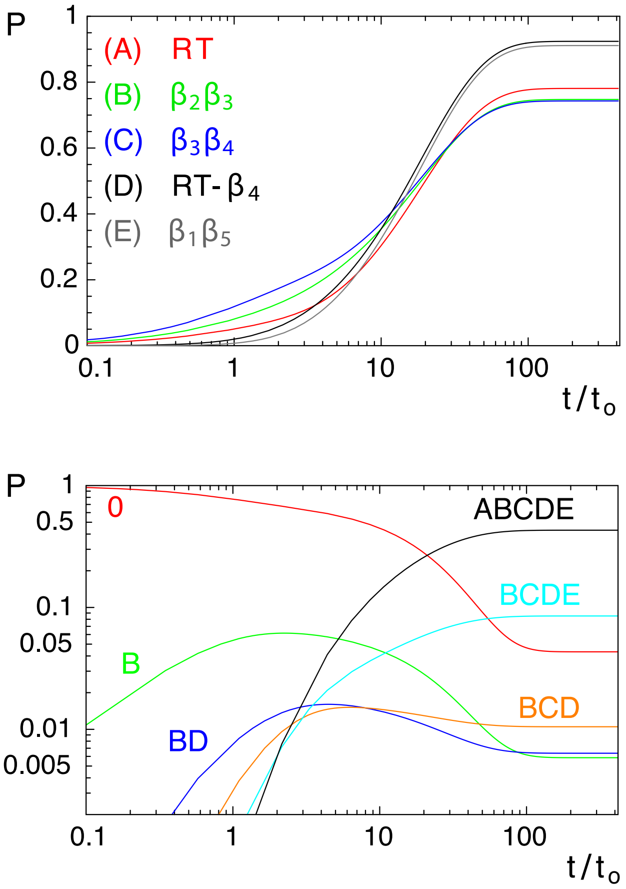

Figure 4.4: (Top) Time evolution of the formation probability P for the major *clusters* of the src SH3 domain during folding (see fig. 4.1). – (Bottom) Time evolution of *state* probabilities for the exemplary path 0 → B → BC → BCD → BCDE → ABCDE of the src SH3 domain (see also fig. 4.2). The initial state at time $t = 0$ is the denatured state in which none of the clusters is formed.

structures stabilizes the overall fold, and thus also the local structures. The barrier arises from the positive free energies in eq. (4.1) due to the formation of local structures and loops (see fig. 4.2). Interestingly, if we set the free energies for local structure formation to be negative by several $k_B T$, we obtain fast multi-exponential downhill folding, without a barrier. Based on experiments and theory, such downhill folding has been recently postulated for the protein BBL [192].

To understand the cooperative folding in the model, it is instructive to turn off the loop-closure term in eq. (4.1) by setting $c = 0$. Then all M clusters are independent of each other, i.e. there is no cooperativity. It can be shown that the matrix $\boldsymbol{W}$ then has the eigenvalues $\lambda = j/t_o$ where j is an integer between 0 and M, the number of clusters. Each of these eigenvalues has a population that is given by the binomial coefficient $j!/[j!(M-j)!]$, which results in a broad non-two-state spectrum of eigenvalues. Hence, the separation of time scales – and the two-state cooperativity – arise in this model from the coupling of the clusters via the loop-closure term in eq. (4.1).

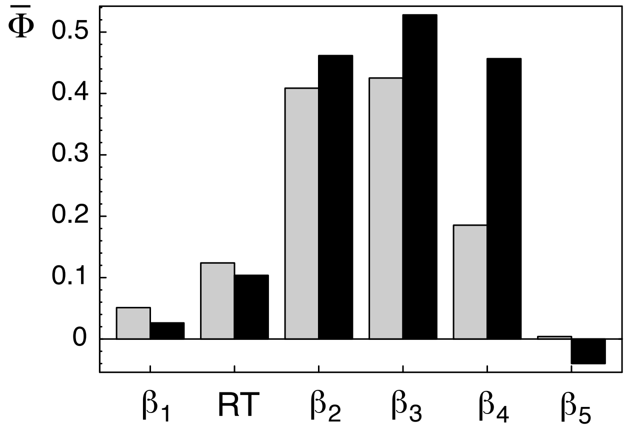

Figure 4.5: (Grey) theoretical and (black) average experimental Φ-values for the secondary structural elements of the src SH3 domain (see also table 3.6). The model parameters are the same as in fig. 4.2.

The folding rate λ_1 is related to the height of the energy barrier on the energy landscape. For comparison, consider a mass-action model with three states D $\leftrightarrow$ T $\leftrightarrow$ N (denatured state, 'transition' state, native state) and transition rates as in eq. (4.5). The folding rate is given, to a very good approximation, by $(1/2)t_o^{-1}\exp[-G_{\text{T-D}}/(k_BT)]$ for barrier energies $G_{\text{T-D}} = G_{\text{T}} - G_{\text{D}} \gg k_BT$. The factor of 1/2 comes from the fact that a molecule in state T can jump both to D and N, with almost equal probability, since both sides of high-barrier transition states are steep downhills. Now, for the energy landscape of the src SH3 domain shown in fig. 4.2, the minimum barrier has free energy 2.4 k_BT for state BD. The corresponding barrier crossing rate of $(1/2)t_o^{-1}\exp[-2.4]$ is in good agreement with the folding rate $\lambda_1 \simeq 0.05/t_o$ (see fig. 4.3).

Experiments have been interpreted either as indicating that burst phases involve structure formation or that burst phases are processes of non-structured polymer collapse, depending on the protein and the experimental method [185, 193–197]. In our model, the burst phase is a process of structure formation. Non-structured collapse is beyond the scope, or resolution, of our model, because the model has only a single fully unstructured state – the state in which none of the clusters is formed. The burst phase in our model captures fast preequilibration events within the denatured state in response to initiating the folding conditions at $t = 0$. In the model, this denatured state is an ensemble of partially folded states on one side of the barrier in the energy landscape (see fig. 4.2). It is reasonable to assume that such preequilibration events within the denatured state exist also for real proteins. However, whether these events can be detected as burst phases in experiments should depend on the initial conditions, experimental probes, etc.

Finally, we show that the model also captures the average Φ-values for the secondary structural elements of the src SH3 domain. To calculate average Φ-values for secondary structures, we consider 'mutations' that change the free energy g_i of a contact cluster according to

$$\Delta g_i(j) = x_{ji}\epsilon \tag{4.7}$$

where x_{ji} is the fraction of residues of the secondary structural element j that are involved in contacts of the cluster i, and ϵ is a small energy. For example, if the secondary structural element j contains m_1 residues, and $m_2 \leq m_1$ of these residues appear in contacts of the cluster i, we have $x_{ji} = m_2/m_1$. Note that $0 \leq x_{ji} \leq 1$, where the value $x_{ji} = 1$ is obtained if the whole secondary structural element j has contacts in cluster i. Thus the Φ-value for the secondary structural element j is given by eq. (1.7) with

$$\ln(k_{\text{wt}}/k_{\text{mut}}) = \ln(\lambda_1/\lambda_1') \tag{4.8}$$

where λ_1' is the smallest nonzero eigenvalue of the mutant with cluster free energies $g_i \rightarrow g_i + \Delta g_i(j)$, and

$$\Delta G_{\text{N-D}} = \sum_i \Delta g_i(j) \tag{4.9}$$

For $\epsilon \ll k_BT$, we find that the calculated Φ-values are nearly independent of ϵ. We choose here $\epsilon = 0.01\ k_BT$. The calculated average Φ-values for the secondary structures of the src SH3 domain are in good agreement with the experimental values, see fig. 4.5.

4.3 Parallel and sequential unfolding events in MD simulations

In this section and the following section 4.4, we analyze unfolding trajectories of the protein CI2 from Molecular Dynamics (MD) simulations with atomistic resolution. In this section, we quantify parallel and sequential processes during unfolding and compare these processes to the minimum-ECO folding route of CI2 shown in fig. 3.5. The minimum-ECO route has been derived from loop-closure dependencies between the structural elements (see section 3.4). In section 4.4, we focus on the correlations of contact unfolding times on the MD trajectories, which reveal a high degree of cooperativity between contacts of the same structural element.

The protein CI2 is a central model system for folding, because of its prominent role as first protein for which two-state kinetics has been observed [14] and an extensive mutational analysis of the kinetics [24, 198].

The folding kinetics of CI2 has been investigated theoretically both with atomistic [48–50, 52, 63, 199–202] and simplified statistical-mechanical models [66–69, 159, 168, 176, 203]. Since atomistic folding simulations are still limited to small or ultrafast folding proteins [64, 80, 204–208], MD unfolding simulations at elevated temperatures are often used to study the kinetics [48–50, 54, 63, 199, 200, 209–216].

In 1968, Levinthal suggested that proteins are guided along sequential pathways into the native structure, since an unguided search of the vast conformational space seemed incompatible with fast and efficient folding [5]. About a decade ago, a 'new view' [10] emerged in which folding is seen as a parallel process on funnel-shaped landscapes, inspired by simple statistical-mechanical models (see also section 1.2). The bias of the funnel landscapes towards the native protein structure ensures efficient folding along a multitude of routes. An intriguing question is whether the apparently contradictory 'old view' of sequential folding and 'new view' of parallel folding can be reconciled [12, 50, 217].

We analyze here parallel and sequential processes on Molecular Dynamics (MD) unfolding trajectories of the protein CI2 at high temperatures. The degree of sequentiality during unfolding will be quantified on two structural levels: on the 'microstructural' level of individual contacts between amino acids, and on the coarser structural level of contact clusters. We consider the sequences of unfolding events of contacts and contact clusters on each trajectory. The unfolding of a pair of contacts or contact cluster is defined as sequential if the same sequence of events is observed on essentially all trajectories. The pairwise unfolding is parallel if contact (or contact cluster) A unfolds prior to contact (or contact cluster) B on some of the trajectories, and later than B on other trajectories [218].

The MD simulations were performed with the CHARMM EFF1 force field [219, 220]. EEF1 is a force field with implicit solvent [221] and has been previously used by several groups to study the unfolding kinetics of proteins [50, 213–215], including the protein CI2 [50]. After minimization of the CI2 crystal structure (protein data bank code: 2ci2), we have performed a 100 ns simulation at the temperature 300 K at which the folded state of the protein is stable [218]. From this simulation trajectory, we took 50 conformations as starting conformations for the thermal unfolding simulations at 400 K, 450 K and 500 K. The length of the individual unfolding simulations depended on the vanishing of the native contacts and was about 100 ns at 400 K, 10 ns at 450 K, and 1 ns at 500K. We have performed 30 unfolding simulations at 400 K, and 50 unfolding simulations at 450 and 500 K.

To define the unfolding times for a specific contact on a trajectory

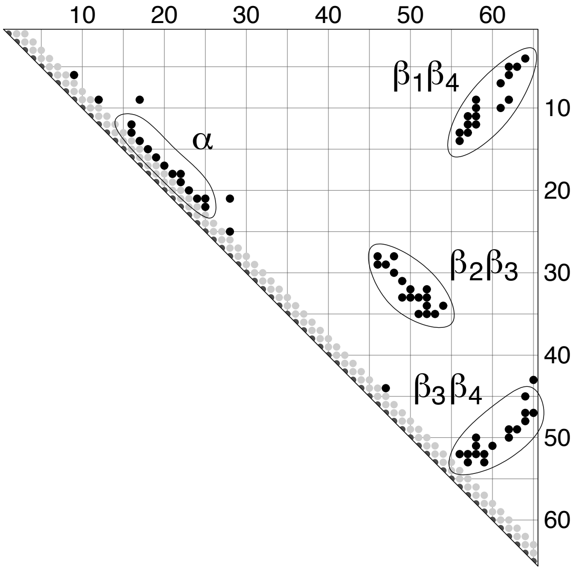

Figure 4.6: Native contacts and contact clusters of CI2. Black dots represent contacts between pairs of amino acids that where present for at least 75% of the simulation time on an exemplary MD-trajectory at 300 K starting from the crystal structure. Two amino acids here are defined to be in contact if the distance between their C_α or C_β atoms is less than 6Å.

at 400 K, we consider time intervals of length 150 ps and determine the probability that the contact is formed during this interval. The unfolding time of this contact is defined as the time at which the probability first falls below the threshold value 0.05. In other words, the unfolding time of a contact is defined as the midpoint of the first 150 ps interval during which the contact was only present 5% of the time. For the trajectories at 450 and 500 K, we use shorter time intervals of length 54 ps and 10.5 ps, respectively, to define contact unfolding times. We consider here as native contacts all contacts that were present during at least 75% of an exemplary trajectory at 300 K (see fig. 4.6). In a given conformation, two residues were taken to be in contact if the distance between their C_α or C_β atoms was less than 6 Å.

Besides contacts, we consider here contact clusters as coarser structural level. The four contact clusters of the protein CI2 correspond to the α-helix and the three strand pairings $\beta_1\beta_4$, $\beta_2\beta_3$, and $\beta_3\beta_4$ (see fig. 4.6). For each of the clusters, we determine the fraction of cluster contacts formed during a trajectory (see fig. 4.7). To define the unfolding sequence of clusters on a trajectory, we consider several threshold values for the fraction of cluster contacts. If two clusters unfold more or less simultaneously, the sequence in which they cross different threshold values can vary. We define a cluster to unfold before another cluster if it crosses all threshold values before that cluster. We have considered here 7 threshold values between 0.05 and 0.2, in intervals of 0.025.

A statistical analysis of the unfolding events on the contact cluster level is presented in fig. 4.8. The numbers indicate the fractions of tra-

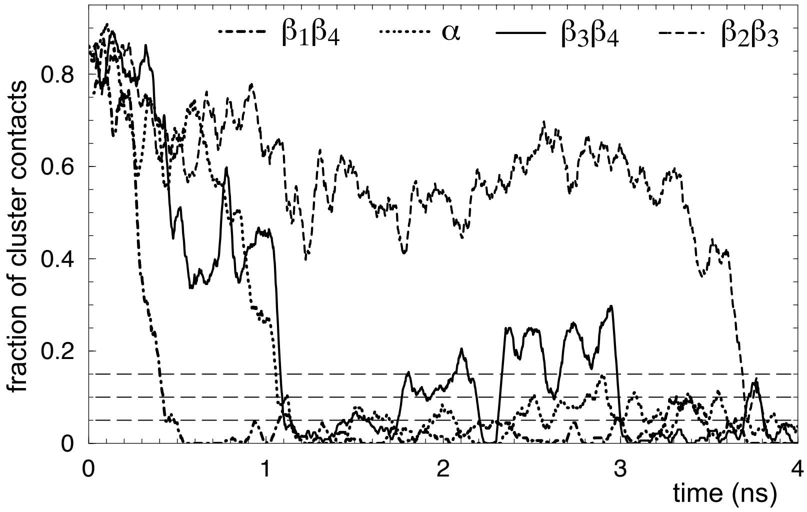

Figure 4.7: Fraction of cluster contacts on an exemplary unfolding trajectory at the temperature 450 K. The four contact clusters α, $\beta_1\beta_4$, $\beta_2\beta_3$, and $\beta_3\beta_4$ of CI2 are defined in fig. 4.6. The contact fractions are averaged over time intervals of 50 ps to integrate out small-time-scale fluctuations. The three dashed horizontal lines represent thresholds used to define the unfolding sequence of the contact cluster. In this example, $\beta_1\beta_4$ is defined to unfold first because its contact fraction crosses all threshold lines prior to the contact fractions of the other clusters. By the same definition, the cluster $\beta_2\beta_3$ here unfolds last. In this example, the clusters α and $\beta_3\beta_4$ unfold 'simultaneously' since α crosses the threshold lines at the contact fraction 0.15 and 0.1 earlier than $\beta_3\beta_4$, but the threshold line at 0.05 later.

jectories on which a given cluster unfolds prior to another cluster. At the temperature 400 K, for example, $\beta_1\beta_4$ unfolds prior to α on 83% of the trajectories, and α unfolds before $\beta_1\beta_4$ on 7% of the trajectories. On the remaining 10% of trajectories, the two clusters unfold simultaneously, i.e. without clear sequence.

At all three temperatures, the cluster $\beta_1\beta_4$ unfolds with high probability prior to the other clusters (the numbers in the second row of the matrices are between 0.83 and 1), and unfolds with low probability after the other clusters (the numbers in the second column are between 0 and 0.07). The unfolding of $\beta_1\beta_4$ with respect to each of the other three clusters thus is sequential. In contrast, the two clusters α and $\beta_3\beta_4$ unfold in parallel. At the temperatures 400 K and 450 K, $\beta_3\beta_4$ unfolds prior to α with probabilities around 0.2, and after α with probabilities slightly

T=400 K unfolds second

unfolds first

	α	$\beta_1\beta_4$	$\beta_2\beta_3$	$\beta_3\beta_4$
α		0.07	0.87	0.43
$\beta_1\beta_4$	0.83		1.00	0.90
$\beta_2\beta_3$	0.00	0.00		0.10
$\beta_3\beta_4$	0.20	0.00	0.77	

T=450 K unfolds second

unfolds first

	α	$\beta_1\beta_4$	$\beta_2\beta_3$	$\beta_3\beta_4$
α		0.02	0.54	0.44
$\beta_1\beta_4$	0.88		0.98	0.90
$\beta_2\beta_3$	0.12	0.02		0.16
$\beta_3\beta_4$	0.22	0.00	0.36	

T=500 K unfolds second

unfolds first

	α	$\beta_1\beta_4$	$\beta_2\beta_3$	$\beta_3\beta_4$
α		0.00	0.22	0.28
$\beta_1\beta_4$	0.94		0.96	0.86
$\beta_2\beta_3$	0.26	0.02		0.14
$\beta_3\beta_4$	0.46	0.04	0.52	

Figure 4.8: Unfolding statistics for the contact clusters of CI2 at the temperatures 400 K, 450 K, and 500 K. The numbers represent the fraction of unfolding trajectories on which contact cluster x unfolds prior to contact cluster y at all considered thresholds (see fig. 4.7). At 400 K, for example, the cluster $\beta_1\beta_4$ opens prior to α on 83% of the trajectories, and α opens prior to $\beta_1\beta_4$ on 7% of the trajectories. On the remaining 10% of trajectories, the two clusters unfold 'simultaneously', i.e. the sequence of unfolding events depends on the considered threshold.

larger than 0.4. On the remaining close to 40% of trajectories, the two clusters unfold simultaneously. At these temperatures, the unfolding of the two clusters is parallel with a 2 to 1 preference for α unfolding prior to $\beta_1\beta_4$ on the trajectories where the two clusters do not unfold simul-

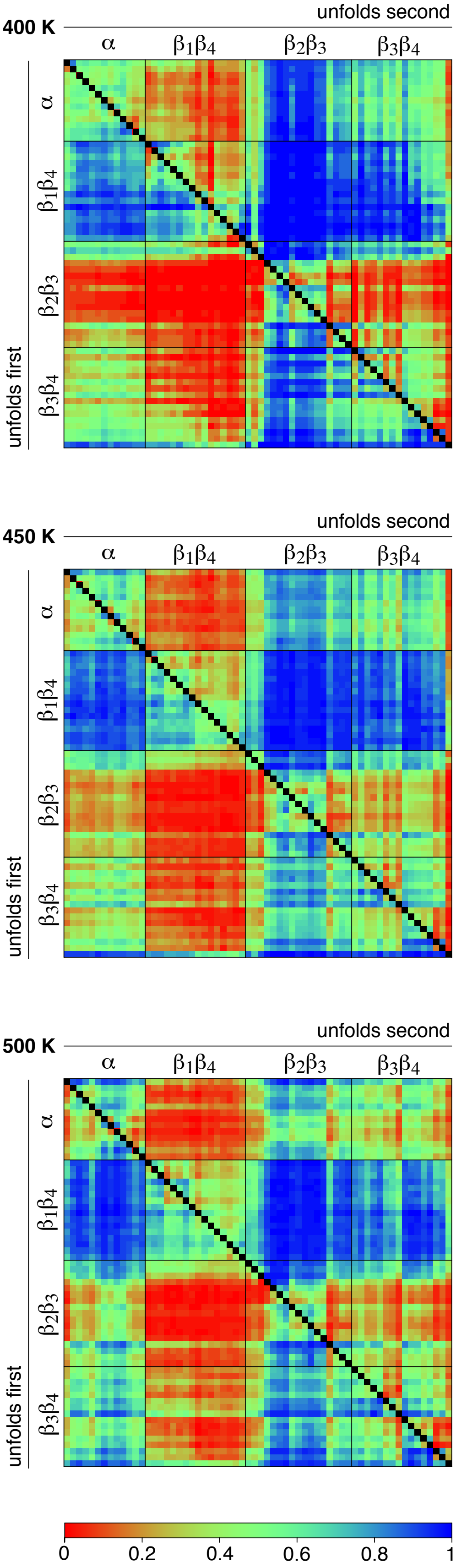

Figure 4.9: Unfolding statistics for native cluster contacts of CI2 at 400 K, 450 K, and 500 K. The colors represent the fraction of unfolding trajectories on which contact x unfolds prior to contact y. For the cluster α, the contacts are arranged in the order 12/16, 13/16, 14/17, 15/18, 16/19, 17/20, 18/21, 18/22, 19/22, 20/23, 21/24, 21/15, 22/25 ('N to C terminus'). For $\beta_1\beta_4$, the contacts are arranged as 14/56, 13/56, 13/57, 12/57, 11/57, 12/58, 11/58, 10/58, 9/58, 10/61, 9/62, 7/61, 6/62, 5/62, 5/63, 4/64 (increasing contact order). For $\beta_2\beta_3$, the order of contacts is 28/46, 28/48, 29/46, 29/47, 30/48, 31/49, 32/50, 32/52, 33/49, 33/50, 33/51, 33/52, 34/52, 34/54, 35/51, 35/52, 35/53. For $\beta_3\beta_4$, the contacts are arranged in the order 52/56, 53/57, 52/57, 52/58, 53/59, 51/58, 52/59, 50/58, 51/60, 50/62, 49/62, 49/63, 48/64, 47/64, 47/65, 45/64 (increasing contact order).

taneously. At the temperature 500 K, the unfolding of α and $\beta_1\beta_4$ is parallel with reversed preferences.

We observe a more pronounced temperature dependence for the unfolding sequences of $\beta_2\beta_3$ and α. The cluster $\beta_2\beta_3$ unfolds sequentially after α at 400 K, and parallel to α at 500 K. Similarly, the unfolding of $\beta_2\beta_3$ and $\beta_3\beta_4$ has a stronger parallel character at the higher temperatures 450 K and 500 K compared to 400 K, with a preference for $\beta_3\beta_4$ opening first.

The overall picture emerging from these statistics is: (1) $\beta_1\beta_4$ unfolds prior to the other three clusters, and (2) the three clusters α, $\beta_2\beta_3$ and $\beta_3\beta_4$ unfold predominantly parallel to each other, with increasing parallelity at higher temperatures. This picture is in agreement with the minimum-ECO route presented for CI2 (see fig. 3.5). On the minimum-ECO route, $\beta_1\beta_4$ is predicted to form after the other three clusters, which fold in parallel. This folding sequence is reversed compared to the MD unfolding sequence.

The unfolding statistics of all pairs of cluster contacts is summarized in fig. 4.9. The precise order in which the contacts of the four clusters are presented is specified in the figure caption. Blue colors indicate high probabilities for unfolding sequencies, and red colors low probabilities. Green colors represent intermediate probabilities, which correspond to parallel events. At all three temperatures, the contacts of $\beta_1\beta_4$ unfold with high probabilities prior to the contacts of the other clusters. At the temperature 400 K, the majority of $\beta_2\beta_3$ contacts have a strong tendency to unfold after the contacts of the clusters α and $\beta_3\beta_4$. This tendency decreases with increasing temperature. The unfolding statistics on the level of individual contacts thus reflects the parallel and sequential events on the cluster level.

A statistical analysis of MD unfolding sequences of the protein CI2 has also been performed by Lazaridis and Karplus [50] and Ferrara et al. [201, 202]. Lazaridis and Karplus [50] have considered the average times for the last appearance of contacts in unfolding simulations of CI2 at the temperature 500 K. They found the smallest average times for contacts between β_1 and β_4, the next-largest average times for contacts between β_3 and β_4 and for contacts within the α-helix, and obtained the largest average times for contacts between β_2 and β_3. Ferrara et al. [201, 202] have considered the average C_α RMSDs of conformations for which groups of contacts disappeared first and appeared last. The C_α RMSD with respect to the native state here served as progress variable for unfolding. Ferrara et al. found the smallest average RMSD values at disappearance, i.e. early unfolding, for the $\beta_1\beta_4$ and $\beta_3\beta_4$ contact groups, followed by RMSD values for the $\beta_2\beta_3$ contact groups, and obtained the largest average RMSD values at disappearance of the contacts of the α-helix. The on average early unfolding of $\beta_1\beta_4$ observed by the two groups is in agreement with our results.

However, our analysis can not be directly compared to sequences of average unfolding times. We identify on each trajectory the unfolding sequences of pairs of contacts or contact clusters, and subsequently estimate probabilities for particular sequences from the numbers of times these sequences appear among all trajectories. The purpose of this analysis is to determine characteristic parallel and sequential unfolding events. Average unfolding times do not reveal this information. For example, a larger average unfolding time for contact A than for contact B is observed if this contact unfolds after contact B on all trajectories (sequential unfolding), but can also be obtained if contact A opens after contact B on some trajectories, and prior to contact B on other trajectories (parallel unfolding).

4.4 Substructural cooperativity

A central assumption of the statistical-mechanical models in the sections 2.1, 2.3, and 4.1 and the minimum-ECO-route model in section 3.4 is that structural elements are either fully formed or not formed in partially folded states of a protein. The structural elements have been identified as contact clusters in native contact maps. In this section, we test this assumption of substructural cooperativity and consider the correlations between the contact unfolding times on the MD trajectories of the protein CI2.

The correlations here are quantified by the Spearman rank correlation coefficient. To calculate the Spearman coefficient, one has to consider the pairs of unfolding times (a_1, b_1), (a_2, b_2), …, (a_N, b_N) of two contacts A and B from all N trajectories. The unfolding times a_i of contact A are then ranked according to their magnitude, and the unfolding times b_i of contact B as well. The Spearman rank correlation coefficient is defined as

$$r = 1 - 6\sum_{i=1}^{N} \frac{d_i^2}{N(N^2-1)} \tag{4.10}$$

where d_i is the rank difference between a_i and b_i. The Spearman rank correlation can attain values between -1 and 1, with 1 representing perfect correlation, and -1 perfect anticorrelation. A value of 1 is obtained if the smallest unfolding time of contact A is paired with the smallest unfolding time of contact B, the next-smallest unfolding time of A with the next-smallest unfolding time of B, etc. The rank difference d_i of all pairs

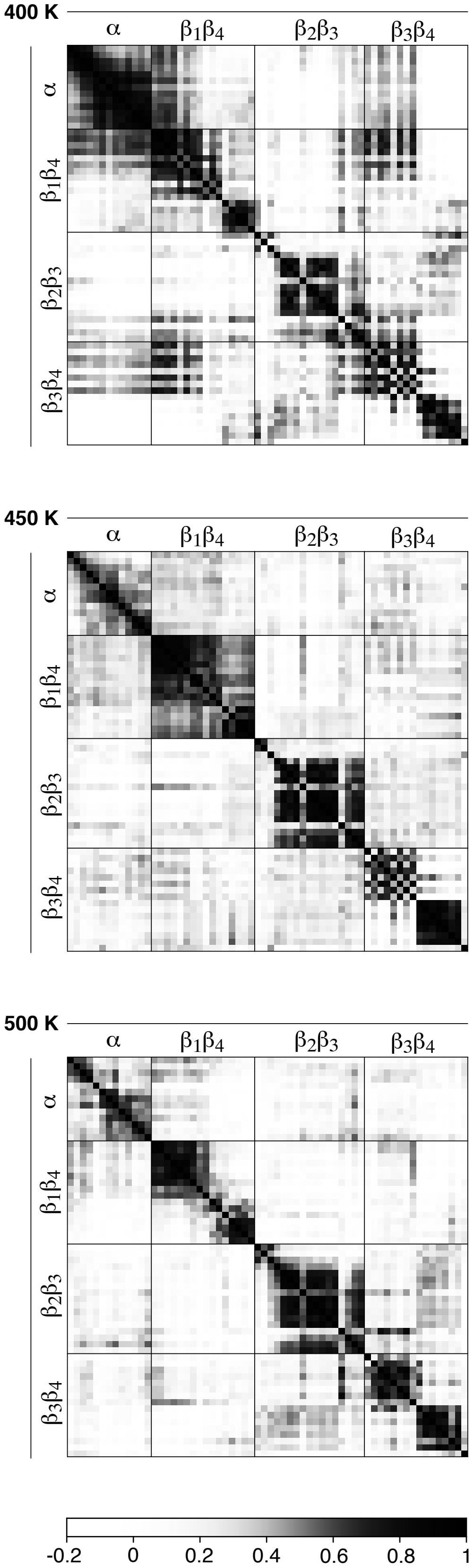

Figure 4.10: Spearman correlation coefficients for the unfolding times of contacts. High correlations between pairs of contacts are represented in black, low correlations in white. High correlations are observed predominantly between contacts of the same contact cluster. The contact clusters thus correspond to cooperative protein substructures. The contacts are presented in the same order as in fig. 4.9. For the 30 trajectories at 400 K, a Spearman correlation coefficient of 0.43 has a p-value of 0.01, and a correlation coefficient 0.55 a p-value of 0.001 [222]. For the 50 trajectories at 450 and 500 K, the correlation coefficients 0.33 and 0.43 have the p-values 0.01 and 0.001, respectively. The p-value of a correlation coefficient is the probability that a similar or higher correlation is obtained by chance. The p-value is a measure for the significance of an observed correlation coefficient. Low p-values indicate high significance.

of unfolding times then is zero. The Spearman correlation coefficient is a simple analogue of the Pearson correlation coefficient. The Spearman correlation is preferable here because it is less sensitive to outliers and, hence, is a more robust measure of correlation.

The obtained correlations of the contact unfolding times are represented in fig. 4.10. The contacts are given in the same order as in fig. 4.9. Here, black indicates high correlations, and white low correlations. We obtain high correlations mostly between contacts of the same contact cluster. The correlations of these contacts are represented in the four sub-matrizes along the diagonal of the matrizes. The high correlations indicate a high degree of substructural cooperativity within contact clusters. For the cluster α, these correlations decrease with increasing temperature. For the clusters $\beta_1\beta_4$ and $\beta_3\beta_4$, the high correlations between contacts of the same cluster mostly appear in two 'sub-blocks'. The contacts of these clusters are ordered according to increasing contact order (see caption of fig. 4.9). The contact order of a contact between residues i and j simply is the sequence separation $|i-j|$. In the contact map shown in fig. 4.6, the cluster $\beta_1\beta_4$ has a small gap between the contacts 14/56 to 9/58 with smaller contact order and the contacts 10/61 to 4/64 with slightly larger contact order. A similar gap also appears in cluster $\beta_3\beta_4$. The two sub-blocks in the correlations between $\beta_1\beta_4$ contacts correspond to high correlations within each of the two groups of contacts. This is also the case for the two sub-blocks in the $\beta_3\beta_4$ correlations. A comparison with fig. 4.9 also reveals a tendency for 'zipping' in $\beta_1\beta_4$ and $\beta_3\beta_4$, i.e. contacts with higher contact order in these clusters have a tendency to unfold earlier than contacts with lower contact order. This can be seen from the dominance of blue colors below the diagonals and red colors above the diagonals of the sub-matrizes in fig. 4.9 that represent the unfolding statistics within the $\beta_1\beta_4$ and $\beta_3\beta_4$ cluster.

In our analysis of MD unfolding trajectories in this and the previous section, we have focused on characteristic unfolding events and correlations, but have not considered transition states for unfolding. The reason is that the unfolding scenario at the high temperatures considered here is not a two-state scenario, but rather resembles a 'downhill-unfolding' scenario. In such a scenario, the initial state of the simulation, the folded state, is instable rather than metastable, see fig. 4.7. The unfolding process is then downhill in free energy and does not involve the crossing of a significant transition-state barrier. Putative transition-state structures have been extracted from high-temperature simulations with a conformational clustering method [49, 63]. At lower temperatures, two-state folding and unfolding has been observed in MD simulations of peptides and small mini-proteins [64, 80, 205–208].

4.5 Summary

In this chapter, we have focused on two different cooperativity aspects of protein folding. First, we have investigated the conditions under which cooperative two-state folding with characteristic single-exponential relaxation (see section 1.2) is obtained in a statistical-mechanical model presented in section 4.1. In this model, partially folded states are described by the structural elements formed in the states. The structural elements are defined via contact clusters in native contact maps (see fig. 4.1). The free energy 4.1 of a state includes loop-closure terms, and terms for forming the cluster contacts. As in section 3.4, the loop-closure dependencies between the contact clusters are captured by ECOs (see fig. 3.1). The folding kinetics of the model is described by a master equation. We find that two-state folding is obtained if the free energies for forming local contact clusters are nonnegative. The formation of these local contact clusters then constitutes a free-energy barrier for folding (see fig. 4.2). The formation of local contact clusters reduces the loop-closure free energy for nonlocal clusters, which form subsequently and stabilize the folded state of the protein. In the model, two-state folding is reflected by a gap in the spectrum of relaxation rates (see fig. 4.3) between the smallest rate λ_1 and all other relaxation rates. On long timescales, the relaxation into equilibrium then is a an effective single-exponential relaxation with rate λ_1 (see section 4.2).

Second, we have analyzed substructural cooperativity by quantifying the correlations between contacts on Molecular Dynamics (MD) unfolding simulations of the protein CI2 (see section 4.4). A correlation analysis of the unfolding times of the contacts reveals high correlations predominantly within contact clusters (see fig. 4.10). The contact clusters thus correspond to cooperative protein substructures. Experimentally, cooperative substructures have been observed during 'cold' unfolding of the protein Ubiquitin [223] and in equilibrium and kinetic hydrogen exchange studies of Cytochrome C [224].

In addition, we have quantified the degree of sequentiality for pairs of contacts and contact clusters on the MD unfolding trajectories of CI2 (see section 4.3). On the level of contact clusters, the characteristic sequential event is the unfolding of $\beta_1\beta_4$ prior to the clusters α, $\beta_2\beta_3$, and $\beta_3\beta_4$ (see fig. 4.8). The unfolding of these other three clusters is predominantly parallel. This unfolding scenario is in agreement with the minimum-ECO folding route of CI2 (see fig. 3.5). On the minimum-ECO route, $\beta_1\beta_4$ forms after the other three clusters, which fold in parallel. This characteristic folding sequence is reversed compared to the MD unfolding sequence. The MD unfolding scenario for the contact clusters is also reflected on the level of individual contacts (see fig. 4.9). On this level, the unfolding process is highly parallel because of the large number of viable unfolding sequences of the 69 contacts.

Publications used in this work

Transition states

- Merlo, C., Dill, K. A., and Weikl, T. R. 2005. Φ-values in protein folding kinetics have energetic and structural components. *Proc. Natl. Acad. Sci. USA* **102**, 10171-10175.
- Weikl, T. R., and Dill, K. A. 2007. Transition states in protein folding kinetics: The structural interpretation of Φ-values. *J. Mol. Biol.* **365**, 1578-1586.
- Weikl, T. R. Transition states in protein folding kinetics: Modeling Φ-values of small β-sheet proteins. *Submitted.*

Loop-closure principles

- Weikl, T. R., and Dill, K. A. 2003. Folding rates and low-entropy-loss routes of two-state proteins. *J. Mol. Biol.* **329**, 585-598.
- Weikl, T. R., and Dill, K. A. 2003. Folding kinetics of two-state proteins: Effect of circularization, permutation, and crosslinks. *J. Mol. Biol.* **332**, 953-963.
- Weikl, T. R. 2005. Loop-closure events during protein folding: Rationalizing the shape of Φ-value distributions. *Proteins* **60**, 701-711.
- Dixit, P. D., and Weikl, T. R. 2006. A simple measure of native-state topology and chain connectivity predicts the folding rates of two-state proteins with and without crosslinks. *Proteins* **64**, 193-197.
- Weikl, T. R. 2007. Loop-closure principles in protein folding. *Arch. Biochem. Biophys.*, in press.

Folding cooperativity

- Weikl, T. R., Palassini, M., and Dill, K. A. 2004. Cooperativity in two-state protein folding kinetics. *Protein Sci.* **13**, 822-829.
- Reich, L. and Weikl, T. R. 2006. Substructural cooperativity and parallel versus sequential events during protein unfolding. *Proteins* **63**, 1052-1058.

Bibliography

[1] Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. 2002. Molecular biology of the cell. Garland, New York.

[2] Branden, C., and J. Tooze. 1999. Introduction to protein structure. Garland, New York.

[3] McPhalen, C., and M. N. James. 1987. Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. *Biochemistry.* 26:261–269.

[4] Dill, K. A., S. B. Ozkan, T. R. Weikl, J. D. Chodera, and V. A. Voelz. 2007. The protein folding problem: when will it be solved? *Curr. Opin. Struct. Biol.* 17:342–346.

[5] Levinthal, C. 1968. Are there pathways for protein folding? *J. Chim. Phys.* 65:44–45.

[6] Levinthal, C. 1969. How to fold graciously. *In* Mössbauer spectroscopy in biological systems. University of Illinois Bulletin. 67:22–24.

[7] Baldwin, R. L. 1999. Protein folding from 1961 to 1982. *Nat. Struct. Biol.* 6:814–817.

[8] Baldwin, R. L., and G. D. Rose. 1999. Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends biochem. sci.* 24:77–83.

[9] Cecconi, C., E. A. Shank, C. Bustamante, and S. Marqusee. 2005. Direct observation of the three-state folding of a single protein molecule. *Science.* 309:2057–2060.

[10] Baldwin, R. L. 1994. Matching speed and stability. *Nature.* 369:183–184.

[11] Matthews, C. R. 1993. Pathways of protein folding. *Annu. Rev. Biochem.* 62:653–683.

[12] Dill, K. A., and H. S. Chan. 1997. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* 4:10–19.

[13] 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.* 21:167–195.

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

[15] Jackson, S. E. 1998. How do small single-domain proteins fold? *Fold. Des.* 3:R81–R91.

[16] Fersht, A. R. 1999. Structure and mechanism in protein science. W. H. Freeman, New York.

[17] Grantcharova, V., E. J. Alm, D. Baker, and A. L. Horwich. 2001. Mechanisms of protein folding. *Curr. Opin. Struct. Biol.* 11:70–82.

[18] Myers, J. K., and T. G. Oas. 2002. Mechanism of fast protein folding. *Annu Rev Biochem.* 71:783–815.

[19] Schuler, B., E. A. Lipman, and W. A. Eaton. 2002. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature.* 419:743–747.

[20] Rhoades, E., M. Cohen, B. Schuler, and G. Haran. 2004. Two-state folding observed in individual protein molecules. *J Am Chem Soc.* 126:14686–14687.

[21] Du, R., V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich. 1998. On the transition coordinate for protein folding. *J. Chem. Phys.* 108:334–350.

[22] Hummer, G. 2004. From transition paths to transition states and rate coefficients. *J. Chem. Phys.* 120:516–523.

[23] Snow, C. D., Y. M. Rhee, and V. S. Pande. 2006. Kinetic definition of protein folding transition state ensembles and reaction coordinates. *Biophys J.* 91:14–24.

[24] Itzhaki, L. S., D. E. Otzen, and A. R. Fersht. 1995. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* 254:260–288.

[25] Villegas, V., J. C. Martinez, F. X. Aviles, and L. Serrano. 1998. Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* 283:1027–1036.

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

[27] Ternström, T., U. Mayor, M. Akke, and M. Oliveberg. 1999. From snapshot to movie: Φ analysis of protein folding transition states taken one step further. *Proc. Natl. Acad. Sci. USA.* 96:14854–14859.

[28] Kragelund, B. B., P. Osmark, T. B. Neergaard, J. Schiodt, K. Kristiansen, J. Knudsen, and F. M. Poulsen. 1999. The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. *Nat. Struct. Biol.* 9:594–601.

[29] Martinez, J. C., and L. Serrano. 1999. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat. Struct. Biol.* 6:1010–1016.

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

[31] Fulton, K. F., E. R. G. Main, V. Daggett, and S. E. Jackson. 1999. Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol.* 291:445–461.

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

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

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

[35] Jäger, M., H. Nguyen, J. C. Crane, J. W. Kelly, and M. Gruebele. 2001. The folding mechanism of a β-sheet: The WW domain. *J. Mol. Biol.* 311:373–393.

[36] Otzen, D. E., and M. Oliveberg. 2002. Conformational plasticity in folding of the split β-α-β protein S6: Evidence for burst-phase disruption of the native state. *J. Mol. Biol.* 317:613–627.

[37] Northey, J. G. B., A. A. Di Nardo, and A. R. Davidson. 2002. Hydrophobic core packing in the SH3 domain folding transition state. *Nat. Struct. Biol.* 9:126–130.

[38] Gianni, S., N. R. Guydosh, F. Khan, T. D. Caldas, U. Mayor, G. W. N. White, M. L. DeMarco, V. Daggett, and A. R. Fersht. 2004. Unifying features in protein-folding mechanisms. *Proc. Natl. Acad. Sci. USA.* 100:13286–13291.

[39] Deechongkit, S., H. Nguyen, E. T. Powers, P. E. Dawson, M. Gruebele, and J. W. Kelly. 2004. Context-dependent contributions of backbone hydrogen bonding to β-sheet folding energetics. *Nature.* 430:101–105.

[40] Garcia-Mira, M. M., D. Böhringer, and F. X. Schmid. 2004. The folding transition state of the cold shock protein is strongly polarized. *J. Mol. Biol.* 339:555–569.

[41] Anil, B., S. Sato, J. H. Cho, and D. P. Raleigh. 2005. Fine structure analysis of a protein folding transition state; distinguishing between hydrophobic stabilization and specific packing. *J. Mol. Biol.* 354:693–705.

[42] Wilson, C. J., and P. Wittung-Stafshede. 2005. Snapshots of a dynamic folding nucleus in zinc-substituted Pseudomonas aeruginosa azurin. *Biochemistry.* 44:10054–10062.

[43] Petrovich, M., A. L. Jonsson, N. Ferguson, V. Daggett, and A. R. Fersht. 2006. Φ-analysis at the experimental limits: Mechanism of β-hairpin formation. *J. Mol. Biol.* 360:865–881.

[44] Matouschek, A., J. T. Kellis, L. Serrano, and A. R. Fersht. 1989. Mapping the transition state and pathway of protein folding by protein engineering. *Nature.* 340:122–126.

[45] Fersht, A. R., and S. Sato. 2004. Φ-value analysis and the nature of protein folding transition states. *Proc. Natl. Acad. Sci. USA.* 101:7976–7981.

[46] Goldenberg, D. P. 1999. Finding the right fold. *Nat. Struct. Biol.* 6:987–990.

[47] de los Rios, M. A., M. Daneshi, and K. W. Plaxco. 2005. Experimental investigation of the frequency and substitution dependence of negative Φ-values in two-state proteins. *Biochemistry.* 44:12160–12167.

[48] Li, A., and V. Daggett. 1994. Characterization of the transition state of protein unfolding by use of molecular dynamics: Chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. USA.* 91:10430–10434.

[49] Li, A., and V. Daggett. 1996. Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by Molecular Dynamics simulations. *J. Mol. Biol.* 257:412–429.

[50] Lazaridis, T., and M. Karplus. 1997. "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science.* 278:1928–1931.

[51] Vendruscolo, M., E. Paci, C. M. Dobson, and M. Karplus. 2001. Three key residues form a critical contact network in a protein folding transition state. *Nature.* 409:641–645.

[52] Li, L., and E. I. Shakhnovich. 2001. Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations. *Proc. Natl. Acad. Sci. USA.* 98:13014–13018.

[53] Gsponer, J., and A. Caflisch. 2002. Molecular dynamics simulations of protein folding from the transition state. *Proc. Natl. Acad. Sci. USA.* 99:6719–6724.

[54] Paci, E., M. Vendruscolo, C. M. Dobson, and M. Karplus. 2002. Determination of a transition state at atomic resolution from protein engineering data. *J. Mol. Biol.* 324:151–163.

[55] Guo, W., S. Lampoudi, and J.-E. Shea. 2003. Posttransition state desolvation of the hydrophobic core of the src-SH3 protein domain. *Biophys. J.* 85:61–69.

[56] Settanni, G., J. Gsponer, and A. Caflisch. 2004. Formation of the folding nucleus of an SH3 domain investigated by loosely coupled molecular dynamics simulations. *Biophys. J.* 86:1691–1701.

[57] Paci, E., K. Lindorff-Larsen, C. M. Dobson, M. Karplus, and M. Vendruscolo. 2005. Transition state contact orders correlate with protein folding rates. *J. Mol. Biol.* 352:495–500.

[58] Salvatella, X., C. M. Dobson, A. R. Fersht, and M. Vendruscolo. 2005. Determination of the folding transition states of barnase by using Φ_I-value-restrained simulations validated by double mutant Φ_{IJ}-values. *Proc. Natl. Acad. Sci. USA*. 102:12389–12394.

[59] Chong, L. T., C. D. Snow, Y. M. Rhee, and V. S. Pande. 2005. Dimerization of the p53 oligomerization domain: Identification of a folding nucleus by molecular dynamics simulations. *J. Mol. Biol.* 345:869–878.

[60] Hubner, I. A., K. A. Edmonds, and E. I. Shakhnovich. 2005. Nucleation and the transition state of the sh3 domain. *J. Mol. Biol.* 349:424–434.

[61] Duan, J., and L. Nilsson. 2005. Thermal unfolding simulations of a multimeric protein − Transition state and unfolding pathways. *Proteins*. 59:170–182.

[62] Daggett, V., A. Li, L. S. Itzhaki, D. E. Otzen, and A. R. Fersht. 1996. Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol.* 257:430–440.

[63] Day, R., and V. Daggett. 2005. Sensitivity of the folding/unfolding transition state ensemble of chymotrypsin inhibitor 2 to changes in temperature and solvent. *Protein Sci.* 14:1242–1252.

[64] Settanni, G., F. Rao, and A. Caflisch. 2005. Φ-value analysis by molecular dynamics simulations of reversible folding. *Proc. Natl. Acad. Sci. USA*. 102:628–633.

[65] Lindorff-Larsen, K., E. Paci, L. Serrano, C. M. Dobson, and M. Vendruscolo. 2003. Calculation of mutational free energy changes in transition states for protein folding. *Biophys. J.* 85:1207–1214.

[66] Alm, E., and D. Baker. 1999. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA*. 96:11305–11310.

[67] Muñoz, V., and W. A. Eaton. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA*. 96:11311–11316.

[68] Galzitskaya, O. V., and A. V. Finkelstein. 1999. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. USA*. 96:11299–11304.

[69] Clementi, C., H. Nymeyer, and J. N. 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. *J. Mol. Biol.* 298:937–953.

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

[71] Alm, E., A. V. Morozov, T. Kortemme, and D. Baker. 2002. Simple physical models connect theory and experiment in protein folding kinetics. *J. Mol. Biol.* 322:463–476.

[72] Kameda, T. 2000. Importance of sequence specificity for predicting protein folding pathways: Perturbed Gaussian chain model. *Proteins*. 53:616–628.

[73] Plaxco, K. W., K. T. Simons, and D. Baker. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* 277:985–994.

[74] Socolich, M., S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan. 2005. Evolutionary information for specifying a protein fold. *Nature*. 437:512–518.

[75] Fernandez-Escamilla, A. M., S. Ventura, L. Serrano, and M. A. Jimenez. 2006. Design and NMR conformational study of a β-sheet peptide based on betanova and WW domains. *Protein Sci.* 15:2278–2289.

[76] Ferguson, N., J. R. Pires, F. Toepert, C. M. Johnson, Y. P. Pan, R. Volkmer-Engert, J. Schneider-Mergener, V. Daggett, H. Oschkinat, and A. Fersht. 2002. Using flexible loop mimetics to extend Φ-value analysis to secondary structure interactions. *Proc. Natl. Acad. Sci. USA*. 98:13008–13013.

[77] Nguyen, H., M. Jäger, A. Moretto, M. Gruebele, and J. W. Kelly. 2003. Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation. *Proc. Natl. Acad. Sci. USA.* 100:3948–3953.

[78] Jäger, M., Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M. E. Bowman, J. P. Noel, M. Gruebele, and J. W. Kelly. 2006. Structure-function-folding relationship in a WW domain. *Proc. Natl. Acad. Sci. USA.* 103:10648–10653.

[79] Bursulaya, B. D., and C. L. Brooks III. 1999. Folding free energy surface of a three-stranded β-sheet protein. *J. Am. Chem. Soc.* 121:9947–9951.

[80] Ferrara, P., and A. Caflisch. 2000. Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc. Natl. Acad. Sci. USA.* 97:10780–10785.

[81] Davis, R., C. M. Dobson, and M. Vendruscolo. 2002. Determination of the structures of distinct transition state ensembles for a β-sheet peptide with parallel folding pathways. *J. Chem. Phys.* 117:9510–9517.

[82] Karanicolas, J., and C. L. Brooks III. 2004. Integrating folding kinetics and protein function: Biphasic kinetics and dual binding specificity in a WW domain. *Proc. Natl. Acad. Sci. USA.* 101:3432–3437.

[83] Bruscolini, P., and F. Cecconi. 2005. Analysis of PIN1 WW domain through a simple statistical mechanics model. *Biophys. Chem.* 115:153–158.

[84] Macias, M. J., V. Gervais, C. Civera, and H. Oschkinat. 2000. Structural analysis of WW domains and design of a WW prototype. *Nat. Struct. Biol.* 7:375–379.

[85] Ranganathan, R., K. P. Lu, T. Hunter, and J. P. Noel. 1997. Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent. *Cell.* 89:875–886.

[86] Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD - Visual Molecular Dynamics. *J. Mol. Graphics.* 14.

[87] Merritt, E. A., and D. J. Bacon. 1997. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* 277:505–524.

[88] van Kampen, N. G. 1992. Stochastic processes in physics and chemistry. North-Holland, Amsterdam.

[89] Weikl, T. R. 2007. Transition states in protein folding kinetics: Modeling Φ-values of small β-sheet proteins. Submitted.

[90] Guerois, R., J. E. Nielsen, and L. Serrano. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* 320:369–387.

[91] Schymkowitz, J., J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano. 2005. The FoldX web server: an online force field. *Nucleic Acids Res.* 33:W382–W388.

[92] Vriend, G. 1990. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* 8:52–56.

[93] de los Rios, M. A., B. K. Muralidhara, D. Wildes, T. R. Sosnick, S. Marqusee, P. Wittung-Stafshede, K. W. Plaxco, and I. Ruczinski. 2006. On the precision of experimentally determined protein folding rates and Φ-values. *Protein Sci.* 15:553–563.

[94] Ruczinski, I., T. R. Sosnick, and K. W. Plaxco. 2006. Methods for the accurate estimation of confidence intervals on protein folding Φ-values. *Protein Sci.* 15:2257–2264.

[95] Akmal, A., and V. Munoz. 2004. The nature of the free energy barriers to two-state folding. *Proteins.* 57:142–152.

[96] Weikl, T. R., and K. A. Dill. 2007. Transition states in protein folding kinetics: The structural interpretation of Φ-values. *J. Mol. Biol.* 365:1578–1586.

[97] Hedberg, L., and M. Oliveberg. 2004. Scattered Hammond plots reveal second level of site-specific information in protein folding: $\Phi'(\beta^{\ddagger})$. *Proc. Natl. Acad. Sci. USA.* 101:7606–7611.

[98] Went, H. M., and S. E. Jackson. 2005. Ubiquitin folds through a highly polarized transition state. *Protein Eng.* 18:229–237.

[99] Sato, S., T. L. Religa, V. Daggett, and A. R. Fersht. 2004. Testing protein-folding simulations by experiment: B domain of protein A. *Proc. Natl. Acad. Sci. USA.* 101:6952–6956.

[100] Teilum, K., T. Thormann, N. R. Caterer, H. I. Poulsen, P. H. Jensen, J. Knudsen, B. B. Kragelund, and F. M. Poulsen. 2005. Different secondary structure elements as scaffolds for protein folding transition states of two homologous four-helix bundles. *Proteins.* 59:80–90.

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

[102] Cota, E., A. Steward, S. B. Fowler, and C. J. 2001. The folding nucleus of a fibronection type III domain is composed of core residues of the immunoglobolin fold. *J. Mol. Biol.* 305:1185–1194.

[103] Serrano, L., A. Matouschek, and A. R. Fersht. 1992. The folding of an enzyme. III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure. *J. Mol. Biol.* 224:805–818.

[104] Jemth, P., R. Day, S. Gianni, F. Khan, M. Allen, V. Daggett, and A. R. Fersht. 2005. The structure of the major transition state for folding of an FF domain from experiment and simulation. *J. Mol. Biol.* 350:363–378.

[105] Zhou, Z., Y. Huang, and Y. Bai. 2005. An on-pathway hidden intermediate and the early rate-limiting transition state of Rd-apocytochrome b562 characterized by protein engineering. *J. Mol. Biol.* 352:757–764.

[106] Scott, K. A., L. G. Randles, and J. Clarke. 2004. The folding of spectrin domains II: Φ-value analysis of R16. *J. Mol. Biol.* 344:207–221.

[107] Muñoz, V., and L. Serrano. 1995a. Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. *J. Mol. Biol.* 245:275–296.

[108] Muñoz, V., and L. Serrano. 1995b. Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence. *J. Mol. Biol.* 245:297–308.

[109] Lacroix, E., A. R. Viguera, and L. Serrano. 1998. Elucidating the folding problem of α-helices: Local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J. Mol. Biol.* 284:173–191.

[110] Merlo, C., K. A. Dill, and T. R. Weikl. 2005. Φ-values in protein folding kinetics have energetic and structural components. *Proc. Natl. Acad. Sci. USA.* 102:10171–10175.

[111] Pace, C. N., and J. M. Scholtz. 1998. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* 75:422–427.

[112] Rao, F., G. Settanni, E. Guarnera, and A. Caflisch. 2005. Estimation of protein folding probability from equilibrium simulations. *J. Chem. Phys.* 122:184901.

[113] Maxwell, K. L., D. Wildes, A. Zarrine-Afsar, M. A. de los Rios, A. G. Brown, C. T. Friel, L. Hedberg, J. C. Horng, D. Bona, E. J. Miller, A. Vallee-Belisle, E. R. Main, F. Bemporad, L. Qiu, K. Teilum, N. D. Vu, A. Edwards, I. Ruczinski, F. M. Poulsen, B. B. Kragelund, S. W. Michnick, F. Chiti, Y. Bai, S. J. Hagen, L. Serrano, M. Oliveberg, D. P. Raleigh, P. Wittung-Stafshede, S. E. Radford, S. E. Jackson, T. R. Sosnick, S. Marqusee, A. R. Davidson, and K. W. Plaxco. 2005. Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci.* 14:602–616.

[114] Kamagata, K., M. Arai, and K. Kuwajima. 2004. Unification of the folding mechanisms of non-two-state and two-state proteins. *J. Mol. Biol.* 339:951–965.

[115] Yang, W. Y., and M. Gruebele. 2003. Folding at the speed limit. *Nature.* 423:193–197.

[116] Kubelka, J., J. Hofrichter, and W. A. Eaton. 2004. The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.* 14:76–88.

[117] Xu, Y., P. Purkayastha, and F. Gai. 2006. Nanosecond folding dynamics of a three-stranded β-sheet. *J. Am. Chem. Soc.* 128:15836–15842.

[118] Kubelka, J., T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter. 2006. Sub-microsecond protein folding. *J. Mol. Biol.* 359:546–553.

[119] Baker, D. 2000. A surprising simplicity to protein folding. *Nature.* 405:39–42.

[120] Bradley, P., K. M. Misura, and D. Baker. 2005. Toward high-resolution de novo structure prediction for small proteins. *Science.* 309:1868–1871.

[121] Weikl, T. R. 2007. Loop-closure principles in protein folding. *Arch. Biochem. Biophys.* In press.

[122] 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. *Proc. Natl. Acad. Sci. USA.* 97:1525–1529.

[123] Zhou, H.-X. 2004. Loops, linkages, rings, catenanes, cages, and crowders: Entropy-based strategies for stabilizing proteins. *Acc. Chem. Res.* 37:123–130.

[124] Ladurner, A. G., and A. R. Fersht. 1997. Glutamine, alanine or glycine repeats inserted into the loop of a protein have minimal effects on stability and folding rates. *J. Mol. Biol.* 273:330–337.

[125] Viguera, A.-R., and L. Serrano. 1997. Loop length, intromolecular diffusion, and protein folding. *Nat. Struct. Biol.* 4:939–946.

[126] Grantcharova, V. P., D. S. Riddle, and D. Baker. 2000. Long-range order in the src SH3 folding transition state. *Proc. Natl. Acad. Sci. USA.* 97:7084–7089.

[127] Otzen, D. E., and A. R. Fersht. 1998. Folding of circular and permuted chymotrypsin inhibitor 2: Retention of the folding nucleus. *Biochemistry.* 37:8139–8146.

[128] Schönbrunner, N., G. Pappenberger, M. Scharf, J. Engels, and T. Kiefhaber. 1997. Effect of preformed correct tertiary interactions on rapid two-state tendamistat folding: Evidence for hairpins as initiation sites for β-sheet formation. *Biochemistry.* 36:9057–9065.

[129] Camarero, J. A., D. Fushman, S. Sato, I. Giriat, D. Cowburn, D. P. Raleigh, and T. W. Muir. 2001. Rescuing a destabilized protein fold through backbone cyclization. *J. Mol. Biol.* 308:1045–1062.

[130] Ainavarapu, R. K., J. Brujic, H. H. Huang, A. P. Wiita, H. Lu, L. W. Li, K. A. Walther, M. Carrion-Vazquez, H. B. Li, and J. M. Fernandez. 2007. Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. *Biophys. J.* 92:225–233.

[131] Burton, R. E., G. Huang, M. A. Daugherty, P. W. Fullbright, and T. G. Oas. 1996. Microsecond protein folding through a compact transition state. *J. Mol. Biol.* 263:311–322.

[132] Gunasekaran, K., S. J. Eyles, A. T. Hagler, and L. M. Gierasch. 2001. Keeping it in the family: folding studies of related proteins. *Curr. Opin. Struct. Biol.* 11:83–93.

[133] Zarrine-Afsar, A., S. M. Larson, and A. R. Davidson. 2005. The family feud: do proteins with similar structures fold via the same pathway? *Curr. Opion. Struct. Biol.* 15:42–49.

[134] Dill, K., K. M. Fiebig, and H. S. Chan. 193. Cooperativity in protein-folding kinetics. *Proc. Natl. Acad. Sci. USA*. 90:1942–1946.

[135] Fiebig, K. M., and K. A. Dill. 1993. Protein core assembly processes. *J. Chem. Phys.* 98:3475–3487.

[136] Kuznetsov, I. B., and S. Rackovsky. 2004. Class-specific correlations between protein folding rate, structure-derived, and sequence-derived descriptors. *Proteins*. 54:333–341.

[137] Gromiha, M. M., and S. Selvaraj. 2001. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction. *J. Mol. Biol.* 310:27–32.

[138] Zhou, H., and Y. Zhou. 2002. Folding rate prediction using total contact distance. *Biophys. J.* 82:458–463.

[139] Makarov, D. E., C. A. Keller, K. W. Plaxco, and H. Metiu. 2002. How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. *Proc. Natl. Acad. Sci. USA*. 99:3535–3539.

[140] Makarov, D. E., and K. W. Plaxco. 2003. The topomer search model: A simple, quantitative theory of two-state protein folding kinetics. *Protein Sci.* 12:17–26.

[141] Debe, D. A., M. J. Carlson, and W. A. Goddard. 1999. The topomer-sampling model of protein folding. *Proc. Natl Acad. Sci. USA*. 96:2596–2601.

[142] Debe, D. A., and W. A. Goddard. 1999. First principles prediction of protein folding rates. *J. Mol. Biol.* 294:619–625.

[143] Gillespie, B., and K. W. Plaxco. 2004. Using protein folding rates to test protein folding theories. *Annu. Rev. Biochem.* 73:837–859.

[144] Wallin, S., and H. S. Chan. 2005. A critical assessment of the topomer search model of protein folding using a continuum explicit-chain model with extensive conformational sampling. *Protein Sci.* 14:1643–1660.

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

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

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

[148] Dixit, P. D., and T. R. Weikl. 2006. A simple measure of native-state topology and chain connectivity predicts the folding rates of two-state proteins with and without crosslinks. *Proteins.* 64:193–197.

[149] Jacobson, H., and W. H. Stockmayer. 1950. Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.* 18:1600–1606.

[150] Chan, H. S., and K. A. Dill. 1990. The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.* 92:3118–3135.

[151] Camacho, C. J., and D. Thirumalai. 1995. Theoretical predictions of folding pathways by using the proximity rule, with applications to bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA.* 92:1277–1281.

[152] Zhou, H.-X. 2003. Effect of backbone cyclization on protein folding stability: Chain entropies of both the unfolded and the folded states are restricted. *J. Mol. Biol.* 332:257–264.

[153] Weikl, T. R., and K. A. Dill. 2003a. Folding rates and low-entropy-loss routes of two-state proteins. *J. Mol. Biol.* 329:585–598.

[154] Weikl, T. R., and K. A. Dill. 2003b. Folding kinetics of two-state proteins: Effect of circularization, permutation, and crosslinks. *J. Mol. Biol.* 332:953–963.

[155] Weikl, T. R. 2005. Loop-closure events during protein folding: Rationalizing the shape of Φ-value distributions. *Proteins.* 60:701–711.

[156] Shea, J. E., J. N. Onuchic, and C. L. Brooks III. 1999. Exploring the origins of topological frustration: design of a minimally frustrated model of fragment B of protein A. *Proc. Natl. Acad. Sci. USA,*. 96:12512–12517.

[157] Clementi, C., P. A. Jennings, and J. N. Onuchic. 2001. Prediction of folding mechanism for circular-permuted proteins. *J. Mol. Biol.* 311:879–890.

[158] Clementi, C., A. E. Garcia, and J. N. Onuchic. 2003. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: All-atom representation study of protein L. *J. Mol. Biol.* 326:933–954.

[159] Hoang, T. X., and M. Cieplak. 2000. Sequencing of folding events in Go-type proteins. *J. Chem. Phys.* 113:8319–8328.

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

[161] Karanicolas, J., and C. L. Brooks III. 2002. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* 11:2351–2361.

[162] Ejtehadi, M. R., S. P. Avall, and S. S. Plotkin. 2004. Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proc. Natl. Acad. Sci. USA.* 101:15088–15093.

[163] Brown, S., and T. Head-Gordon. 2004. Intermediates and the folding of proteins L and G. *Protein Sci.* 13:958–970.

[164] Wallin, S., and H. S. Chan. 2006. Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple native-centric polymer model. *J. Phys.* 18:S307–S328.

[165] Qin, M., J. Zhang, and W. Wang. 2006. Effects of disulfide bonds on folding behavior and mechanism of the β-sheet protein tendamistat. *Biophys. J.* 90:272–286.

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

[167] Garbuzynskiy, S. O., A. V. Finkelstein, and O. V. Galzitskaya. 2004. Outlining folding nuclei in globular proteins. *J. Mol. Biol.* 336:509–525.

[168] Bruscolini, P., and A. Pelizzola. 2002. Exact solution of the Muñoz-Eaton model for protein folding. *Phys. Rev. Lett.* 88:258101.

[169] Karanicolas, J., and C. L. Brooks III. 2003. The importance of explicit chain representation in protein folding models: An examination of Ising-like models. *Proteins.* 53:740–747.

[170] Henry, E. R., and W. A. Eaton. 2004. Combinatorial modeling of protein folding kinetics: Free energy profiles and rates. *Chem. Phys.* 307:163–185.

[171] Nelson, E. D., and N. V. Grishin. 2006. Alternate pathways for folding in the flavodoxin fold family revealed by a nucleation-growth model. *J. Mol. Biol.* 358:646–653.

[172] Zwanzig, R. 1995. Simple model of protein folding kinetics. *Proc. Natl. Acad. Sci.* 92:9801–9804.

[173] Hilser, V. J., and E. Freire. 1996. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. *J. Mol. Biol.* 262:756–772.

[174] Hilser, V. J., B. Garcia-Moreno, T. G. Oas, G. Kapp, and S. T. Whitten. 2006. A statistical thermodynamic model of the protein ensemble. *Chem. Rev.* 106:1545–1558.

[175] Zimm, B. H., and J. K. Bragg. 1959. Theory of the phase transition between helix and random coil. *J. Chem. Phys.* 31:526–535.

[176] Weikl, T. R., M. Palassini, and K. A. Dill. 2004. Cooperativity in two-state protein folding kinetics. *Protein Sci.* 13:822–829.

[177] Shmygelska, A. 2005. Search for folding nuclei in native protein structures. *Bioinformatics.* 21:I394–I402.

[178] Karplus, M., and D. L. Weaver. 1976. Protein-folding dynamics. *Nature.* 260:404–406.

[179] Karplus, M., and D. Weaver. 1994. Protein-folding dynamics: The diffusion-collision model and experimental data. *Protein Sci.* 3:650–668.

[180] Islam, S. A., M. Karplus, and D. L. Weaver. 2002. Application of the diffusion-collision model to the folding of three-helix bundle proteins. *J. Mol. Biol.* 318:199–215.

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

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

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

[184] Teilum, K., F. M. Poulsen, and M. Akke. 2006. The inverted chevron plot measured by NMR relaxation reveals a native-like unfolding intermediate in acyl-CoA binding protein. *Proc. Natl. Acad. Sci. USA.* 103:6877–6882.

[185] Englander, S. W. 2000. Protein folding intermediates and pathways studied by hydrogen exchange. *Annu. Rev. Biophys. Biomol. Struct.* 29:213–238.

[186] Bai, Y., T. R. Sosnick, L. Mayne, and S. W. Englander. 1995. Protein folding intermediates: Native-state hydrogen exchange. *Science.* 269:192–197.

[187] Juneja, J., and J. B. Udgaonkar. 2003. NMR studies of protein folding. *Current Science.* 84:157–172.

[188] Krishna, M. M., L. Hoang, Y. Lin, and S. W. Englander. 2004. Hydrogen exchange methods to study protein folding. *Methods.* 34:51–64.

[189] Bai, Y. W. 2006. Protein folding pathways studied by pulsed-and native-state hydrogen exchange. *Chem. Rev.* 106:1757–1768.

[190] Wales, T. E., and J. R. Engen. 2006. Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrom. Rev.* 25:158–170.

[191] Ivankov, D. N., and A. V. Finkelstein. 2001. Theoretical study of a landscape of protein folding-unfolding pathways. folding rates at midtransition. *Biochemistry.* 40:9957–9961.

[192] Garcia-Mira, M. M., M. Sadqi, N. Fischer, J. M. Sanchez-Ruiz, and V. Munoz. 2002. Experimental identification of downhill protein folding. *Science.* 298:2191–2195.

[193] Callender, R. H., R. B. Dyer, R. Gilmanshin, and W. H. Woodruff. 1998. Fast events in protein folding: The time evolution of primary processes. *Annu. Rev. Phys. Chem.* 49:173–202.

[194] Gruebele, M., J. Sabelko, R. Ballew, and J. Ervin. 1998. Laser temperature jump induced protein refolding. *Acc. Chem. Res.* 31:699–707.

[195] Eaton, W. A., V. Munoz, P. A. Thompson, E. R. Henry, and J. Hofrichter. 1998. Kinetics and dynamics of loops, α-helices, β-hairpins, and fast-folding proteins. *Acc. Chem. Res.* 31:741–753.

[196] Parker, M. J., and S. Marqusee. 2000. A statistical appraisal of native state hydrogen exchange data: Evidence for a burst phase continuum? *J. Mol. Biol.* 300:1361–1375.

[197] Ferguson, N., and A. R. Fersht. 2003. Early events in protein folding. *Curr. Opin. Struct. Biol.* 13:75–81.

[198] Otzen, D. E., L. S. Itzhaki, N. F. elMasry, S. E. Jackson, and A. R. Fersht. 1994. Structure of the transition state for the folding/unfolding of the barley chymotrypsin inhibitor 2 and its implications for mechanisms of protein folding. *Proc. Natl. Acad. Sci. USA.* 91:10422–10425.

[199] Kazmirski, S. L., K. B. Wong, S. M. Freund, Y. J. Tan, A. R. Fersht, and V. Daggett. 2001. Protein folding from a highly disordered denatured state: The folding pathway of chymotrypsin inhibitor 2 at atomic resolution. *Proc. Natl. Acad. Sci. USA.* 98:4349–4354.

[200] Day, R., B. J. Bennion, S. Ham, and V. Daggett. 2002. Increasing temperature accelerates protein unfolding without changing the pathway of unfolding. *J. Mol. Biol.* 322:189–203.

[201] Ferrara, P., J. Apostolakis, and A. Caflisch. 2000a. Targeted molecular dynamics simulations of protein unfolding. *J. Phys. Chem. B.* 104:4511–4518.

[202] Ferrara, P., J. Apostolakis, and A. Caflisch. 2000b. Computer simulations of protein folding by targeted molecular dynamics. *Proteins.* 39:252–260.

[203] Ozkan, S. B., G. S. Dalgýn, and T. Haliloglu. 2004. Unfolding events of chymotrypsin inhibitor 2 (CI2) revealed by Monte Carlo (MC) simulations and their consistency from structure-based analysis of conformations. *Polymer.* 45:581–595.

[204] Duan, Y., and P. A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science.* 282:740–744.

[205] Zagrovic, B., C. D. Snow, M. R. Shirts, and V. S. Pande. 2002. Simulation of folding of a small α-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* 323:927–937.

[206] Snow, C. D., N. Nguyen, P. V. S., and M. Gruebele. 2002. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature.* 420:102–106.

[207] Snow, C. D., L. Qiu, D. Du, F. Gai, S. J. Hagen, and V. S. Pande. 2004. Trp zipper folding kinetics by molecular dynamics and temperature-jump spectroscopy. *Proc. Natl. Acad. Sci. USA.* 101:4077–4082.

[208] Cavalli, A., M. Vendruscolo, and E. Paci. 2005. Comparison of sequence-based and structure-based energy functions for the reversible folding of a peptide. *Biophys. J.* 88:3158–3166.

[209] Tirado-Rives, J., M. Orozco, and W. L. Jorgensen. 1997. Molecular dynamics simulations of the unfolding of barnase in water and 8 m aqueous urea. *Biochemistry.* 36:7313–7329.

[210] Wang, L., Y. Duan, R. Shortle, B. Imperiali, and P. A. Kollman. 1999. Study of the stability and unfolding mechanism of BBA1 by molecular dynamics simulations at different temperatures. *Protein Sci.* 8:1292–1304.

[211] Tsai, J., M. Levitt, and D. Baker. 1999. Hierarchy of structure loss in md simulations of src sh3 domain unfolding. *J. Mol. Biol.* 291:215–225.

[212] Gsponer, J., and A. Caflisch. 2001. Role of native topology investigated by multiple unfolding simulations of four sh3 domains. *J. Mol. Biol.* 309:285–298.

[213] Sham, Y. Y., B. Ma, C.-J. Tsai, and R. Nussinov. 2002. Thermal unfolding molecular dynamics simulation of escherichia coli dihydrofolate reductase: Thermal stability of protein domains and unfolding pathway. *Proteins.* 46:308–320.

[214] Ma, B., and R. Nussinov. 2203. Molecular dynamics simulations of the unfolding of β_2-microglobulin and its variants. *Protein Eng.* 16:561–575.

[215] Morra, G., M. Hodoscek, and E.-W. Knapp. 2003. Unfolding of the cold shock protein studied with biased molecular dynamics. *Proteins.* 53:597–606.

[216] Merlino, A., G. Graziano, and L. Mazzarella. 2004. Structural and dynamic effects of α-helix deletion in Sso7d: Implications for protein thermal stability. *Proteins.* 57:692–701.

[217] Pande, V. S., T. Tanaka, and D. S. Rokhsar. 1998. Pathways for protein folding: Is a new view needed? *Curr. Opin. Struct. Biol.* 8:68–79.

[218] Reich, L., and T. R. Weikl. 2006. Substructural cooperativity and parallel versus sequential events during protein unfolding. *Proteins.* 63:1052–1058.

[219] Lazaridis, T., and M. Karplus. 1999. Effective energy function for proteins in solution. *Proteins.* 35:133–152.

[220] Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and K. M. 1983. Charmm: A program for macromolecular energy, minimization, and dynamics calculation. *J. Comp. Chem.* 4:187–217.

[221] Feig, M., and C. L. Brooks III. 2004. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.* 14:217–224.

[222] Zar, J. H. 1972. Significance testing of the Spearman rank correlation coefficient. *J. Am. Stat. Assoc.* 67:578–580.

[223] Babu, C. R., V. J. Hilser, and A. J. Wand. 2004. Direct access to the cooperative substructure of proteins and the protein ensemble via cold denaturation. *Nat. Struct. Mol. Biol.* 11:352–357.

[224] Maity, H., M. Maity, M. M. G. Krishna, L. Mayne, and S. W. Englander. 2005. Protein folding: The stepwise assembly of foldon units. *Proc. Natl. Acad. Sci. USA*. 102:4741–4746.

Acknowledgements

Finally, I would like to thank all those who helped and contributed to this thesis, in particular:

- Reinhard Lipowsky for his continuous support on my scientific route, especially during my last five years as a group leader in the department of Theory and Bio-Systems, which was a firm basis for the projects summarized in this habilitation thesis.
- Ken Dill for introducing me to essentially all topics touched in this thesis, and for inspiring discussions during my postdoc stay at the University of California, San Francisco, and in succeeding and ongoing projects.
- My co-workers in the scientific projects summarized here, Ken Dill, Purushottam Dixit, Claudia Merlo, Matteo Palassini, and Lothar Reich, for providing essential parts and ideas of this thesis.
- My co-workers in other projects, David Andelman, Mesfin Asfaw, Marion Becker, Jay Groves, Wolfgang Helfrich, Shige Komura, Misha Kozlov, Heinrich Krobath, Reinhard Lipowsky, Roland Netz, Bartosz Rozycki, Gerhard Schütz, and Robert Seckler.
- All members of the Department of Theory and Bio-Systems at the MPI of Colloids and Interfaces, and the members of the Dill lab at the University of California, San Francisco, for providing creative and stimulating environments.
- Last but not least, my partner Anja Brietzke and our families and friends for their encouragement and support.