

Stony Brook University



OFFICIAL COPY

The official electronic file of this thesis or dissertation is maintained by the University Libraries on behalf of The Graduate School at Stony Brook University.

© All Rights Reserved by Author.

Protein Dimerization Mechanisms Study with Molecular Dynamics Simulation

A Dissertation presented

by

Yuan Yao

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

December 2014

Stony Brook University

The Graduate School

Yuan Yao

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

Jin Wang - Dissertation Advisor
Associate Professor, Department of Chemistry

Xiaolin Li - Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Thomas MacCarthy
Assistant Professor, Department of Applied Mathematics and Statistics

Charles M. Fortmann
Associate Professor, Department of Material Science and Engineering

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Protein Dimerization Mechanisms Study with Molecular Dynamics Simulation

by

Yuan Yao

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2014

Protein dimerization is involved in many essential biological processes like gene expression, allosteric regulation, enzymatic activation and signal transduction. However, our understanding about dimerization mechanisms is limited despite of the progress made. In this work, we applied molecular dynamics simulation with all-atom structure-based model to the investigation of ten commonly studied regulatory dimers. Through the combinational analysis of intrinsic energy funnels, density of states, thermodynamic free energy landscapes, phi-values of transition states, and kinetic simulations, new and detailed mechanisms of dimerization and their coupling with monomer folding are revealed, in good agreements with existing experimental evidences and suggestions. We found five distinct representative strategies from ten dimers, demonstrating the high diversity and uniqueness in protein dimerization mechanisms, which challenge, or complement, the conventional concepts of two-state (obligatory) and three-state (non-obligatory) dimers. The all-atom structure-based model used here is demonstrated to have better representation for hydrogen bonds and hydrophobic packing, be able to distinguish disulfide bonds from native contact pairs, and give more detailed agreements with experiments than the previous course-grained structure-based model.

Lambda Cro repressor is one of the most studied dimeric transcription factors. But there is still an unsettled debate for decades about whether it is a two-state dimer or three-state dimer. We provide a new mechanism model that can reconcile these seemingly conflicting (mutually exclusive) experimental results. From simulations with all-atom structure-based model, we observe that the dimerization process of Lambda Cro repressor starts from one folded monomer with one unfolded monomer. Intra-subunit folding and inter-subunit binding are half-coupled, in a fly-casting manner.

Contents

1	General Introduction	1
1.1	Bio-molecular structure-function relationship	1
1.2	Conventional understanding of dimer association.	1
1.2.1	Classification of protein dimers	2
1.2.2	Coupling between binding and folding in dimerization process	3
1.3	Funneled landscape theory	4
1.4	Relationship between binding and folding	5
1.5	Folding landscape and topography ratio Λ	5
1.6	All-atom structure-based model	5
2	Methods	7
2.1	Simulations	7
2.1.1	Replica exchange molecular dynamics simulations (REMD)	7
2.1.2	Weighted histogram analysis (WHAM)	7
2.1.3	Software packages and computation resources	8
2.2	Calculations	8
2.2.1	Order parameters selection	8
2.2.2	Native contacts definition and calculation.	8
2.2.3	Entropy S calculation	9
2.2.4	Topography ration Λ calculation	10
2.2.5	Energetic roughness calculation	10
2.2.6	Folding temperature T_f calculation	10
2.2.7	Thermodynamic stability	11
2.2.8	Kinetic folding speed	11
2.3	Potential function	11
2.4	Characterization and visualization	12
2.4.1	3D free energy landscapes	12
2.4.2	Contacts formation maps along minimal free energy paths.	12
2.4.3	Contacts formation evolutionary map	13
2.4.4	Phi value calculations.	13
3	Characteristic funneled binding-folding landscapes	15
3.1	One dimensional intrinsic landscapes	15
3.2	Quantify the intrinsic landscape parameters	21
3.3	Two dimensional density of states landscape $\Omega(Q_i, Q_{a+b})$	21
4	Neither two-state nor three-state: Dimerization of Lambda Cro repressor	25
4.1	Free energy profiles with four states	26
4.2	Structural evolution in dimerization	28
4.3	Phi-values for two transitions	35
4.4	Experimental evidences	37

5	Diverse mechanisms of coupled binding-folding.	39
5.1	Inversion stimulation factor (1f36)	41
5.2	Arc repressor (1arr)	48
5.3	Lambda repressor (1lmb)	55
5.4	Streptomyces subtilisin inhibitor (3ssi)	61
5.5	LFB1 transcription factor (1lfb)	67
6	Landscape topography of dimerization	70
6.1	Density of state of the funneled landscape	70
6.2	Binding-folding mechanism in Free energy profile	71
6.3	Correlation between λ and thermodynamic stability	83
7	Kinetics about binding-folding	86
8	Discussion	89
8.1	Structure-based model with minimal frustration	89
8.2	Other factors influencing protein dimerization	89
8.2.1	Involvement of DNA molecule	89
8.2.2	Co-translational folding/binding	89
8.2.3	Chaperone	90
8.2.4	Slower dimerization $F + U \rightleftharpoons D$	90
9	Summary	91

List of Abbreviations

REMD:	Replica Exchange Molecular Dynamics
WHAM:	Weighted Histogram Analysis Method
MD:	Molecular Dynamics
1arr:	Arc repressor
1cta:	Troponin C site III
1cop:	Lambda Cro repressor
1f36:	Factor for inversion stimulation
1lmb:	Lambda repressor
1flb:	LFB1 transcription factor
1bet:	beta nerve growth factor
2gvb:	Gene V protein
2oz9:	Trp repressor
3ssi:	Streptomyces subtilisin inhibitor
Tf:	Transition temperature (including both folding and binding is not specified)
Tg:	Glass transition temperature (including both folding and binding is not specified)
n():	number of state (density of state)
DOS:	density of states
Λ :	landscape topography ratio
δE :	energetic gap between native state and non-native ensemble
ΔE :	energetic roughness of non-native ensemble
Qi:	inter-chain native contacts between two monomers of the dimer; binding descriptor
Qa(Qb):	intra-chain native contacts within chain A (chain B); folding descriptor
Q:	general reference to any of Qi, Qa, Qb, or all of them
2S:	two states dimer, in which folding and binding happen together/coupled
3S:	three states dimer, in which folding and binding processes is separable/distinguishable

Acknowledgements

I would like to express my great gratitude to my advisor Dr. Jin Wang, who gives me great freedom as well as guidance to match up my math and physics background, to explore and develop my research interests, and offers invaluable advices in both research and life.

I would like to thank all my lab mates for their supports, especially Dr. Zhiqiang Yan, who introduced me to programming, Dr. Xiakun Chu, who helped a lot in developing analysis methods, and Cong Chen, who went through my defense presentation and offered very constructive suggestions.

I want to say thank you to my teammates of XChange coop club: You Quan Chong, Jun Duan, Hardy Wu, Yiwen Pan and Si Wen. I feel very lucky to meet them in my graduate years.

I also want to thank my supportive family, especially my elder sister. It is their love that encourages me to overcome any hard time and achieve this far.

Finally, thanks to funding support from National Science Foundation, this work becomes possible.

1 General Introduction

1.1 Bio-molecular structure-function relationship

As the major carrier of biological functions, proteins adopt diverse and specific three dimensional structures, in order to implement the genetic hereditary information encoded in nucleotide sequences. The spatial arrangement of specified atoms (structure) gives a protein geometric shape, surface, electrostatic charge distribution, chemical characters, and underlying topology, on which a protein's dynamic behavior, i.e. function, relies. The deep rooted maxim that structure determines function, has been well accepted and followed during the development of experimental technologies, especially X-ray crystallography and nuclear magnetic resonance (NMR), which enable people to reveal the atomic structure of large biomolecules.

Resolved bio-molecular structures in protein data bank (PDB) has passed 100,000 by early 2014. A great amount of valuable information has been and will be learned from these structures, including activation sites and catalytic mechanisms of enzymes, key residues for protein functions, possible targets for drug design, dimer association and complex assembly mechanisms. However, in order to fully understand structure-function relationship, static structures from PDB database are still not enough. Since the resolved structures are mostly about the native state of a protein or occasionally very stable intermediate states, which are at low energetic state, and deprived from solvent environment, a PDB structure is likely possessing a different conformation from its functional conformation in vivo. More importantly, the dynamic behavior of these folded structures and the folding processes of achieving these end-point structures can carry more meaningful information in terms of interpreting mechanisms and designing new desired functions. The relatively low temporal and spacial resolution of common biophysical and biochemical experiments can only capture the overall crude structural evolution of bio-molecules, as an ensemble average most of the time. Molecular dynamics (MD) simulation, on the contrary, can give single molecule atomic level resolution. Given the recent rapid improvement of empirical molecular force fields and fast growth of computational power, molecular dynamics simulation provides a new way to study the dynamic properties of bio-molecules.

1.2 Conventional understanding of dimer association.

Protein dimerization [1, 2, 3, 4] is involved in many essential cellular events, including gene expression, signal transduction, protein transportation, cell skeleton assembly etc. The DNA-binding domains of transcription factors usually take the form of homodimer in order to recognize palindromic DNA-sequence motifs. [5, 6]. Many cross membrane signal receptors are dimers, so that they can mutually activate each other when proper signal comes [7, 4]. The theoretical models used to explain protein association processes have been evolving towards understanding about structure-function relationship of biomolecules. Fischer, as early as 1894, first proposed the rigid docking model for protein ligand association (lock and key model) [8]. Later, local flexibility of biomolecules was taken into consideration, resulting in conformational selection model [9] and induced fit model [10, 11]. The former emphasizes on the multitude of biomolecular conformations and random search for the correct interface; while the later suggests a more actively guided search and optimization of binding interface from conformational change induced from binding. More recently, the discovery of the ubiquitous existence of intrinsically disordered proteins (IDPs)

[12, 13, 14] revolutionized the conventional paradigm of the structure dependent protein functions. Being highly flexible before encountering the right partner, an intrinsically disordered protein can increase its searching diameter through fly-casting mechanism [15, 16], and at the same time, free its target-binding specificity from the necessity of high affinity required in structured proteins. The specificity without high affinity (i.e. thermodynamic stability) further enables IDPs to unbind easily. This subtle balance between affinity and specificity makes IDPs especially suitable for signal transduction and gene regulations [17, 18, 19, 20]. Following this major progress, our work here further revealed the diversity and complexity of protein association mechanisms, even with homodimers. New mechanisms with finer details are revealed, which agree with current experimental evidences and suggest new hypotheses for future experimental tests.

1.2.1 Classification of protein dimers

There are a few common ways to classify protein dimers. They emphasize on different aspects of their dimerization mechanisms. Since overlapping commonly exists among these classification methods, equivalences and implications are often made among them. However, as we will demonstrate later, the subtle differences of these classification terminologies should be better understood with greater caution.

Based on whether a thermodynamically stable intermediate state exists or not, between unfolded unbound monomers ($2U$) and structured dimers (D), people classify protein dimers into 2 state dimers and 3 state dimers [21, 22, 23]. A 2 state dimer has only one transition $D \rightleftharpoons 2U$, without any intermediate state. It does not have a structured monomeric form, and thus requires binding to its partner to get folded. While a 3 state dimer possesses a folded monomeric intermediate state ($2F$), or a partially structured dimeric intermediate (I_2), and thus has two transitions $D \rightleftharpoons 2F \rightleftharpoons 2U$ or $D \rightleftharpoons I_2 \rightleftharpoons 2U$.

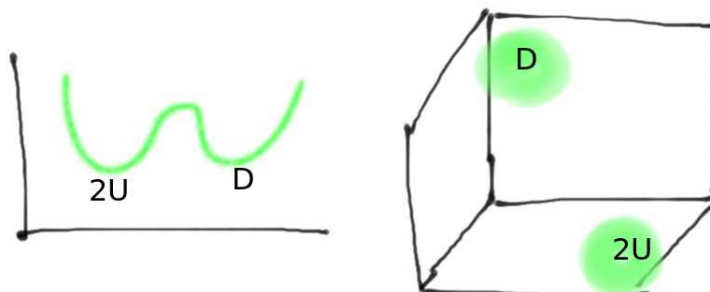


Figure 1.1: 2 state (obligatory dimer) $D \rightleftharpoons 2U$

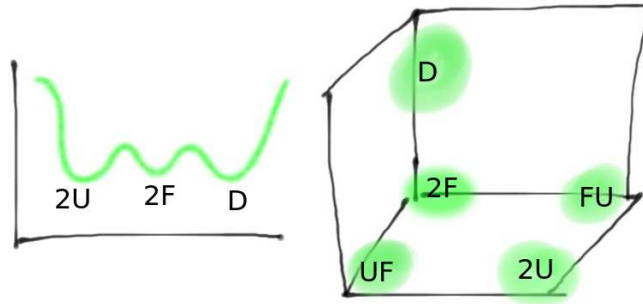


Figure 1.2: 3 state (non-obligatory) $D \rightleftharpoons 2F \rightleftharpoons 2U$

Based on whether a dimeric form is required/obligatory or not for the protein to get structured (folded), protein dimers are classified into obligatory dimers and non-obligatory dimer. Equivalences are often made between 2 state dimer and obligatory dimer, and between 3 state dimer and non-obligatory dimer. However, one clear inconsistency is that a 3 state dimer with dimeric intermediate state I_2 is not a non-obligatory dimer, since it does not have a folded monomeric form either.

Based on whether the binding interaction is strong enough to keep the dimer stable or not, dimers can be classified into permanent dimers and transient/temporary dimers. Based on whether the two partners in a dimer are the same protein or not, dimers are classified into homodimers and heterodimers.

1.2.2 Coupling between binding and folding in dimerization process

The discovery of ubiquitous intrinsically disordered proteins (IDPs) directly demonstrated that the folding process of a protein can be closely associated with its binding process. The association mechanisms of IDPs, according to the temporal order of binding and folding, are classified simply and intuitively into three classes: cooperative "coupled binding-folding", non-cooperative decoupled "fold first then bind" and "bind first then fold" scenarios [24]. The first two classes are most commonly observed in protein dimers. The third class "bind first" mechanism is much rarer, but still observed in small peptide ligands binding to their bigger protein partners [25].

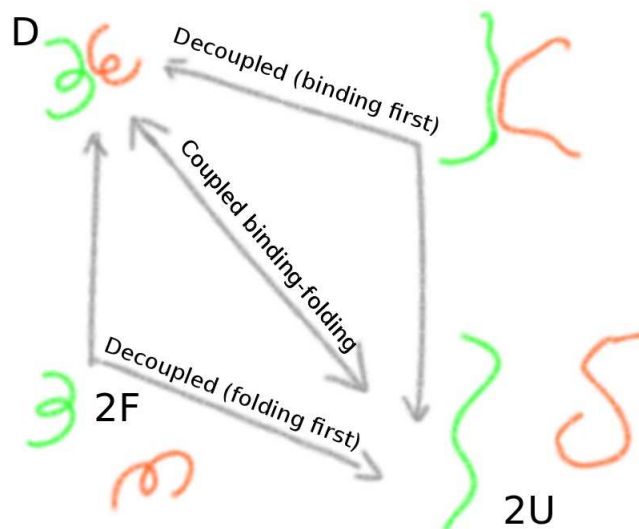


Figure 1.3: Coupling/decoupling between binding and folding.

Implications are commonly made that 2 state dimers always follow a coupled binding-folding mechanism, while 3 state dimers follow a "fold first then bind" mechanism at the presence of the intermediate state of structured monomers (2F). However, in our simulations, these implications does not always hold. Dimerization mechanisms display a much higher diversity in the ten dimers studied here, as demonstrated in the results section.

1.3 Funneled landscape theory

The development of protein folding funneled landscape theory has been providing us deeper insights into protein folding mechanisms for more than two decades [26, 27, 28, 29, 30, 31, 32, 33]. The previous concept of discrete sequential folding pathways connecting transition states and intermediates has gradually been replaced by the multiple parallel pathways among ensembles of states on the protein folding funneled landscape. [31].

Funneled landscape theory vividly incarnates current theoretical understanding about protein folding [26, 34, 27, 31]. Energy landscape profile of a structured protein in its physiological environment is thought of taking the shape of a funnel, with the bottom tip as the native folded state, and the wide open mouth as the ensemble of non-native states. Protein folding can be imagined as a ball rolling down the funnel, from a vast (high entropy, unstructured) high energy ensemble narrowing down to a tiny (low entropy, well structured) low energy native state. The wall of the funnel can be smooth or bumpy due to the intrinsic topological or energetic frustrations; and there can be multiple routes going down the funnel [35, 29, 36]. Since the possible conformational space of a protein is astronomically vast and thus requires infinite time to perform a brutal random search, (Levinthal paradox [37]), there exists a selective necessity for a structured protein to be able to fold into its functional form within biological timescales. This evolutionary pressure determines that the bumps and traps (i.e. roughness of the landscape) on the funnel wall, scaled by the size of the funnel, must be small enough, relative to the overall funnel shape, so that they can not delay folding process (i.e. going down the funnel) significantly. This evolutionary smoothness is

formulated into minimal frustration assumption [33, 38, 39], based on which structure-based model [40, 41], is developed and has given many satisfying agreements with experimental observations so far [33, 21, 42, 22, 43, 24, 19, 44, 25].

But the quantitative interpretation and application of funneled landscape are still largely immature. Previous studies have demonstrated that the topography ratio ($\Lambda = \delta E / (\Delta E \sqrt{2S})$) of funneled landscape is a good quantitative descriptor for the thermodynamics and kinetics of protein folding [45].

1.4 Relationship between binding and folding

Sharing the same type of energetic interactions among residues which drive the dynamics of protein folding, binding process differs from folding process only in the entropy part, due to different peptide chain connectivity. Therefore, an analog can be made between folding and self-binding, or between binding and multi-domain protein folding [17, 24]. Under a similar selective pressure to approach partner and form final native binded state within physiological timescales, the underlying energy landscape of binding also has to be funneled. Moreover, similar to the hydrophobic core of a structured monomeric protein, which composes the major contribution to folding stability [46], the commonly existing hydrophobic binding interfaces in protein dimers, especially in homodimers [47, 48], like the ten dimers we studied here, makes structure-based models equally applicable in studying coupled binding-folding processes. Coarse-grained model at residue level has successfully explained the dimer topology-mechanism relationship [21], short peptide sequence binding [25, 19], protein assembly [49] and fly-casting behavior in IDPs [50].

1.5 Folding landscape and topography ratio Λ

Previous work has shown three essential quantities in characterizing funneled folding landscape [45]. They are energy gap, energetic roughness and entropy. The energy gap δE is the energy difference between native/folded state and non-native/unfolded states, which is the assumed driving force/bias toward folding and stabilizes the native state after that. The depth of the funnel is reflected by this energy gap. The energetic roughness ΔE is energy fluctuation of the non-native state ensemble, which describes how rugged/bumpy the surface/wall of the funnel is. This roughness can cause the polypeptide chain, during its configurational searching, to be trapped in non-native local minima, and therefore delay or even prevent folding. The third quantity, entropy S , is the entropy of the non-native states ensemble, representing the volume of the configuration space need to be searched by the polypeptide chain, and thus depicts the width/horizontal scale of the funnel. The topography ratio $\Lambda = \delta E / (\Delta E \sqrt{2S})$, which is defined as the dimensionless ratio of the energy gap over energy roughness and entropy, has been demonstrated to be a robust metric that quantifies “the degree of funneledness”, which can be used to accurately predict the thermodynamic stability and kinetic speed of folding [45], with large Λ implying better thermodynamic stability against trapping, and faster folding speed.

1.6 All-atom structure-based model

In accordance with the concept of minimal frustration, which is shaped under natural selection pressure on folding speed and robustness in evolution [33], structure-based model [40], with only

native structure stabilizing interactions encoded, has enjoyed great successes in the study of protein folding mechanisms [33]. The commonly used structure based model only have C_{α} atoms in an amino acid represented, and therefore is described as coarse-grained model.

In recent years, Paul C. Withford, Jeffery K. Noel, Jose N. Onuchic etc. have developed an structure-based model that includes every single non-hydrogen atom of a protein [41]. This model combines the merits of being computationally economical with minimal frustrated Hamiltonian as well as being able to offer atomic level details, and has been demonstrated to be able to reproduce folding mechanisms in agreement with coarse-grained model and all-atom empirical forcefield results [41]. The developers also kindly provide one convenient web sever SMOG [51], which is used here, to help prepare our simulations.

2 Methods

2.1 Simulations

Our funnel topography ratio Λ calculation is based on protein’s density of states (DOS) in configurational and energetic space. Because the density of states is about micro-canonical ensemble, and therefore independent of temperature, it can serve as the fundamental intrinsic description of the system (protein folding-binding landscape here). In order to get the temperature independent intrinsic potential energy and entropy profiles, we need density of states from micro-canonical ensemble. To get density of states, we first performed all-atom structure based simulation with replica exchange method (REMD) [52] in canonical ensembles to sufficiently sample the dimerization configuration space. Then the temperature dependent energy distributions from REMD is transformed into micro-canonical density of states by weighted histogram analysis method (WHAM) algorithm [53, 54].

Free energy profiles are built directly from the configurational sampling at certain temperature replicas as the log of reverse probability ($G = -\log P + \text{const.}$), or from the density of states calculated by WHAM ($P = \sum \Omega_i * e^{-E_i/kT}$, Ω_i is number of states at energy level i , E_i is potential energy at level i). Two methods give very close free energy profiles. The unit of free energies calculated in this work is in Boltzmann factor (kT), where k is Boltzmann’s constant and T is temperature in reduced unit of structure-based model.

2.1.1 Replica exchange molecular dynamics simulations (REMD)

For each dimer, we set up 84 temperature replicas (with exceptions of 48 replicas for 1cta, and 60 replicas for 1arr due to their smaller sizes), and each replica is propagated for 50 nanoseconds. Starting, from native PDB structure, the first 10 nanoseconds of simulation is used to allow the structure to fully reach its equilibrium at each temperature, and then sampling of frames were collected from the last 40 nanoseconds, with a time interval of 0.5 picoseconds. Conformation exchange between neighboring replicas is attempted every 0.5 picoseconds. Success exchange rate between any two neighboring replicas is at least 60% around transition temperatures, and at least 20% everywhere else.

2.1.2 Weighted histogram analysis (WHAM)

Density of states, $\Omega(E)$, or microcanonical partition function, gives the number of microstates with energy E . $\Omega(E)$ is independent of temperature, and can be theoretically calculated from canonical sampling (constant temperature) at any temperature. However, one single constant temperature sampling is narrowly focused on a limited region of energy/conformation space (around the averages), and can give huge statistical errors at the under-sampled regions (tails of a distribution at a constant temperature). Thus, Ferrenberg-Swendsen reweighting method (WHAM) is applied to reconstruct density of states from replica exchange sampling at multiple temperatures [53, 54]. Through an iteration of reweighting, the statistical errors is minimized as a least-square optimization result. The algorithm used is as following.

In this work, a macrostate is defined by the number of native contacts formed (Q_i , Q_a , Q_b), which represents the formation of binding interface and folding of two monomers respectively.

Since energy is another necessary dimension in WHAM process, we calculate density of states as a function of four variables $\Omega(E, Qi, Qa, Qb)$. Probability of one state at replica t , $P_t(E, Qi, Qa, Qb)$, can be estimated directly from MD sampling. N is the number of replicas. T is the temperature in replica t . k is boltzmann constant. E is the potential energy of that state. The following two equations are iteratively solved, with input P_t from MD sampling and f_t initialized to be zero, until the change of f_t between two iteration cycles is less than 0.0001. (Different ending criteria like 0.01, 0.001 and 0.00001 are tested, and 0.0001 is chosen since there is no further improvement from lower values.)

$$\Omega(E, Qi, Qa, Qb) = \frac{\sum_{t=1}^N P_t(E, Qi, Qa, Qb)}{\sum_{t=1}^N \exp\{-E/kT - f_t\}}$$

$$f_t = \ln[\sum_E \Omega(E, Qi, Qa, Qb) e^{-E/kT}]$$

2.1.3 Software packages and computation resources

All molecular dynamics simulations are performed with Gromacs [55]. Input files are prepared with the help of SMOG server [51]. Center of mass constraint potential is applied in the form of a flat bottom well by PLUMED library [56]. WHAM procedure is coded in Fortran, and results are plotted and analyzed in Matlab.

Lonestar super-computing cluster in the University of Texas at Austin is used to implement the majority of our simulations. Replica exchange sampling makes an ideal parallel job in this cluster. Each replica takes one core in the cluster, giving a nice 90%+ parallel efficiency.

2.2 Calculations

2.2.1 Order parameters selection

To specifically define configuration states of a protein, as well as to visualize the landscape, we need to introduce specific order parameters, or reaction coordinates in the system. For dimers, we need at least two reaction coordinates to profile the landscape; one is for monitoring folding of the monomers, and the other is for monitoring binding of the two monomers. Several combinations of native contact number, root mean square deviation (RMSD) and center of mass (COM) distance have been tested as order parameters. The native contact number (Q) proves to be the only suitable candidate to monitor binding. Both RMSD and Q can monitor monomer folding, but using RMSD yields a streaky behavior in WHAM procedure. So, the formation of native inter-chain contacts Qi , native intra-chain contacts Qa and Qb was finally chosen as the reaction coordinates to describe binding and folding of two monomers respectively.

2.2.2 Native contacts definition and calculation.

Native contacts are determined by shadow map method [57]. The native state (PDB structure from X-ray crystallography or NMR) is used as the reference to define the native contacts. A cut off distance of 6 angstroms is first used to determine whether two atoms can possibly be in contact or not. And then, a filter removes contacts having a third atom blocking in between, who casts a shadow on one of the two atoms 2.1. All native contacts are defined by the native structure

before the simulations start. They can get lost in the simulation when their distance goes beyond their native distance times 1.2. By definition, the native state of a dimer will have the largest contact numbers, since once the structure deviates from the native structure, some of the native contacts will be lost and no new "native" contacts can be formed. That's the reason we use the number/percentage of native contacts formed to serve as the reaction coordinate of folding and binding.

The previous paragraph defines the atomic level contact between two atoms, the number of atomic native contacts can be huge, ranging from few hundreds to well over one thousand. As the developer of the model did, we reduced atomic level contacts to residue level contacts. As long as two residue has one pair of atoms in native contact, we say that two residues are in contact. By doing so, we get the residue level contact numbers, ranging from a few dozen to three hundreds. Then we rescale the residual level contact numbers by one or two, to make them stay below 150 for 2D histogram for free energy calculation. For 3D free energy calculations, we need to scale down the contact numbers further to make the number of bins in each dimension (Q_i , Q_a and Q_b) smaller than 50. For example, lambda Cro repressor has 532 atomic-level contacts in monomer A, 513 atomic-level contacts in monomer B; and 174 on interface between A and B. They add up to over one thousand contacts, which is way too many to visualize. So, we degenerate atomic-level contacts in to residue-level contacts. Any two residues that have at least one atomic contacts are regarded as being in contact. This results in 140 contacts in monomer A, 140 in monomer B, and 57 in between. In 2D free energy calculations, we use these as the maxima of Q_a , Q_b and Q_i . In 3D free energy calculations, we further reduce them into 46, 46 and 28.

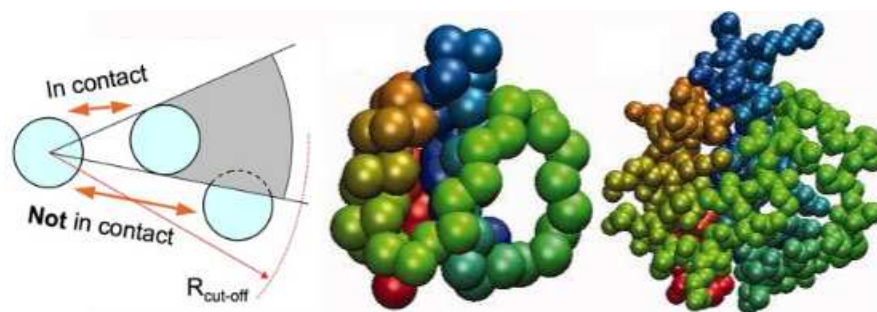


Figure 2.1: Shadow map method of defining native contacts

2.2.3 Entropy S calculation

After we got the density of state (DOS) from WHAM, entropy is defined as $S(E) = \ln[n(E)]$ (Boltzmann's formula), where $n(E)$ is the number of states with energy E , i.e. the scaled density of state. $S(E)$ describes how many available states are there with energy E , and thus measures the area of the cross-section of the funnel at energy level E .

In order to estimate the entropy of the non-native ensemble S_{non} , we set a cutoff criteria of the chosen reaction coordinates to separate native state configurations (folded and bound) from non-native states configurations (unfolded and/or unbound). Then a second WHAM procedure is performed on those non-native state configurations, which gives us the density of states for the non-native ensemble. The entropy of the non-native ensemble is therefore defined as $S_{non} = \ln\{\Sigma n(E)\}$, where $\Sigma n(E)$ is the sum of number of states over the non-native ensemble.

2.2.4 Topography ration Λ calculation

With ΔE and S_{non} derived, Λ is calculated in the formula $\Lambda = \delta E / (\Delta E \sqrt{2S})$. A illustration about how λ , Tf and Tg are calculated is given in figure 2.2.

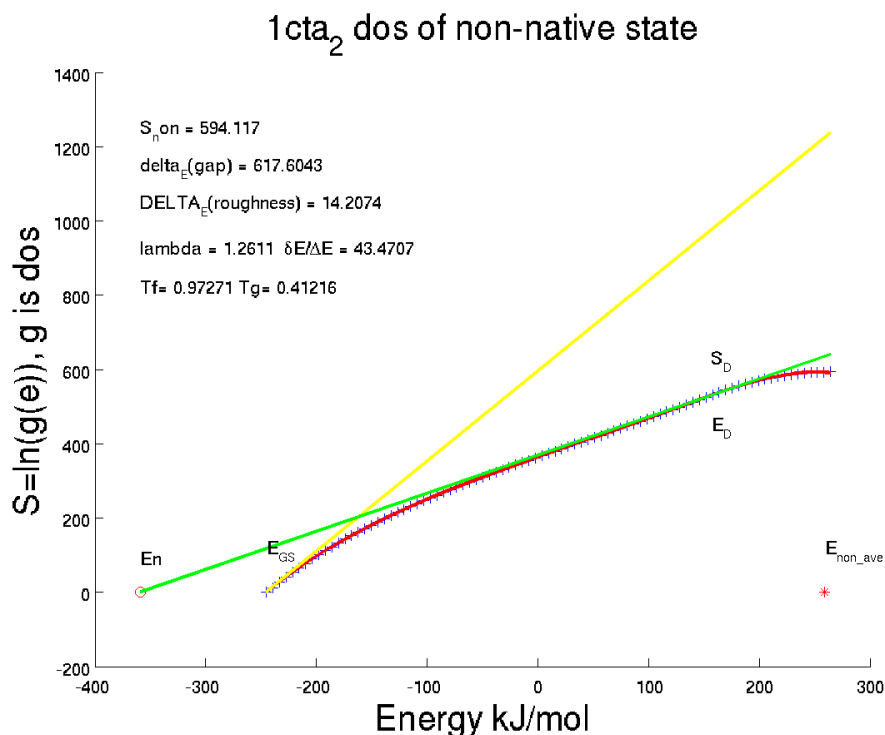


Figure 2.2: An example of how Tf and Tg are calculated. '+' signs indicate DOS of unfolded state at each energy level, and the red line is a 6 order polynomial fit of these data; Green line is the tangent line of the red curve which goes through native state (red 'o' on x-axis, labeled E_n), whose slope gives $1/T_f$; Yellow line is a tangent line which intersect with x-axis at the lower boundary of non-native ensemble, whose slope gives $1/T_g$; Red '*' indicates the average energy of non-native ensemble, with gap $\delta E = E_{non_ave} - E_n$.

2.2.5 Energetic roughness calculation

The energetic roughness ΔE is determined from the relation $T_g = \sqrt{\Delta E^2 / 2S_{non}}$, where the T_g is the glassy-like trapping temperature (glass transition temperature) [30]. T_g is the temperature at which the entropy of the non-native states vanishes, and below which the system becomes trapped or frozen. According to the thermodynamic Maxwell relationship $\partial S / \partial E = 1/T$, if we plot the entropy of non-native state curve as a function of E ($S(E) = \ln[n(E)]$), the inverse of the slope of the tangent line, at the intersection of the entropy curve and x-axis (where $S=0$), will give us the value of T_g . This process is illustrated by figure 2.2.

2.2.6 Folding temperature T_f calculation

The folding temperature T_f (Here, T_f refers to the transition temperature between native state ensemble and non-native state ensemble in general, including both folding and binding transitions.)

can be calculated in two ways. First, T_f is corresponding to the peak of heat capacity curve as a function of temperature. This is because the protein folding/unfolding or binding/unbinding event comes with a huge fluctuation of energy, and therefore gives a large heat capacity there. A second way utilizes the concept that the entropies of native and non-native states are equal at T_f , and thus $1/T_f$ is the slope of the tangent line of the non-native entropy curve, the same one mentioned in previous paragraph, which goes through the native ensemble (it is simply taken as the native PDB structure, which is a single point in my work, see Fig. 2.2). The second approach gives a T_f value depending on the chosen cutoff criteria which separates native configurations from non-native configurations. This criteria can be quite arbitrary as long as it meets the common sense in defining native state; say “within 2 angstrom RMSD from PDB structure” for example. But this cutoff makes much more sense if it gives a T_f value which matches the T_f derived from heat capacity curve in the first method. So, this gives us a rule of thumb in choosing the proper cutoff criteria value, by calibrating the two T_f s, which turns out to be tediously labor demanding when being carried out.

2.2.7 Thermodynamic stability

If we take folding transition temperature T_f as a reflection of the strength of native interactions in native state and glass transition temperature T_g as the strength of non-native interactions in other states, The thermodynamic stability of protein folding against trapping is characterized by the ratio of T_f over T_g [33]. And simulation studies has demonstrated that sequences with higher T_f/T_g ratio tend to fold faster [58, 59]. As we will show later, the T_f/T_g is positively correlated with Λ .

2.2.8 Kinetic folding speed

Constant temperature folding simulations are performed to get folding-binding speed for each dimer. For each dimer, we hope to fold 252 trajectories starting with different configurations and velocities. The initial unfolded and unbound structures are generated from high temperature trajectories in REMD. The temperature we surveyed for folding speed is T_χ [60], at which 80% of the population is in native state. We also performed a much thorough folding speed scanning process for these dimer. Folding of 256 random initial structures are attempted at a temperature range from 10 to 110 with a increasing interval of 10. Five of the smaller dimers give enough folding events, and their results are discussed in section 7.

2.3 Potential function

In this all-atom structure-based model, each non-hydrogen atom is represented by a single point of unit mass. Harmonic potential is used to maintain the bond lengths, bond angles, regular dihedral angles and improper dihedral angles at their equilibrium values found in the native PDB structure. Non-bonded native contacts are determined by shadow map method [57], and is represented in the form of Lennard-Jones potential. And all non-native contacts are repulsive. Details of the potential function and parameters are carefully discussed in developers’ publications [41].

The potential function is of the form:

$$V = \sum_{bonds} \epsilon_r (r - r_0)^2 + \sum_{angles} \epsilon_\theta (\theta - \theta_0)^2 + \sum_{improper} \epsilon_\chi (\chi - \chi_0)^2$$

$$\begin{aligned}
& + \sum_{backbone} \epsilon_{BB} F_D(\phi) + \sum_{sidechain} \epsilon_{sc} F_{SC}(\phi) \\
& + \sum_{contact} \epsilon_C \left[\left(\frac{\sigma_{ij}}{r} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r} \right)^6 \right] + \sum_{non-contact} \epsilon_{NC} \left(\frac{\sigma_{NC}}{r} \right)^{12}
\end{aligned}$$

where

$$F_D(\phi) = [1 - \cos(\phi - \phi_0)] + \frac{1}{2}[1 - \cos(3(\phi - \phi_0))]$$

with coefficients $\epsilon_r = 100$, $\epsilon_\theta = 100$, $\epsilon_\chi = 10$, $\epsilon_{NC} = 0.01$ and equilibrium values r_o , θ_o , χ_o , ϕ_o , σ_{ij} taken from native structure; and repulsive $\sigma_{NC} = 2.5\text{\AA}$. Setting of ϵ_{BB} and ϵ_{sc} need to take into account the numbers backbone dihedral, side chain dihedral and total native contacts. Please see the original publication for reference [41].

2.4 Characterization and visualization

2.4.1 3D free energy landscapes

To better reveal the mechanism of coupled binding-folding, 3 dimensional free energy landscapes are constructed from replica exchange molecular dynamics simulation. (e.g. Fig. 5.10).. In each 3D figure, the x-axis is the number of native contacts formed within monomer A, the y-axis is the number of native contacts formed in monomer B, and the z-axis is the number of native contacts formed on the interface between monomer A and B. Along the z direction (binding coordinate, Q_i), several representative cross-section layers are chosen, and, on each layer (i.e. an xy-plane with fixed z value), 2D free energy contours (describing monomers folding) are plotted. The red line is the minimal free energy path, connecting small red balls, which denote the minimal free energy points on those contour slices. (A straight diagonal red line in the space shall indicate perfect coupling between binding and folding. i.e. the red path in Figure 5.1). Note here, we restricted the red dots to the region where $Q_a \geq Q_b$, which is on the right side of the diagonal if we follow the diagonal from bottom to top. In this way, we avoid non-informative fluctuations around the cubic diagonal, which is due to the trivial free energy difference between imaging spots about the diagonal. So, in all the 3D landscapes, red minimal free energy path is always to the right of the diagonal; and, in the structural analysis, monomer A always have no less contacts than monomer B.

2.4.2 Contacts formation maps along minimal free energy paths.

For each free energy minimum of an xy-plane with an integer z value (Q_i), we collected structures around it from REMD simulation trajectories, and then constructed the contacts formation probability map, showing how likely each residue-pair contact is formed around that minimum spot. Though we have more detailed information about atom-pair contacts available, we have to degenerate it into residue-level contacts in order to reduce the number of contacts to a degree that is plot-able on a moderate sized map. The reduction is from 1000+ atomic contacts to less than few hundreds residue level contacts, varying case by case according to the size and topology of a dimer.

2.4.3 Contacts formation evolutionary map

To spot the potentially key residues/contacts that influence binding/folding processes, and to get a temporal understanding of the functions they take, we built the evolution map of contact formation (an indicator of nativeness of a residue's surroundings) along folding/binding process Fig 2.3. To read the map, we notice that the x-axis is the number of native contacts formed, from 0 to the maximal number of native contacts in PDB structure, representing the binding/folding process; and the y-axis is the residue number index. We often see a white stripe in the middle, showing the terminal residues of a monomer may not have any native contact. The color of each point in the map indicates the percent of native contacts formed for certain residues (read from y) at certain stage of the overall transition (read from x). Red means all contacts of that residue is completely formed; dark blue means none of them formed and white indicate missing sample points (non-exist). We can see clearly from most of the maps that there are outstanding strips/bands of bright colors (yellow to red) spanning over most part of x-axis, showing the corresponding residues form contacts early and sustain native-like contacts/shape most of the time. For more informative presentation, we placed colored bars between maps, with dark red indicating regions in which residues only form contacts within the same subunit (not involved in the interface), yellow color for residues form contacts with both subunits (on the interface), and blue for residues only form contacts with the other subunit (only covalently bonded to its subunits, very rare).

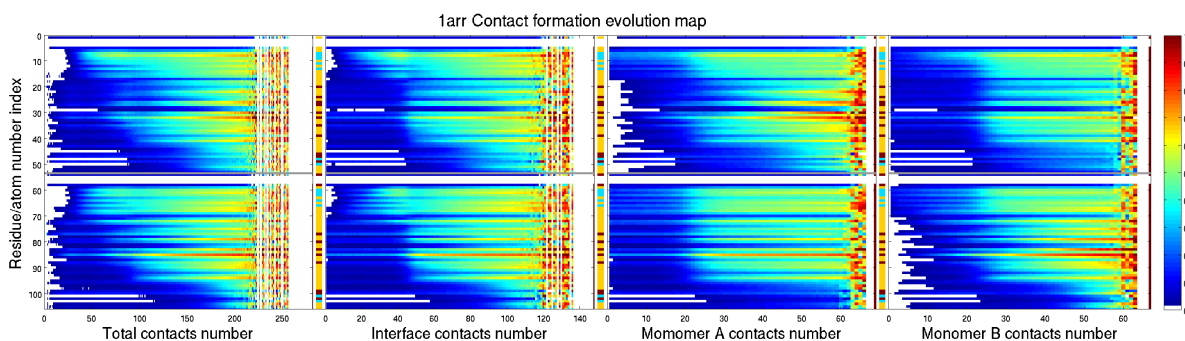


Figure 2.3: Contact evolution of 1arr

We build four different evolution maps to describe four transitions, the whole dimer, the interface, subunit A and subunit B, respectively. Basically, the contact evolution maps shows the conditional probability of each residue forming its native contacts at given native-likeness of the overall structure (In the four maps, the overall structure means the whole dimer, the interface, subunit A and subunit B respectively.)

To infer dynamic information from these evolution maps, we can use the location bar between maps to identify sequence locations (interface, monomer core, or both). Then we could read out a temporal order of the formation of different parts from the location and intensity of the high probability bands/strips or plateaus.

2.4.4 Phi value calculations.

Phi-value analysis in protein folding experiments is used to characterize the transition state structure. Through the comparison of the change of free energy difference between native state

and non-native state, and between transition state and non-native state, due to residue substitution (mutation), the phi value of a residue reveals how likely that residue obtains its native interactions at folding transition state. In simulation studies, the same purpose can be realized without actually implementing residue mutations. The commonly adapted method to calculate phi values from MD simulation is formulated as the ratio of the probability differences between transition state (TS) and unfolded state (U), over the probability difference between folded (N) state and unfolded state (U).

$$\Phi = \frac{P_{TS} - P_U}{P_N - P_U}$$

In this sense, having a phi value close to 1 means that native contacts of that residue is well-formed (exist) in transition state; while having a phi value close to 0 indicates that its native contacts barely exist in transition state. In our all-atom structure-based model, each atom (residue) can have multiple native contacts with other atoms, the phi value of a residue is the average of the phi-values of its contacts. In the calculating process, improper phi values, which are not within the range of 0 ~ 1, are filtered out. These improper phi-values appear because of large fluctuations of the rarely sampled contacts in transition state. To further reduce the error from statistical fluctuation, another filter is added to achieve stable phi values. We require that, for any contact pair to be reckoned reliable, its presence in folded state is higher than 40%, in transition state is higher than 20%, and in unfolded state is less than 30%. Contact pairs that do not pass these filters are not included in our contact phi-value map. During this process, we notice that the filtering criteria that unfolded state presence should be less than 30% is a very critical requirement, which filters out most noises (low counts in sampling). The rationalization for the 2nd filter is that, we want to focus our attention on the contact pairs that are involved in the transition process, which means they have to be not well-structured before the transition and better structured after the transition; at the same time have to be sampled reasonably well in transition state (>20%).

3 Characteristic funneled binding-folding landscapes

3.1 One dimensional intrinsic landscapes

The funneled protein folding landscape theory illustrates that the conformational entropy and intrinsic potential energy decrease as a protein approaches its native state. Recently, our group and other researchers have extended the funneled landscape theory to protein association [61, 62, 63, 39, 25, 17, 21, 49, 50, 44, 24, 16, 18] and demonstrated the interplay between underlying intrinsic binding landscape and folding landscapes [24]. One of the results and expectations is that less funneled binding energy landscapes is more likely to lead to non-cooperative (non-coupled) binding-folding. Here, we plotted the decreasing of entropy (i.e. density of state, Figure 3.3) and temperature independent intrinsic potential energy landscape (Figure 3.4) for monomer folding and dimer binding respectively (left half and right half). These decreasing slopes can be perceived as degenerated one dimensional landscapes, following the reaction coordinate of native contact formation, in monomer A and B, or on the interface between them. We place these slopes in a funnel like pattern with folding on the left hand side and binding on the right hand side, creating the incarnation of funnel analog and facilitating easier comparison. We observed a clear tendency, in agreement with other's results and expectations, that 3 state dimers (decoupled binding-folding) have less funneled and more rugged binding landscape, and 2 state dimers (coupled binding-folding) have less funneled and more rugged folding landscape.

We classify dimers into two groups: coupled binding-folding dimers and decoupled binding-folding dimers. This classification is based on how closely one's minimal free energy path, at transition temperature, follows the diagonal of the free energy cube (Fig. 1), which representing the perfect coupling between binding and folding. Four dimers have minimal path tracing the diagonal rather well, and thus are categorized into coupled binding-folding dimers; their PDB IDs are 1arr, 1f36, 1cta and 2oz9 (Figure 3.1). Another four fall into decoupled category; they are 2gvb, 1lmb, 3ssi and 1cop (Figure 3.2). Another dimer (1flb) is an extreme case of decoupled dimer. It even has two clearly separated transition temperatures for folding and binding respectively.

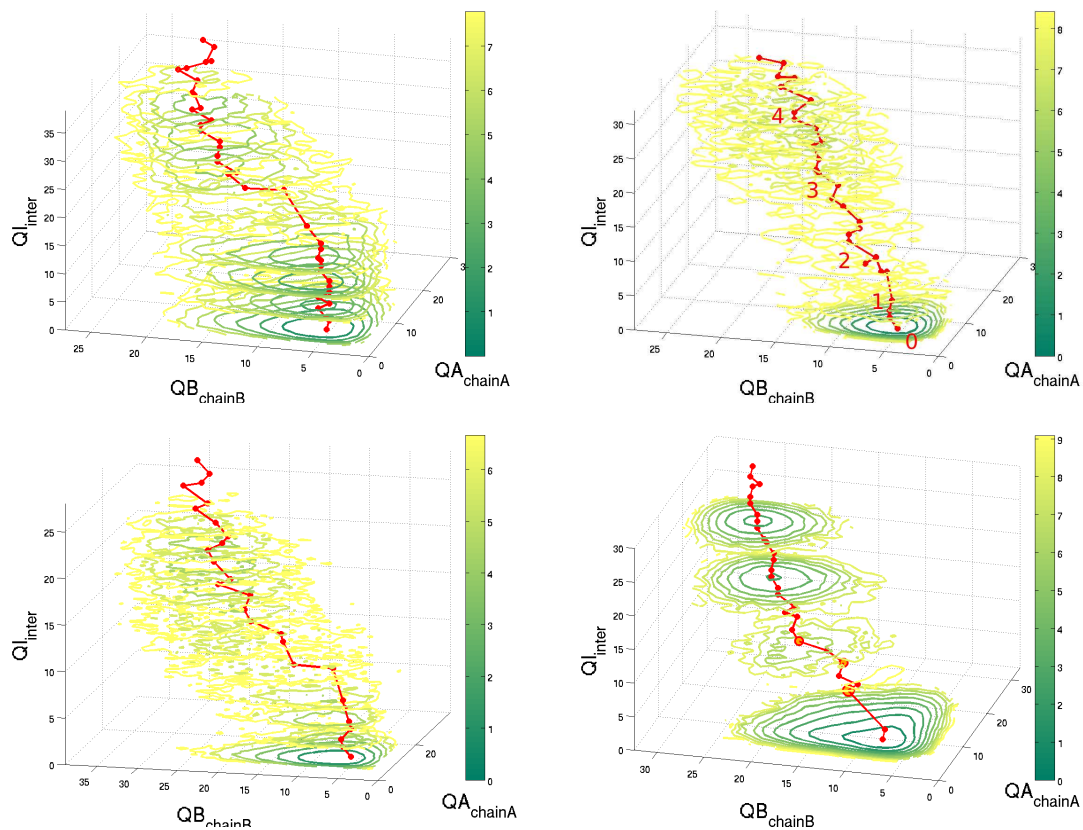


Figure 3.1: Free energy profiles of dimers with coupled binding-folding mechanisms. TTT label names to each figure.

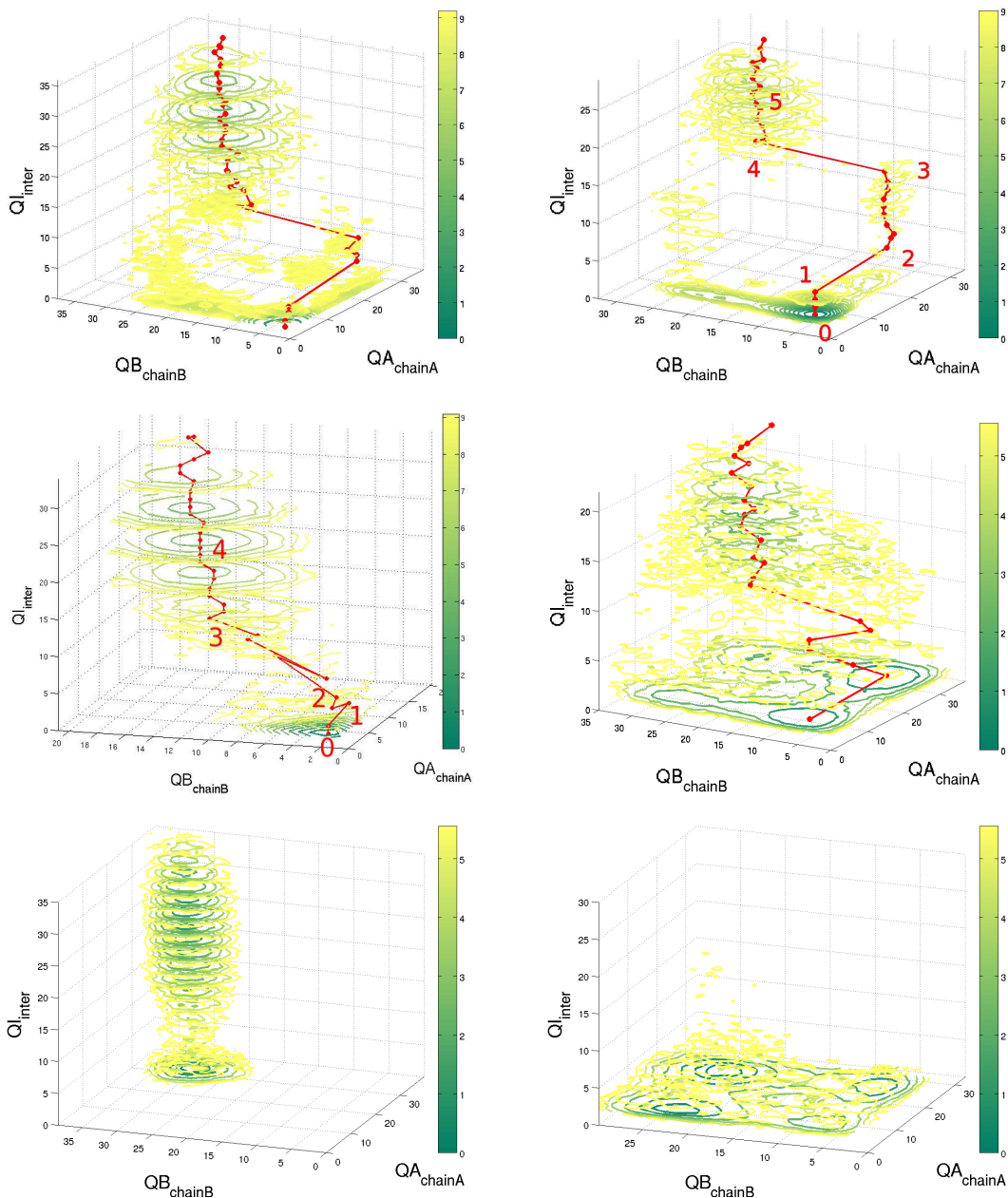


Figure 3.2: Free energy profiles of dimers with decoupled binding-folding mechanisms.

For both groups of dimers, their characteristic patterns are distinctively clear. As the most extreme case of decoupled binding-folding dimers, 1fb (purple solid line in Fig. 3.3 and 3.4) has a folding temperature (111) that is much higher than its binding temperature (103). We can see that its binding slope is overall much lower and flatter than its folding slope, especially after a big steep plunge at $Q_i \sim 0.2$ (which may correspond to the initial encounter of two subunits.); For 1cop (green solid line in Fig. 3.3 and 3.4), another 3-state decoupled dimer, we don't observe altitude or gradient difference between binding slope and folding slope, but it is clear that binding slope on

the right is much more rugged than folding slope on the left, indicating bigger bumps and traps on the binding funnel; For 1gvb (cyan solid line) and 1lmb (black dash line), another two decoupled dimers, the binding slope is not only significantly shallower and flatter, but also more rugged than its folding counterpart. Conversely, well characterized 2-state dimers with coupled binding-folding, like 1f36, 1cta, 2oz9 and 1arr, the binding slope and folding slope are more comparable in shape and smoothness (Fig. 3.3 and 3.4). We notice that, 1arr (red solid line), whose monomer folding follows a phase of significant initial binding between its beta strands, has a rougher surface on its folding slope, with two clear bumpy spikes.

Moreover, if we compare the folding slope of coupled dimers with those of decoupled dimers, it is clear that coupled binding-folding dimers (1f36, 1cta, 2oz9 and 1arr) also have a less funneled slopes with rougher surfaces for monomer folding than decoupled dimers (1cop, 1flb, 1gvb, 1lmb and 3ssi). This is especially obvious on the intrinsic energy landscape (Fig. 3.4). This agrees with the behavior of IDPs, which have a more rugged folding landscape, and only fold upon binding to proper partner [24]. Conversely, the binding slope of coupled binding-folding dimers are much more funneled and smoother than those of decoupled dimers.

On the funneled landscape of intrinsic potential energy (Figure 3.4), the red dot at the center of x-axis represents the potential energy of the native states of all dimers (after shifting). If we imagine that the red dot is the bottom of a golf hole, then the two slopes flanking it would be the terrain of the golf field around that hole. We can perceive that a golf ball on the decoupled binding-folding terrain will have a much harder time in reaching the golf hole from the binding side. While on the other hand, a coupled binding-folding dimer enjoys an overall smoother binding-folding funnel, which represents less fluctuation and more robustness. There are three possible evolutionary reasons/benefits of this coupled binding-folding approach in IDPs. First, if homodimer binding can be similar to, or nearly as easy and fast as monomer folding, half of genome coding can be saved, and at the same time the chance of errors from mutations is reduced. Second, binding speed is increase by fly-casting mechanism and high specificity is achieved without high affinity. Third, dimerization make 1st order kinetics into 2nd order kinetics, which may compose another layer of activity regulation base on the concentration.

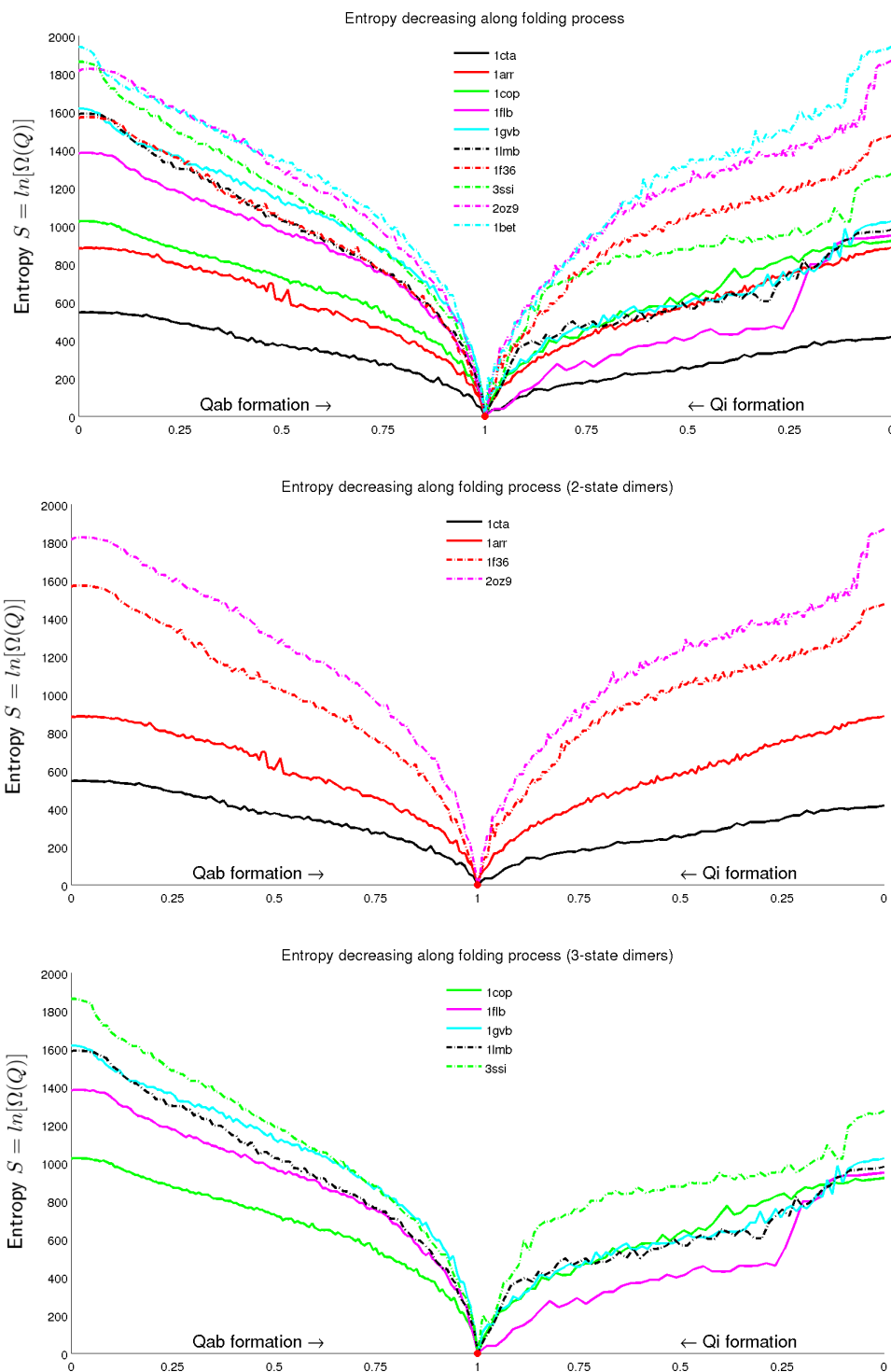


Figure 3.3: Entropy (density of state) decrease as folding or binding proceeds. The first figure has all ten dimers studied here included; the 2nd figure contains only 4 coupled binding-folding dimers; the third has only 5 decoupled dimers. We can observe that (1) binding landscapes and folding landscapes are more comparable in shape and smoothness for coupled dimers; (2) binding slopes is shallower, flatter, less funneled and more rugged than folding slopes in decoupled dimers; (3) decoupled dimers have more funneled and smooth folding landscapes than coupled dimers.

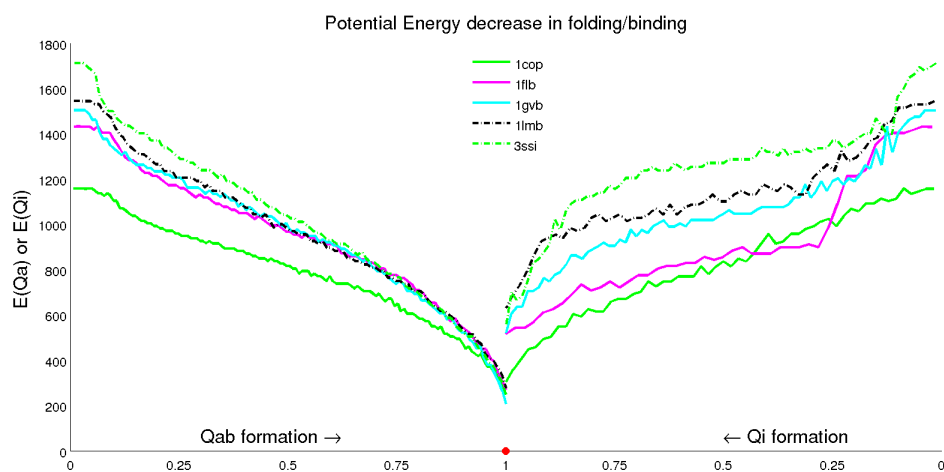
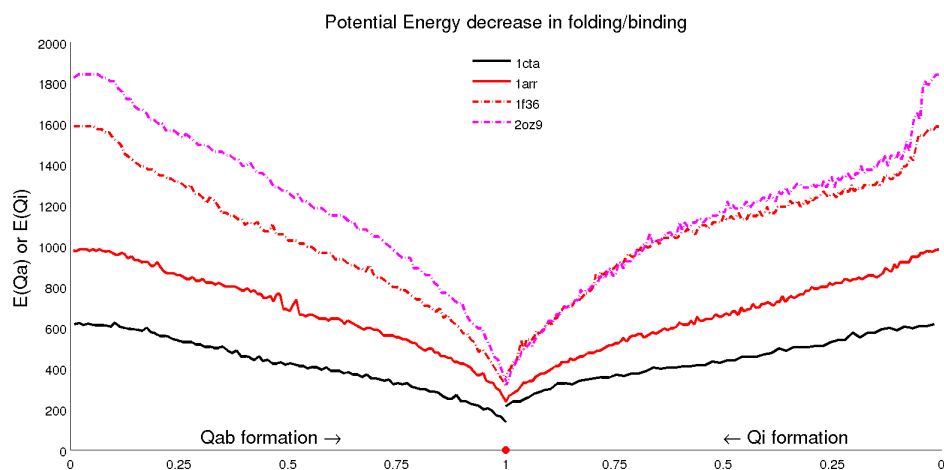
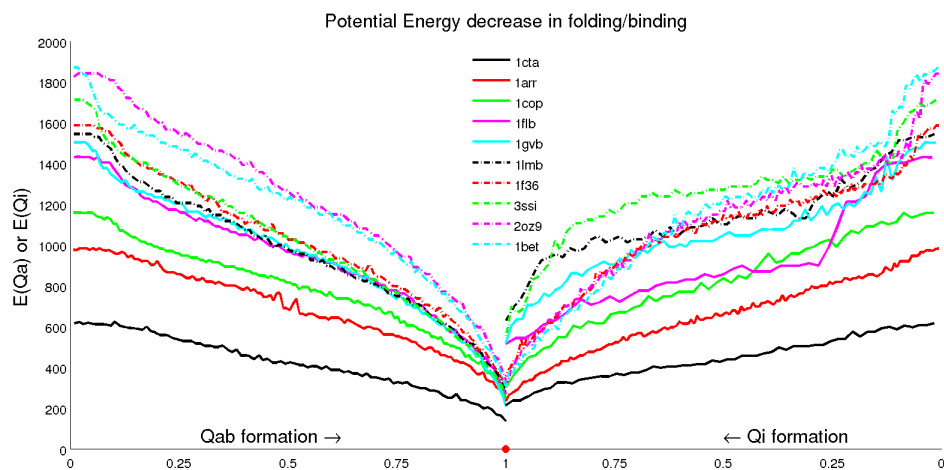


Figure 3.4: Intrinsic potential energy (dihedral + contacts) decrease as folding or binding proceeds. Similar observations as in figure 3.3.

3.2 Quantify the intrinsic landscape parameters

Several quantitative measures about the intrinsic funneled landscapes of dimers are calculated and listed below (table 1). We notice many of them are positively correlated with the size of the protein, including energy gap, roughness, ratio of gap to roughness, non-native state entropy, as expected. But only landscape topographic ratio λ is positively correlated with thermodynamic stability as measured by T_f/T_g (section 6.3).

Notice that the roughness calculated here is about the overall roughness of the non-native state of a dimer, including roughness from both folding and binding. Since the intrinsic landscapes displayed in previous part (Fig. 3.4) reveals clear differences between binding energetic roughness and folding energetic roughness, a rational next step in the future will be to calculate the energetic roughness for binding and folding respectively, which has been performed with residue-level structure based model [24].

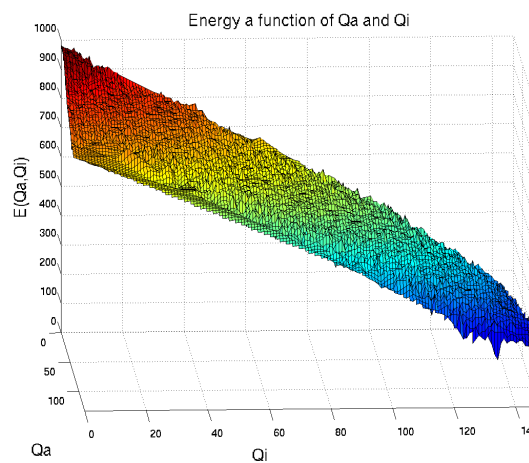
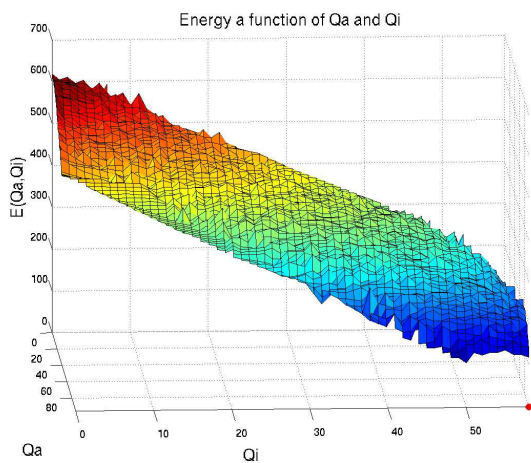
3.3 Two dimensional density of states landscape $\Omega(Q_i, Q_{a+b})$

Two dimensional density of states as a function of binding (Q_i) and folding (Q_a+Q_b) are constructed similarly for 2-state dimers (Fig. 3.5) and 3-state dimers (Fig. 3.6). It's hard to estimate the roughness or steepness of these landscapes directly by eye, but one trend is very clear between these two groups. 3-state dimers have broader surfaces than 2-state dimers. This may indicate more independence between binding and folding in 3-state dimers. While 2-state dimers have narrower stripe-like surfaces, suggesting strong binding-folding coupling.

name	Tf_Cv	Tf_tangent	Tg	lambda	roughness	size	Gap	S_non	dE/DE	#atom	EN
1cta	113.3	0.97271	0.41216	1.2611	14.2074	68	617.6	594.1	43.4704	538	-358.667
1arr	105.5	0.8727	0.36031	1.294	16.5433	106	982.9	1054	59.4149	870	-580
1cop	113.2	0.90782	0.37812	1.3	18.3498	132	1157.6	1177.5	63.0862	1036	-690.671
1flb	111.1	0.92686	0.38824	1.2725	20.8977	154	1431.3	1448.6	68.4923	1280	-853.346
1gvb	111.8	0.9512	0.40114	1.2461	22.0014	174	1503.6	1504.1	68.3434	1360	-906.657
1lmb	102.7	0.87175	0.37987	1.2235	21.9239	179	1548	1665.4	70.61	1380	-919.991
1f36	106	0.93829	0.41146	1.2026	23.3048	178	1587.4	1604	68.1136	1416	-944
3ssi	106.2	0.91615	0.39894	1.2119	23.7379	216	1711.8	1770.3	72.1126	1544	-1029.34
2oz9	109	0.87382	0.37163	1.2679	23.9667	208	1842	1954.7	79.2743	1656	-1104
1bet	110.5	0.91355	0.38028	1.2679	23.6936	214	1871.6	1941	78.996	1680	-1120

Table 1: Values of intrinsic landscape parameters. Name used here is the PDB ID of a dimer; Tf_Cv is transition temperature read from heat capacity curve; Tf_tangent is transition temperature read from the density of state curve; Tg is glass trapping temperature; lambda is landscape topographic ratio; roughness is for the overall energetic roughness of the non-native ensemble; size is dimer size in number of amino acids; Gap is the energy gap between native state and non-native ensemble average; S_non is the entropy of non-native state, measured by $\ln(\text{DOS})$; dE/DE is the ratio of energy gap to energy roughness; #atom is the number of atoms in a dimer; EN is the potential energy of native state. The units of temperature and roughness are in reduced units; the unit of energy is kcal; entropy and lambda have no units.

1cta (left) and 1arr (right)



1f36 (left) and 2oz9 (right)

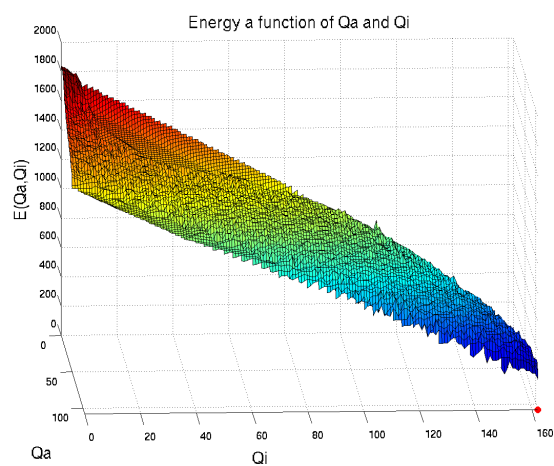
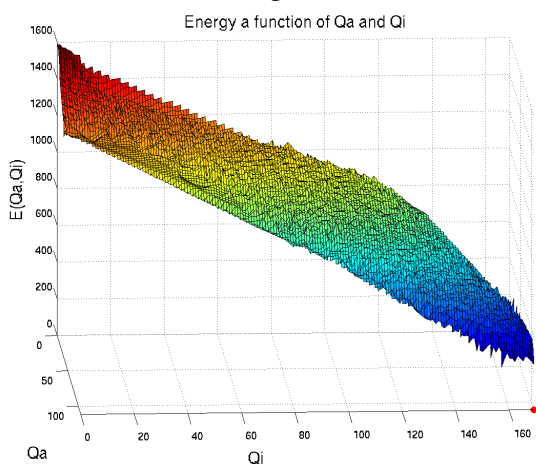
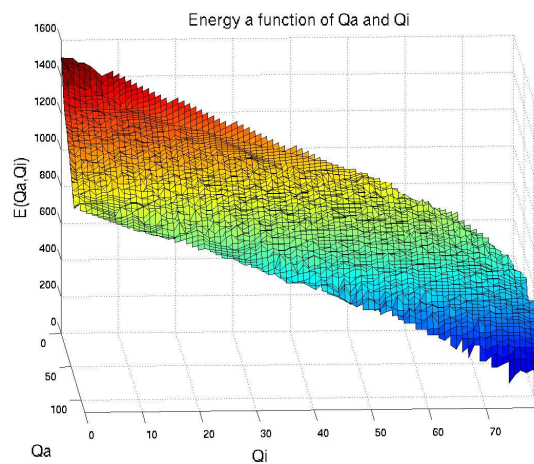
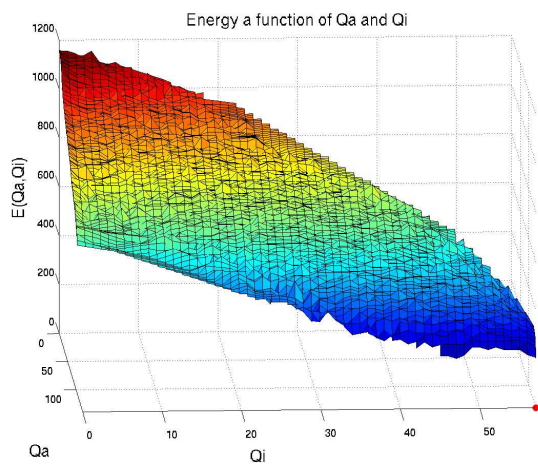
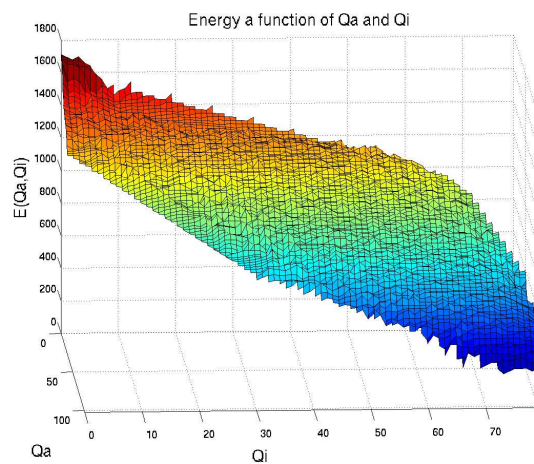
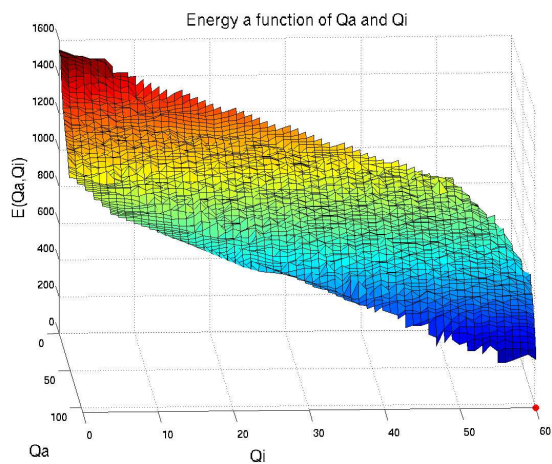


Figure 3.5: 2D density of state for two state dimers. Narrower ribbon, compared with three state dimers, indicates the strong coupling between binding and folding.

1cop (left) and 1gvb (right)



1lmb (left) and 3ssi (right)



1fbb

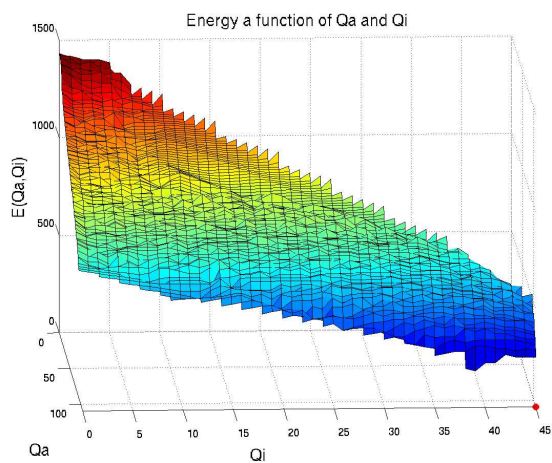


Figure 3.6: 2D density of state for three state dimers. Broader spectrum/surface, especially in the Q_a direction, suggest the higher independence between binding and folding.

4 Neither two-state nor three-state: Dimerization of Lambda Cro repressor

Lambda Cro repressor is one of the best studied dimeric transcription factors. But there is still an unsettled debate for decades about whether it is a 2 state dimer or 3 state dimer. We provide a new mechanism model that can reconcile these seemingly conflicting (mutually exclusive) experiment results. From simulations with all-atom structure-based model, we observe that the dimerization process of Lambda Cro repressor starts from one folded monomer with one unfolded monomer. Intra-subunit folding and inter-subunit binding are half-coupled, in a fly-casting manner.

Cro protein is a transcription repressor in bacteriophage Lambda. It is needed for lambda phage to enter lytic cycle, and competes with another repressor cI in determining the fate of host bacteria [64, 65]. It has been shown by X-ray crystallography and NMR experiments that Cro binds to DNA operators as a homodimer, which has a dimerizing interface between the third beta strands on the C terminal parts of two monomers [66, 67, 68, 69]. But its dimerization mechanism is not clear yet, despite the progress made. In thermal denaturation studies [70, 71, 72] and guanidine hydrochloride (GdnHCL) denaturation study [73], Cro dimer follows a one stage transition between only two significantly populated states, folded dimers (D) and unfolded monomers (2U). However later evidence from urea induced denaturation suggests the N-terminal part of a Cro monomer is partially structured [74]. And linear extrapolation of GdnHCL denaturation results with a 2-states dissociation model predicts a nanomolar dissociation constant at zero denaturant concentration, which is inconsistent with (unrealistically smaller than) the micromolar dissociation constant measured by DNA binding hydrodynamic experiments [75, 73]. These conflicts may result from different monitoring methods and/or nonidentical experimental conditions, but at the same time suggest the complexity involved in Cro dimerization [76, 75, 74, 77, 78, 71]. The conventional classification of 2 state dimer or 3 state dimer may not be sufficient to describe/characterize the real dimerization process of Cro repressor.

Protein homodimers are widely involved in many fundamental cellular functions, including enzyme activation, gene expression, and signal transduction. The current two most used models in experimental research to characterize dimerization mechanisms are 2-states model (also known as obligatory dimer) and 3-states model (non-obligatory dimer) [3]. 2-states model involves only one transition between one native structured dimer and two unbound unstructured monomers ($D \rightleftharpoons 2U$); 3-states model includes two transitions with an additional monomeric or dimeric intermediate state ($D \rightleftharpoons 2F \rightleftharpoons 2U$ or $D \rightleftharpoons I_2 \rightleftharpoons 2U$; F is folded monomer, I_2 is dimeric intermediate). In more recent computational simulation studies, thanks to the much finer spacial and temporal resolution in simulation, new terms like cooperative/coupled binding-folding, non-cooperative/decoupled binding-folding, binding-prior-folding, or folding-prior-binding are used to characterize dimerization process [24], emphasizing more on the kinetic part. These existing terms are more or less conceptually overlapping with each other, and in many cases, equivalency and implications are made among them in discussing dimerization mechanisms. For example, it is intuitive to assume non-obligatory dimers (3 state dimers) have a decoupled binding-folding process, with an intermediate state of folded yet unbound monomers. But in this work, as you will see later, this implication does not hold for Cro dimer. In our simulation, the folded yet unbound monomers are actually off-path intermediate states for dimerization process. Cro monomer folding can be decoupled from binding; but binding has to be coupled with the folding of one of the two

monomers in a dimer.

Our molecular dynamics simulations are based on a newly developed all-atom structure-based model [41, 51, 79]. This model only encodes native contacts found in PDB structure, as tradition course-grained structure-based model does. But this model defines native interaction at atomic level, with every heavy atom included. Companioned by shadow-map contacts identification method [79], which is specially designed to capture protein folding dynamics, this atomic-level structure-based model has been demonstrated to be able to reproduce the protein folding results achieved from course-grained model and empirical force-field MD simulations [41, 51, 79]. We will also demonstrate that this model captures hydrogen bonding and hydrophobic packing better than course-grained model. Replica exchange molecular dynamics (REMD) sampling is applied, with 84 replicas covering a wide temperature range from 20 degree to 280 degree. (Temperature is measured in reduced units. The transition temperature for Cro dimer is 113.2 degree at reduced units). Each replica trajectory is 58 nanoseconds long and gives 116,000 configurations (every 0.5 picoseconds). Exchange rate is maintained at least 20% between any two neighboring replicas. 10 replicas around transition temperature is closely arranged within 2 degrees, in order to have exchange rate as high as 60%, and to guarantee sufficient sampling of the transition process. Flat-bottom well potential is used to constrain the center of mass of two monomers within 55 angstroms from each other. Free energy profiles are built directly from certain temperature replicas as the log of reverse probability.

4.1 Free energy profiles with four states

Free energy profiles from 1 dimension to 3 dimensions are constructed (Fig 1 to 3). Together, they review the dimerization mechanism of lambda Cro repressor. At transition temperature, where dimer dissociation and monomer unfolding happen concurrently [80, 70], there are 4 significantly populated states: one structured dimer (D), two unfolded and unbound monomers (2U), two folded but unbound monomers (2F), and two unbound monomers with one folded and one unfolded (FU). These 4 states are observed universally in simulations of 3 state dimers [22, 23, 21]. Because the FU state is often ignored in discussion of dimerization mechanisms, people refer to this scenario as 3 state dimerization, even though the fourth state (FU) clearly exists given the coexistence of 2F and 2U states.

Reaction coordinate chosen to measure structural evolution is the number of formed native contacts (Q_i is inter-subunit contacts between two subunits; Q_a is intra-subunit contacts within monomer A; Q_b is intra-subunit contacts within monomer B). There are 532 atomic-level contacts in monomer A, 513 atomic-level contacts in monomer B; and 174 on interface between A and B. They add up to over one thousand contacts, which is way too many to visualize. So, we degenerate atomic-level contacts into residue-level contacts as the developers of all-atom model did. Any two residues that have at least one atomic contacts are regarded as being in contact. This results in 140 contacts in monomer A, 140 in monomer B, and 57 in between. You may have noticed that the number of native atomic-level contacts in monomer A is slightly higher than that in monomer B, because the native structure from NMR experiment is not completely symmetric [69]. And since the Hamiltonian of structure-based model relies on the input native structure, this asymmetry may contribute to the slight asymmetry between monomer A and monomer B in the following free energy profiles.

In one dimensional free energy profile, the only reaction coordinate is the total number of

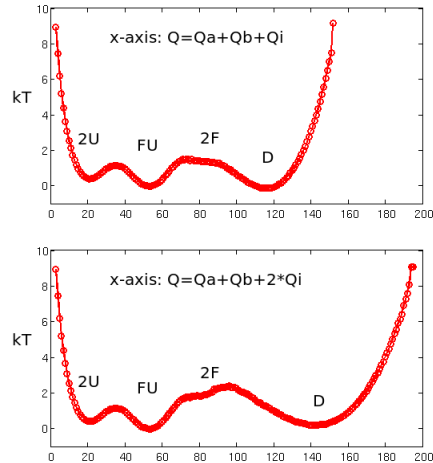


Figure 4.1: 1D free energy

formed native contacts ($Q_a+Q_b+Q_i$). Because the number of inter-subunit contacts is small compared with the intra-subunit contacts, the 2F state has a well merged with D state (Fig. 4.1 upper). Multiplying Q_i by 2 will push the D well further, so that 2F well is better revealed (Fig. 4.1 lower). Because the transition temperature is identified as the peak of heat capacity curve and more than two states exist here, there is no guarantee that these four wells will be of the same depth. Here, the states of FU and D are mostly predominant, have about the same free energy values. Noticing that the barrier height between FU state and D state are not independent of the choice of reaction coordinate ($\sim 2kT$ in the upper line and $\sim 3kT$ in the lower line), we suggest that 1 dimensional free energy profile should not be used to estimate transition barrier heights.

In two dimensional free energy profile (Fig. 4.2), we use one dimension to capture folding (x-axis: $Q_a/2+Q_b/2$) and the other dimension to describe binding (y-axis: Q_i). Because the unbound states have 0 inter-subunit contacts, 2F, 2U and FU are all located at the boundary of the figure (on the surface of $Q_i=0$, which gives the blue ribbon). We can see that the barrier between monomer folding is about $1\sim 2kT$, while the barrier height between state D and state FU is much higher ($6\sim 7kT$). More surprisingly, on this landscape, the lowest saddle point to reach the native dimer state (D) is from the state of one folded monomer and one unfolded monomer (FU). The barrier between 2F state and D state is higher than that. We do observe a dimerization tendency from the folded monomers (2F state), as demonstrated by the small indent on the edge of the 2F well, which indicates the formation of a few inter-subunit contacts. But this indent doesn't reach the dimer well. This interesting feature suggests there may be steric clashes between two monomers when they try to bind each other in folded conformation. One possible source of this clash can come from PHE58 located on the extend third beta strand. Experimental studies suggest that this interface residue PHE58 in Cro dimer, may fold back to the same monomer and stabilize the monomer core [73, 74].

In three dimensional free energy landscape (Fig. 4.3), the transitions among these four states become even more clear. Cro repressor demonstrates a very special coupled binding-folding pathway. Although it is a classical 3-state dimer, or often equivalently called non-obligatory dimer, its binding is not about two folded monomers coming together as intuition may tell us. Instead, in both our thermodynamic and kinetic simulations, the binding transition process avoids (does not start

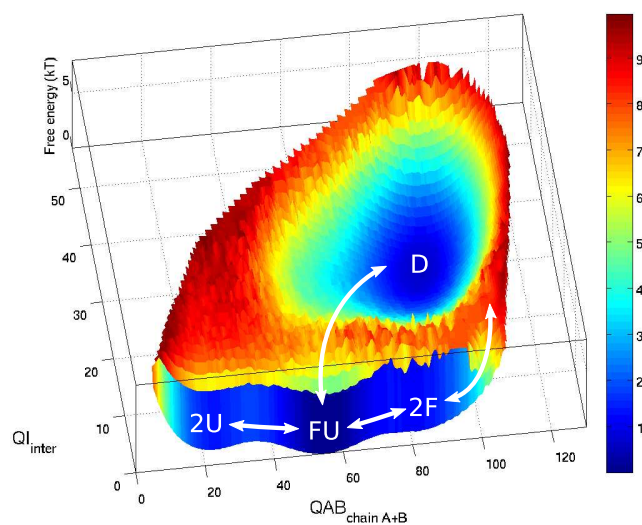


Figure 4.2: 2D free energy

from) two well folded monomers. That is to say, the coexistence of two folded monomers (2F) is an off-pathway intermediate state for association process. To start binding, one of the monomer need to be completely unfolded while the other relatively well folded. As demonstrated in the structural evolution below, the relatively well folded monomer (called A) has an one-side exposed hydrophobic core, which serves as the target of the extended “sticky” fly-casting arm (containing residue PHE58) from its unfolded partner (called B), that completes the partially formed hydrophobic core of A. When this happens (when the second unfolded monomer is casted onto the first folded one, and on its way to complete its partner’s core as well as the interface), the competition from the first monomer (A) makes the second one (B) to sacrifice some of its intra-contacts temporarily for the formation of the binding interface. After that, with the interface formed and “sticky” arms from the first monomer (A) right around, the second monomer (B) readily gets folded.

4.2 Structural evolution in dimerization

We will demonstrate the structural evolution in 5 stages described below. Each stage is labeled in 3D free energy landscape (Fig. 4.3). The red line connecting red dots is the minimal free energy path (saddle line). We can see the transition barrier for the individual monomer folding is relatively low (between stage 0 and stage 1, height 1~2 kT). While the coupled folding-binding transition from stage 1 to stage 4 has a higher barrier of ~5kT, and thus is the speed limit step in Cro dimerization. For each stage, a representative structure frame is picked, and the average contact formation probability map is built. The color of the dot encodes the probability that there is a native contact formed between two residues. Key residues which give high probability spots on the map are displayed in the structure, and colored according to there chemical property (while-hydrophobic, green-polar, blue-basic, and red-acidic). Secondary structure are labeled as well (H for alpha helix, B for beta strand).

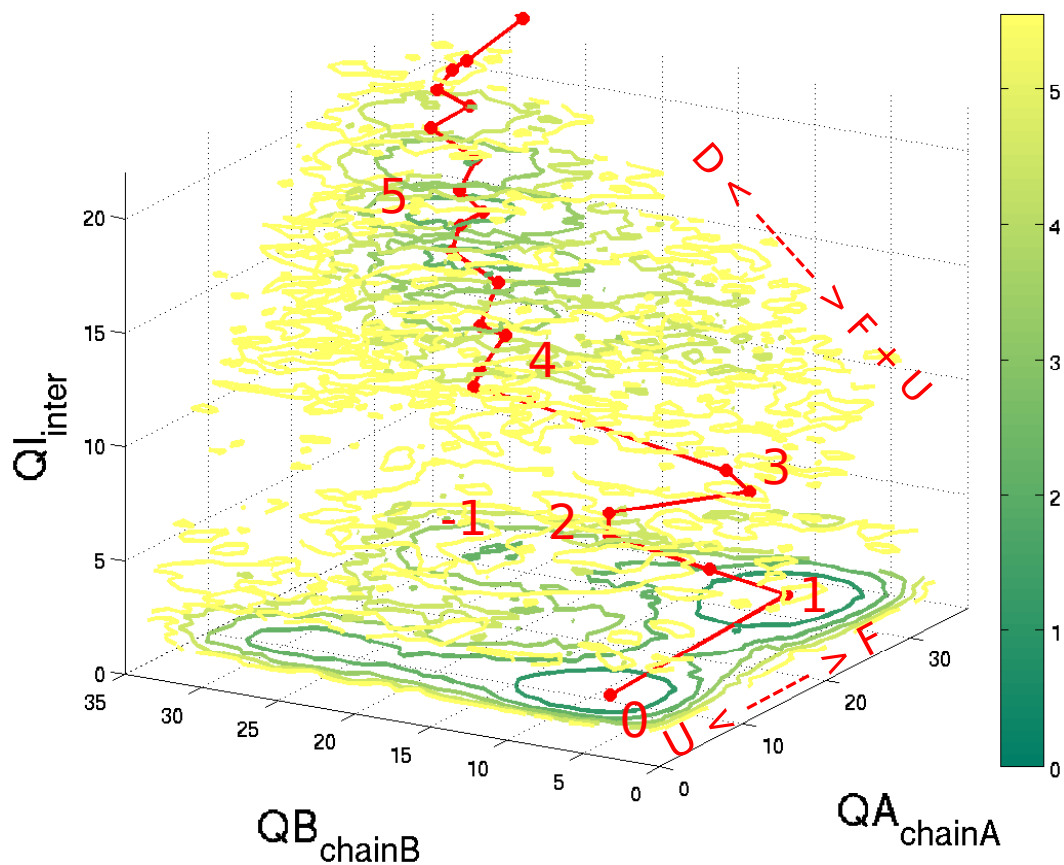


Figure 4.3: 3D landscape of 1cop

Stage (-1), Fig. 4.4. The off pathway intermediate state of two folded monomers. We say this is an off pathway intermediate state because the free energy barrier from this state to the native dimer state is much higher than the barrier from the state of one folded monomer and one unfolded monomer. This off-pathway intermediate state is actually corresponding to a trap on the funneled landscape of dimerization.

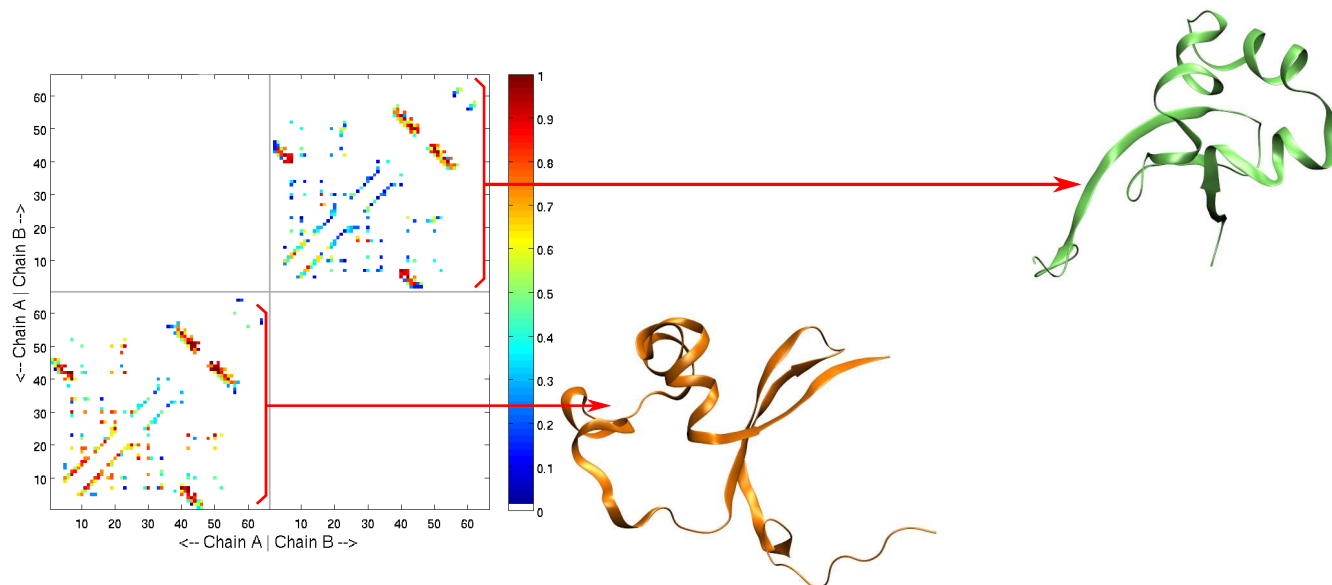


Figure 4.4: Stage (-1). Off-pathway thermodynamic trap, with two folded monomers. Monomer A is colored orange; and monomer B is colored lime. The contact map gives well developed intra-contacts quadrants (bottom left for A, and top right for B), but blank inter-contact quadrants.

Stage (0), Fig. 4.5. Before folding-binding transition starts, neighboring residues meet and form temporary contacts. Among these temporary contacts, residue ASN45, SER49, ILE44 VAL50, THR43, TYR51, LYS8, MET12, PHE14 and LYS18 are relatively more actively involved, indicated by the warm colored spots they have on the heat map, and illustrated in the snapshot beside the map. These regions involve the turn between $\beta 2$ and $\beta 3$, the middle part of $\alpha 1$ and interface between $\alpha 1$ and $\alpha 2$.

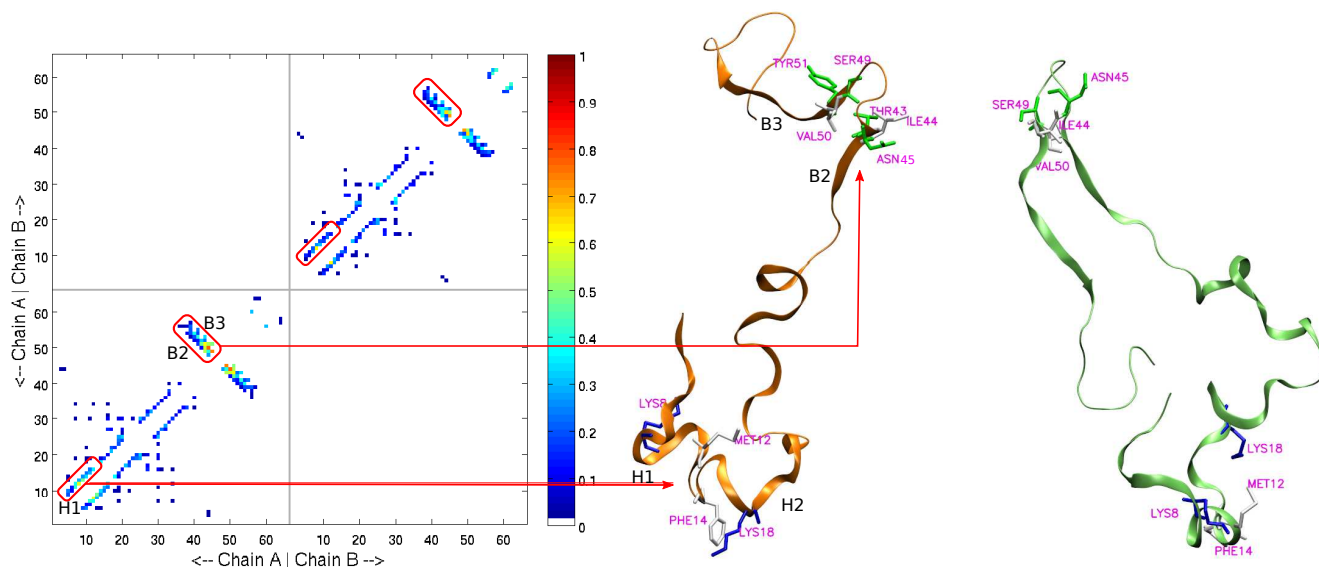


Figure 4.5: Stage (0). Temporary local contacts. Key residues are displayed in Licorice form.

Stage 1, Fig. 4.6. One monomer (A) gets folded with all major components formed (all three helices and all three beta strands), while the other monomer is still largely unstructured. In the folded monomer, very stable (strong, with dark red spots on the contact map) contacts are made among residue LEU7, TYR10, THR19, ILE30, ILE40, PHE41, LEU42, THR43, ILE44, VAL50, and TYR51 (all these residues are shown in the figure in Licorice drawing form, but not all labeled). Most of these stabilizing interactions are hydrophobic packing of the core.

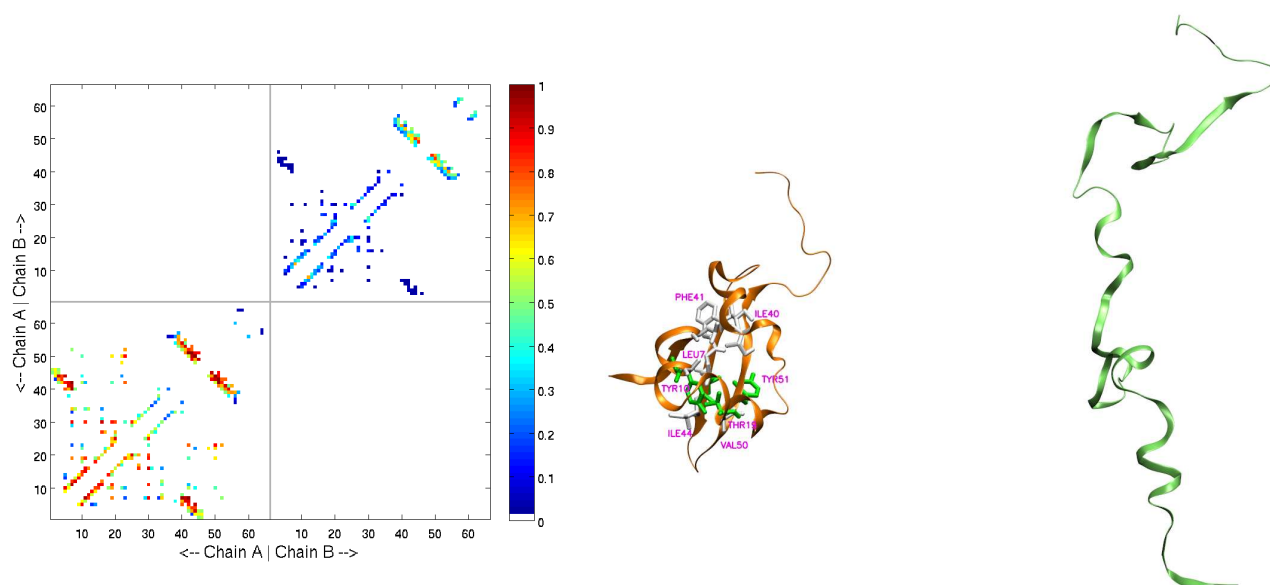


Figure 4.6: Stage (1). One monomer has to get unfolded to start binding.

Stage 2, Fig. 4.7. Inter-subunit contacts begin to form between beta strand 3 of the C terminus. Among these contacts, as can be seen from the map, PHE58 (of monomer B) interacts with several residues (LEU23, ALA29, ILE40, TYR51, ALA52 and GLU53) at various locations on one side of the other monomer (A), demonstrating its important participation in the hydrophobic core of monomer (A) and the binding interface. Its side-chain interactions with LEU23, ALA29, ILE40, and ALA52 complete the hydrophobic core in monomer A; and its backbone interactions with TYR51, ALA52 and GLU53 contribute to the binding interface between two beta strand 3 from two C terminus. The second monomer (B) also forms some temporary intra-contacts within itself, which get sacrificed soon when more of the binding interface develops. These temporary intra-contacts exist between alpha helix 1 and 2 (residues LEU7, LYS8, ALA11, MET12 and GLN16), and between beta strand 2 and 3 (residues ASN45, SER49, ILE44, VAL50, THR43, and TYR51).

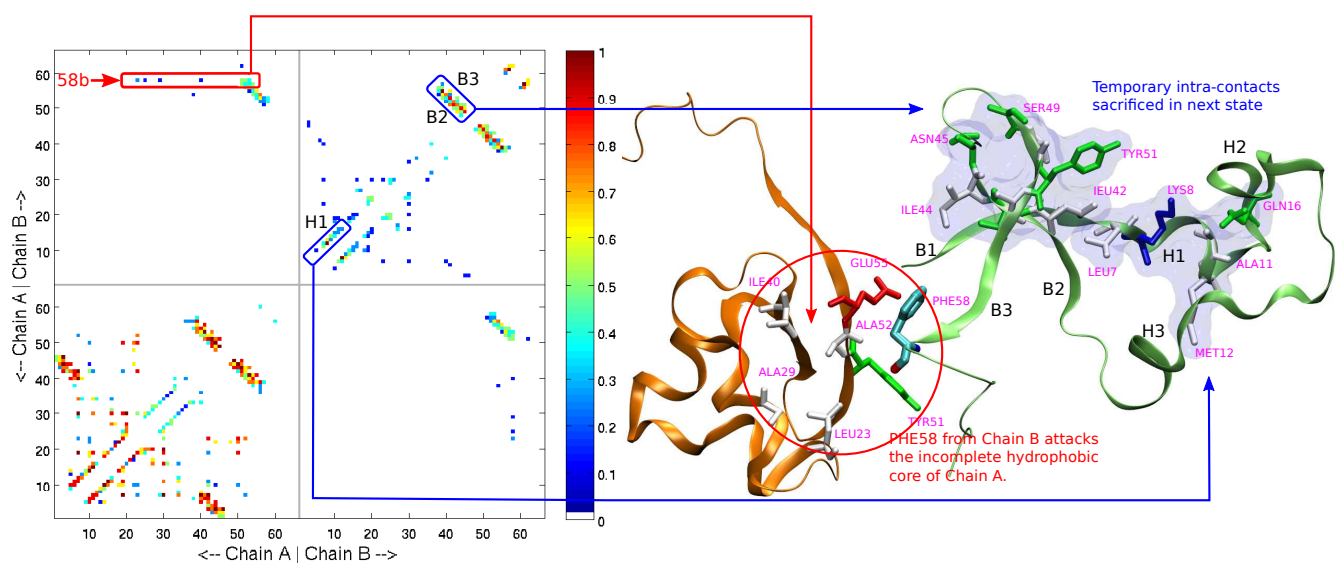


Figure 4.7: Stage (2). Fly-casting of the second monomer. PHE58 from Chain B attacks the incomplete hydrophobic core of Chain A.

The important role of PHE58 is more about initiating binding process rather than contributing to the folding of the hydrophobic monomer core, since it is located on the surface layer of the monomer core, as a patch to cover an exposed hydrophobic region on monomer A. Also, we find instances with two folded but unbound monomers, suggesting PHE58 from its partner is not essential for one monomer to fold itself, and thus thermodynamically 1cop is a non-obligatory dimer (3 state dimer). However, from minimal free energy path, and our constant temperature kinetic simulations, single folded monomer binds to its unfolded partner at C terminus is widely observed. Based on these two observations, we hypothesize that the one unfolded monomer is required when binding occurs in 1cop. That is to say, although independent folded monomers generally exist, they do not necessarily bind with each other in folded form, due to the harsh requirement of precise docking of well structured interface and smaller partner searching diameter. However, if one of the monomer gives up folded structure and becomes extended coil, fly casting

can readily occur, with its PHE58 as a “sticky” arm targeting at the partially exposed hydrophobic core of the other already folded monomer.

This hypothesis questions the conventional suggestion/assumption that a three-states dimer possesses a prerequisite intermediate state of folded monomers in its dimerization process. Being three states dimer only tells us that there is a third local minimum well ($2F$ or I_2), other than the unstructured monomers ($2U$) or structured dimer (D), on the free energy landscape, but does not and should not infer that two folded monomers are required for binding to occur. Here, in our REMD and constant temperature kinetic simulations of Cro repressor, the combination of one folded monomer with another unfolded monomer is clearly preferred when binding occurs.

Stage (3), Fig. 4.8. The binding interface comes into being. We can see a rather native like binding interface, composed by the two anti-parallel beta strand 3 (residues: TYR51, GLU53, GLU54, LYS56 and PHE58 in monomer A; and GLU53, GLU54, LYS56, PRO57 and PHE58 in monomer B). The intra-contacts of monomer B get sacrificed when they compete with the formation of binding interface, i.e. the intra-contacts between beta strand 3 and strand 2 (of monomer B) get lost when beta strand 3 from the other monomer (A) competes for the same beta strand 3 from the other side. The closing-up of PHE58 from the folded monomer A may also disturb the temporary contacts in partially folded monomer B.

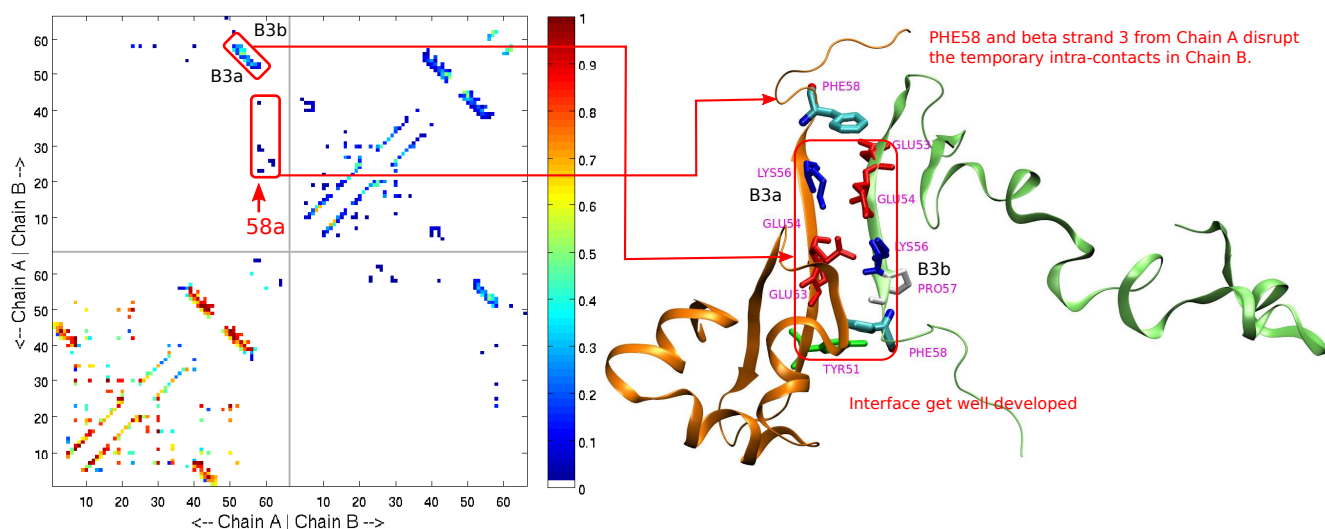


Figure 4.8: Stage (3). Binding interface formed.

Stage (4), Fig. 4.9. Second monomer folds and binding interface gets fully developed. We also get the phi-value map of contact pairs for this transition (binding coupled with second monomer folding, Fig. 4.11 lower), from which we identified the key residues with high phi-values: LEU7, ILE40, PHE41, LEU42, THR43, ILE44, GLU53 and GLU54 of monomer B; and GLU54 and LYS56 for monomer A.

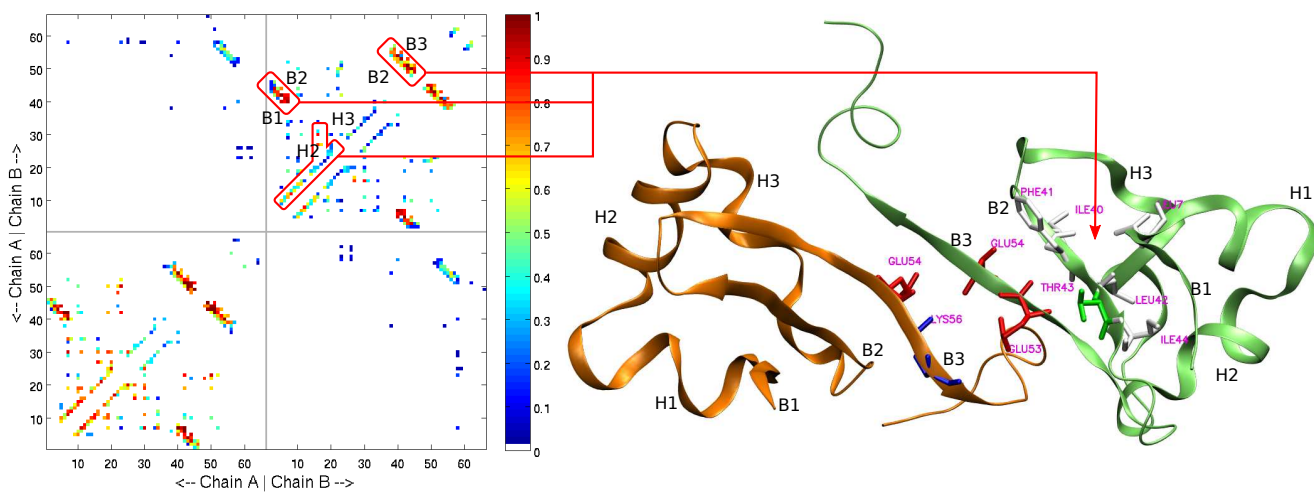


Figure 4.9: Stage (4). Second monomer folds.

Stage (5), Fig. 4.10. Gradual refinement of binding interface and packing of secondary structures.

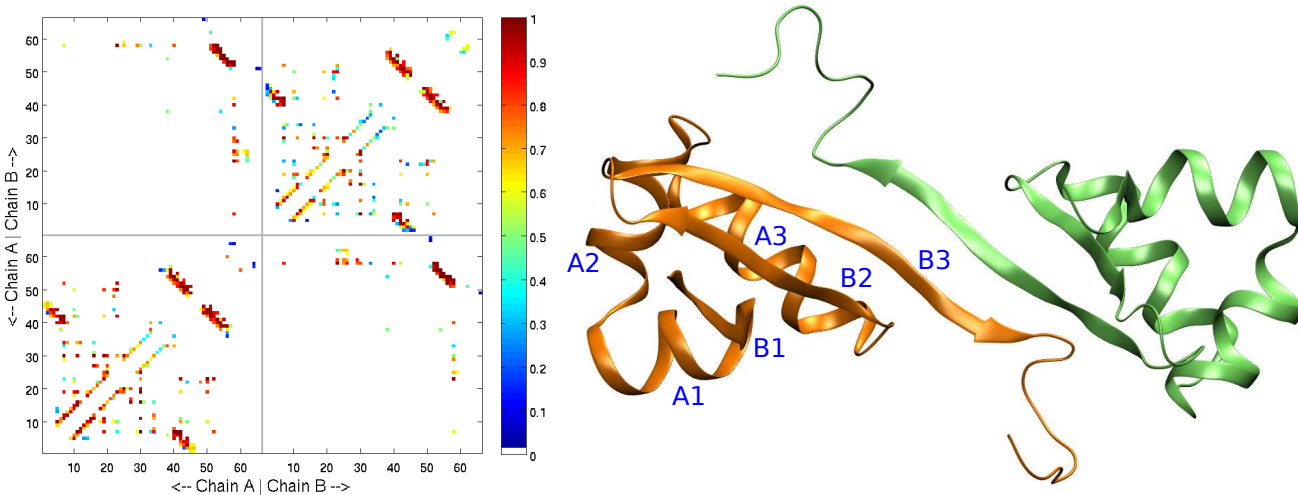


Figure 4.10: Detail refinement

4.3 Phi-values for two transitions

Phi-value analysis is applied to characterize the two transition states in Cro repressor dimerization (Fig. 4.11). The phi-value of a residue reveals how likely that residue obtains its native interactions at transition state [22]. In the first transition from stage 0 to stage 1, the high phi-value residues stay within the first monomer itself, and primarily locate on its three-strands beta sheet. Also, several intra-monomer contacts of medium phi-value are observed among the three helices (the scattered light blue/green spots on the map, with phi-values around 0.5). One representative pair of them is between residue THR19 in the 2nd helix and ILE30 in the 3rd alpha helix.

The second transition has high phi-value residues primarily located on the binding interface between beta strand 3 from each monomer (Figure 4.11 bottom). Also, another two relatively high phi-value spots on the contact phi-value map is between beta strand 1 and 2, and beta strand 2 and 3 within monomer B. Residue THR19 and ILE30 in the 2nd and 3rd alpha helix, which likely have their native interactions formed in the first monomer folding transition, are not that significant any more in the second transition. The map region for interactions among helices get darker (lower formation probability) and sparser (less in number). However, the interface residues (GLU54, VAL55 and LYS56) keep high phi-values during the second transition, which agrees with the 3D energy landscape (i.e. we can see that the second transition from stage 1 to stage 4 on the landscape, involves the development of the binding interface contacts, and thus is very different from the first pure folding transition from stage 0 to stage 1).

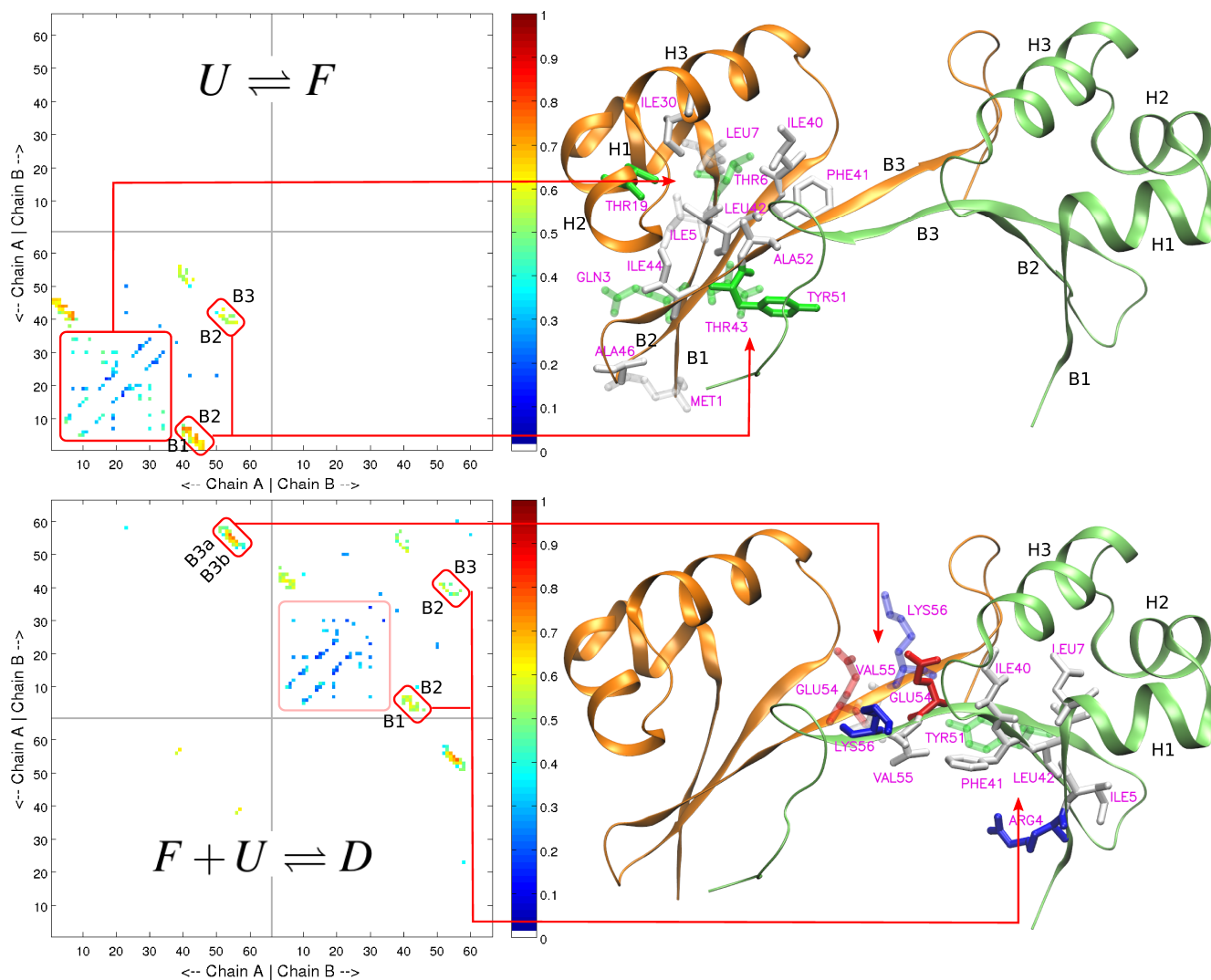


Figure 4.11: 1cop phi-values. The light red colored square in the second map represents attenuated contact phi-values among the alpha helices of monomer B in the second transition.

4.4 Experimental evidences

The hydrophobic core in each subunit is composed by three alpha helix and three anti-parallel beta strands, one of which is beta strand 3 from the other subunit. Phe58 on the third beta strand has been identified as essential component for the packing of the hydrophobic core and contributes significantly to the core's stability [81, 82, 83, 78, 76]. This is likely the reason that dimer dissociation and monomer denaturation are found at the same time when temperature is increased to 45 °C [80, 70]. Remarkably, this concurrent/syn-chronic behavior is well observed in our simulation. There is only one peak in heat capacity curve of 1cop at transition temperature, but the free energy landscape shows two distinct transitions (one for monomer folding, and one for coupled binding-folding of the second monomer).

Multidimensional NMR had shown extensive inter-subunit hydrogen bonds between the C terminus of beta strand-3 (residues GLU53 to PRO57) [69]. We non-ambiguously found that the corresponding region gives high contact formation probability as well as high phi-values. (Fig. 4.8 and 4.11). The same feature is also captured along the structural evolution along minimal free energy pathway, from which we observe the contact between C terminus starts the binding process (Fig. 4.7). This suggests that all-atom structure based model used here well preserves the hydrogen bonding interactions among residues, due to its finer grained topology, which is capable of creating configuration-based inter-atomic contacts, like hydrogen bonds here. This is a fundamental advantage over residue level coarse-grained structure-based model. Also, improvement in characterizing hydrophobic packing can be expected for the same reason. In our study of another dimer (unpublished yet), the all-atom structure-based model is even able to distinguish regular spacial non-bonded contacts from covalent disulfide (-S-S-) bonds.

Given the integrated monomer hydrophobic core, with Phe58 from its partner, and the extensive inter-subunit hydrogen bond networks between the C terminus (beta strand-3, residues GLU53 to PRO57), the fly-casting behavior observed for the second monomer become more reasonable. The C terminus from one monomer is essential for the other monomer to fold into a stable hydrophobic core, and at the same time serves as a sticky arm of fly-casting for dimerization. Recent theoretical study shows that the fly-casting is more effective when a protein possesses a small folding barrier and relative rigid extended part towards its target [84]. A relative rigid triad of PRO57-PHE58-PRO59 [74] is located in the C terminal of Cro repressor, which meets the extended rigidity requirement suggested in theoretical study. And the monomer folding barrier observed here is low (1~2kT), which indicates the folding and unfolding transitions of a monomer occur fairly often. The prerequisite of both folded and unfolded monomers at the same time, can also alternatively explain the rather high micromolar dissociation constant and the slowness of Cro dimerization [76, 75, 73].

In summary, we proposed a new dimerization mechanism based on our simulation on lambda Cro repressor. For the first time, this mechanism reveals the importance of unfolded monomer (U) in the dimerization process of a non-obligatory dimer. Cro dimerization is observed as an half coupled binding-folding transition between folded dimer state (D) and one hybrid state of one folded monomer and one unfolded monomer (F+U). Two transitions among four state are involved. The first is a low barrier monomer folding transition $U \rightleftharpoons F$; the second is a high barrier half coupled binding-folding transition $F + U \rightleftharpoons D$, which can be well characterized as a fly-casting approach. Conflicting experimental results about whether Cro repressor is 2 state dimer or 3 state dimer is reconciled in this way.

This new possibility questions the clarity of current terms in describing dimerization mechanisms. 3 state dimer actually often have 4 populated states (D, 2F, 2U, FU), with the last one of them commonly ignored. When a 3-state dimer (often equivalent to non-obligatory dimer) possesses a dimeric intermediate state (I_2) instead of a folded monomeric state (2F), it stops being a non-obligatory dimer, because being "non-obligatory" requires the existence of structured but unbound monomers (2F). Moreover, base on these thermodynamic terms (3-states or 2-states, obligatory or non-obligatory), assumptions of kinetic process should not be made. Being 3-state dimer does not necessarily suggest decoupled binding-folding; Being 2-state dimer does not suggest coupled binding-folding either. More precision, caution and pondering are needed in describing dimerization mechanisms as we discover and understand more and more about protein dimerization in the future.

5 Diverse mechanisms of coupled binding-folding.

In this work, we applied all-atom structure-based model to study the coupled binding-folding mechanisms of 10 homodimers of distinct topologies [21]. These ten dimers have been classified before into two groups, 2 state dimers and 3 state dimers, according to whether populated intermediate states or pre-folded monomers exists or not in the transition between native dimer and unfolded monomers. In this study, we used a new all-atom structure-based model, which promisingly bridges the atomic structural resolution and long protein dynamics timescale [41, 51, 79]. We are now able to examine more details of the coupling mechanisms between dimer binding and monomer folding, and analyze their relationship with the underlying topologies. Five distinct dimerization mechanisms are observed from these 10 dimers, ranging from completely decoupled folding-then-binding strategy in 1flb, to perfectly coupled binding-folding of 1f36, complicated by non-uniform mix in 1arr, fly-casting of 1cop and 1lmb, and folding-first 3ssi. The three ideal coupling mechanisms between binding and folding are illustrated schematically by the thick bright-colored lines (Figure 5.1), together with our newly observed diverse coupling mechanisms (Figure 5.1, thinner gray lines).

With the finer resolution of all-atom structure-based model, we find that these generally classified two-state or three-state dimers actually display a much more diverse set of patterns in the way that folding and binding are coupled. These findings even question the conventional suggestion that being three-states dimers usually requires pre-folded monomers, as intermediate state, during association [22]. As we will see later, 1cop is a clear three-state dimer, with well-populated unbound pre-folded monomers existing before binding. Its association process is, however, not initiated from (does not go through) two folded monomers. On the contrary, one of the monomer has to be unfolded in order to start binding process. In other words, instead of being on-pathway intermediate state for association, the two-folded-monomers state is actually an off-pathway intermediate state for the dimerization process. Thus, being a three-state dimer does not necessarily infer that pre-folded monomers are obligatory/necessary for dimerization.

(Meanwhile, as will be seen in the case of 3ssi, being decoupled binding-folding dimer does not necessarily infer been 3 state dimer. In our simulation, binding and folding of 3ssi are decoupled from each other, but there is no stable/populated intermediate state during its transition. Based on these results, we suggest making better discrimination/distinction between thermodynamic and kinetic descriptions about dimerization. Thermodynamically, the number of populated/stable macro-states observed determines whether a dimer is 2 state dimer, 3 state dimer, or even higher order multi-states dimer. Kinetically, the correlation in the development of monomer folding and binding interface decides whether the folding and binding process are coupled or not. There is no certain equivalence between 2 state dimer and coupled binding-folding. Nor is there prerequisite for pre-folded monomers in the association of 3 state dimers. Actually, it should be 4 states dimers (i.e. 2U, FU, 2F and D).)

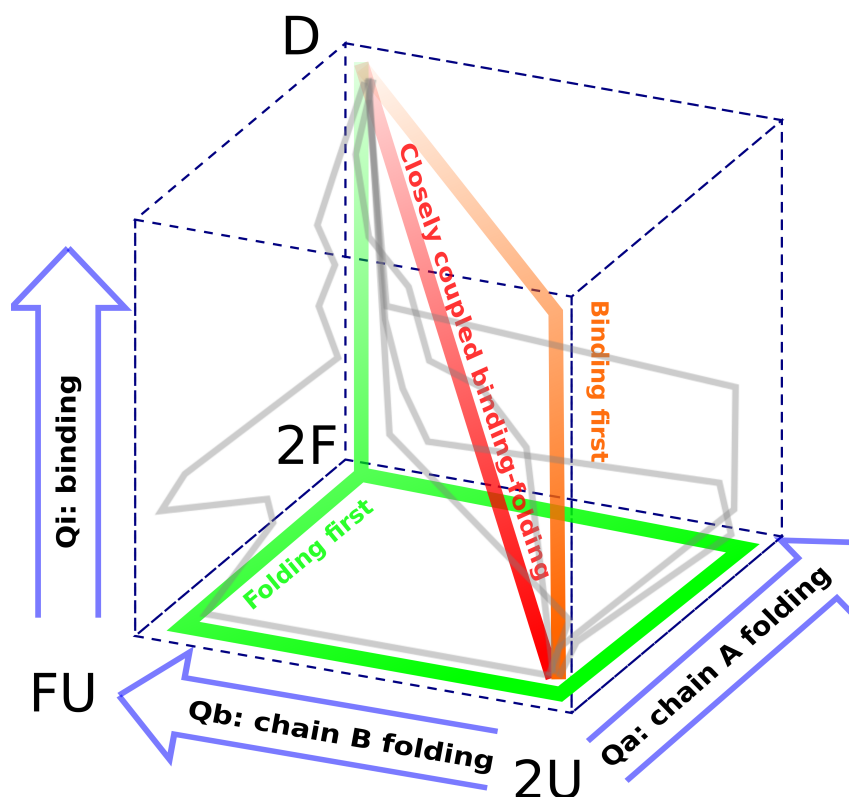


Figure 5.1: Schematic diagram of three ideal dimerization mechanisms. The axes measures the development of binding or folding by the number of native contacts formed on the interface (Q_i) or within the two monomers (Q_a or Q_b). Closely coupled binding-folding (cooperative development of inter and intra-monomer contacts) are illustrated as the red diagonal line. The green path along the edges of the cube is folding prior to binding scenario; while the orange path shooting up to the top is binding prior to folding scenario (though exist in vary rare cases, it do represent at least a possibility). The thinner gray lines represent the diverse binding-folding pathways revealed in this study.

Combining the minimal free energy path in the 3D free energy profile and the evolution of contact-pair formation maps along it, we are able to reveal specific and detailed contacts evolution in the diverse binding-folding transition processes of dimers. Representative structures are chosen along the minimal free energy path in order to illustrate the corresponding structural development. Newly formed contacting residues at each state, especially those of high probabilities, are shown in Licorice drawing in the structure figures and labeled. (coloring scheme: backbone of monomer A is colored orange; backbone of monomer B lime/green; key residues are colored according to their chemical properties: red for the acid, blue for the basic, green for the polar and white for the non-polar. Some residues in the back are made transparent to increase distance contrast. Some times, if imaging key residues (the same residue of the two monomers) exist on both monomers, only one of them is labeled in order to avoid over crowdedness, and often the other one is made transparent.)

It's worth noting that the minimal free energy path is always slightly away from the exact diagonal, suggesting the lost of entropy for two identical monomers to behave synchronously. This stray-away is also observed in other simulation results [22]. Also, since the free energy landscape for our homodimers are mostly symmetric about the diagonal, we restricted the red minimal free energy line to the right side of the diagonal.

Eventful/distinctive transition stages along the minimal free energy path are labeled from being random coils (stage 0) to being fully native state (stage 5 or 6). Corresponding residue-level contact pair maps are made for each labeled stage in the 3D free energy landscapes. Development of contacts are circled on these maps and matched to the structures beside them. Red colored arrows are used for contacts formed in that stage; and blue color arrows are used for contacts that disappear in that stage or that will disappear in the next stage (detailed by case later). Note that this step-wise description may not be corresponding to the real temporal order of structure development, but very likely is true if we assume the development of native contacts is gradual or continuous. (Also, by scrutinizing the many kinetic folding trajectories we collected around transition temperatures, these step-wise progressions illustrated below are generally followed in kinetic binding-folding transitions, with certain back and forth kinetic fluctuations.)

5.1 Inversion stimulation factor (1f36)

For inversion stimulation factor (1f36), which is a very entangled/inter-winded dimer with bundles composed of helices from both monomers, folding and binding are well coupled all the way from unstructured monomers to well structured dimer (Fig. 5.2). Only native dimer and unfolded monomers are presented in this ideal two-state transition ($D \leftrightarrow 2U$). The structural entangling of this dimer is so complete that equal numbers of native contacts are found in one monomer and on the interface ($Q_i = Q_a = Q_b = 169$), calculated by shadow map method described in the method part. Even though, it is still amazing to find out how closely, except for small fluctuations, the minimal free energy path traces the diagonal of the 3D free energy cube, which represents uniformly coupled binding-folding behavior. The same suggestion has also been made from urea-induced equilibrium denaturation experiment monitored by circular dichroism and tyrosine fluorescence [85]. Previous computation work with residue-level course-grained structure-based model has suggested that a dimer's topology determines its binding mechanism [21, 22]. Here, the 1f36 serves as another good example.

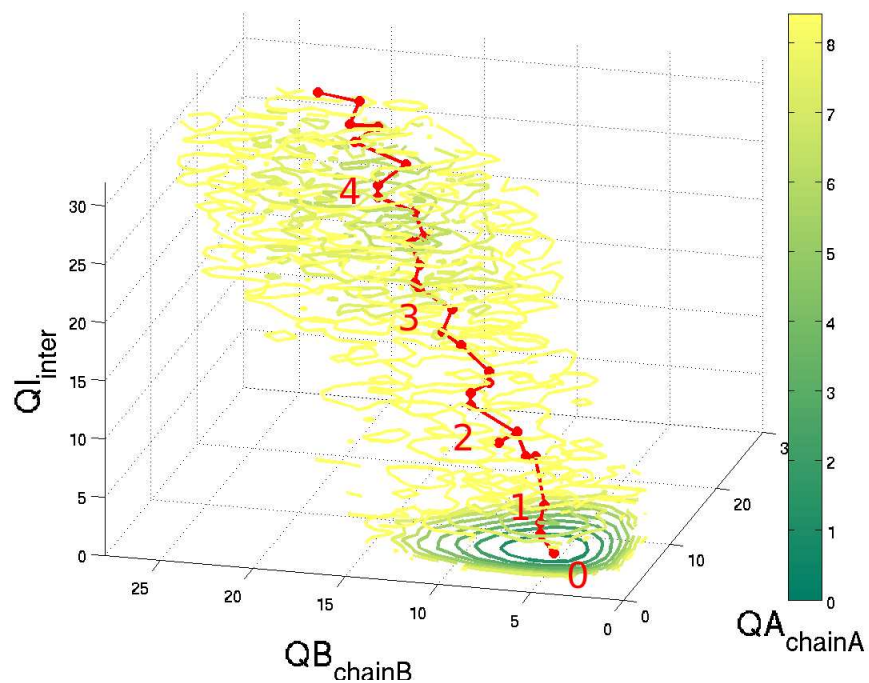


Figure 5.2: 3D landscape of 1f36

Stage (0), Fig. 5.3. Before binding starts, helix 3, helix 4 and C terminal (second half) of helix 2 have a tendency to cluster together (LEU55, GLU59, MET67, ALA77 and LEU92), as can be seen from the short range non-local contact pairs formed on the map (figure 5.3). Also, high secondary structure tendency is observed for the hairpin head loop (VAL16, ANS17 and GLN21). These two regions are not involved in the dimer interface and thus could possibly be formed before binding starts.

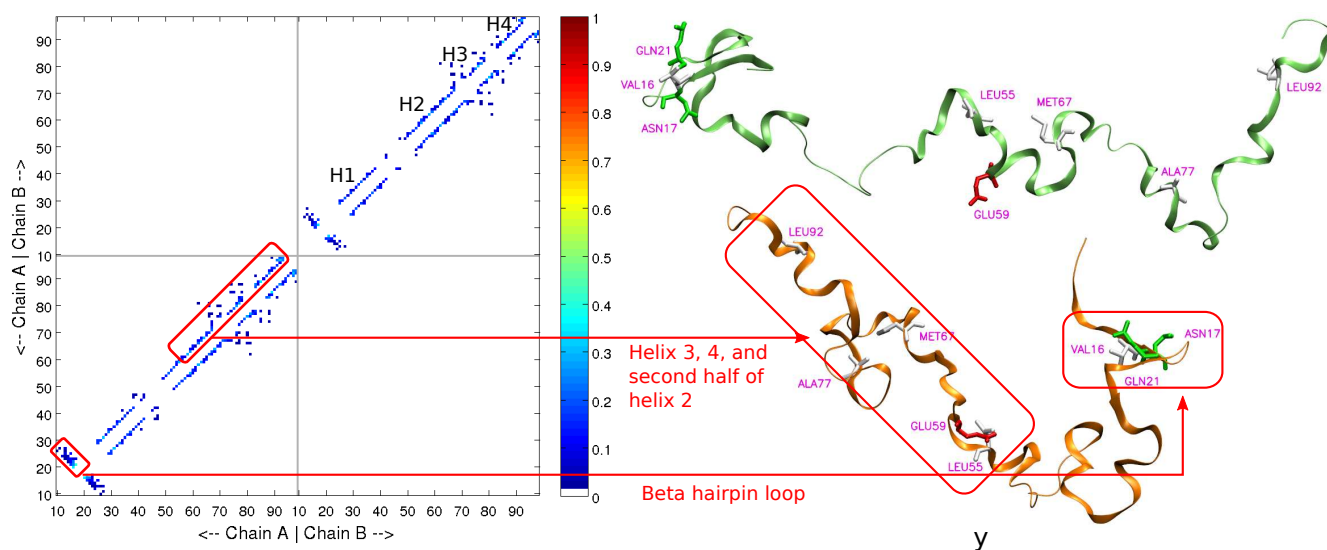


Figure 5.3: Secondary tendency

Stage (1), Fig. 5.4. The inter-monomer contacts starts to form, most of these contacts are between helix 1 of one monomer and helix 2 of its partner. These four helices compose the entangled hydrophobic core of the 1f36 dimer. The following residues are involved in the initial formation of binding interface: ARG28, VAL31, LEU35 in chain A helix 1; TYR51, LEU55, PRO61, MET65 in chain A helix 2; MET80 in chain A helix 3; and ARG28, VAL31, GLU36 in chain B helix 1; LEU50, LEU55, GLU57, PRO61, MET65 in chain B helix 2; and MET80 in chain B helix 3.

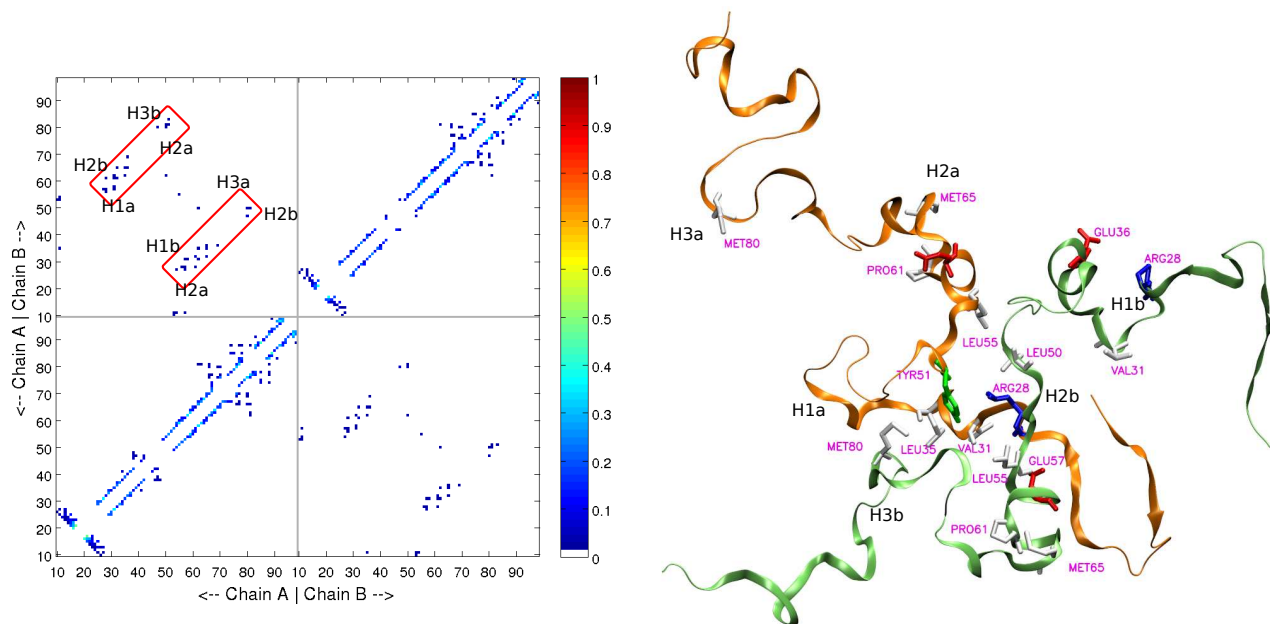


Figure 5.4: Dimer core binding. (H1a, H2b) denote Helix 1 from monomer A and Helix 2 from monomer B respectively.

Stage (2), Fig. 5.5. The foot of the beta hairpin forms with residue contact pairs LEU11-LEU35B and VAL13-GLN33B. Meanwhile the dimer core gets extended and strengthened. (Notice that residue 11 is actually located at the very left/bottom of the contact map, since the PDB structure of 1lmb starts with residue 10.). Also, the relatively independent module of alpha helix 3 and 4 gets into good shape in both monomers.

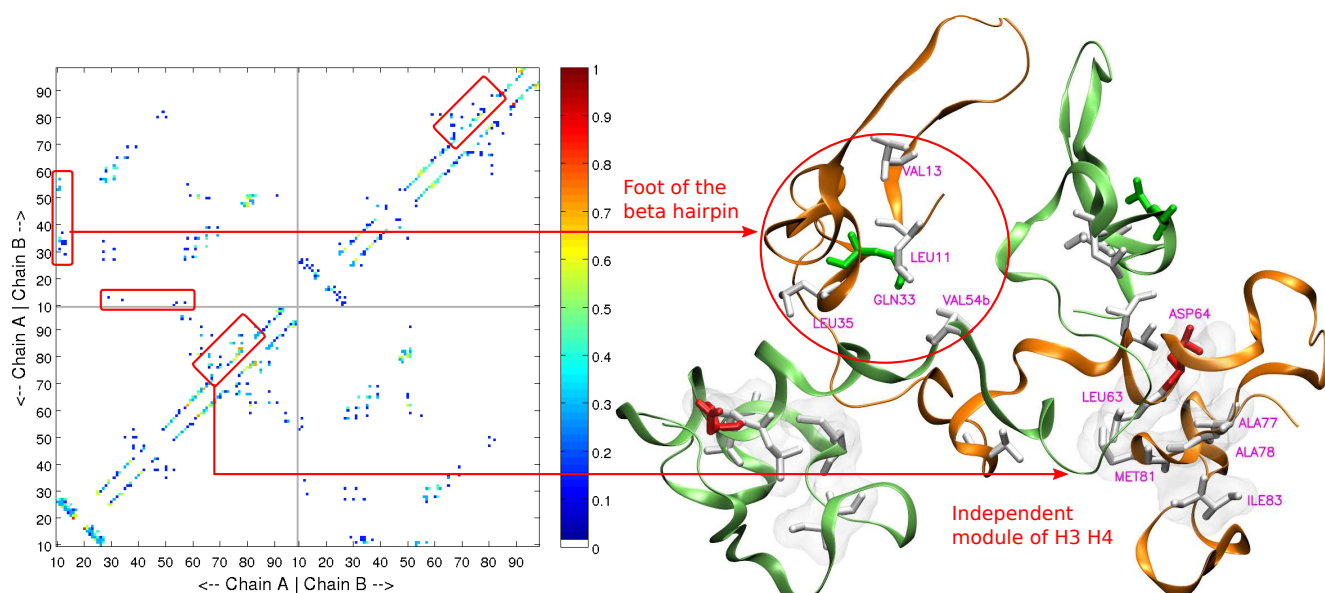


Figure 5.5: Hairpin foot/base formation; helix 3 and 4 module forms.

Stage (3), Fig. 5.6. The refinement of binding interface and local secondary structures follows. We notice that this happens slightly differently between two monomers. In monomer A, it is the contacts from helix 3 (e.g. ALA77, MET80) and the second half of helix 2 (e.g. GLU57, VAL58, GLN60, ASP64, VAL66) that get strengthened; however, in monomer B, it is helix 1 (e.g. LEU27B, ASP29B, VAL31B), the first half of helix 2 (e.g. LEU55B and ALA56B) and the turn between them (e.g. VAL47B) that get strengthened. This asymmetry may be presumably due to the entropy cost of being exact symmetric and the geometrical benefits (or energetic advantage) of being complimentary to each other. Local structure refinement may be required by and coincide with the improvement of binding interface.

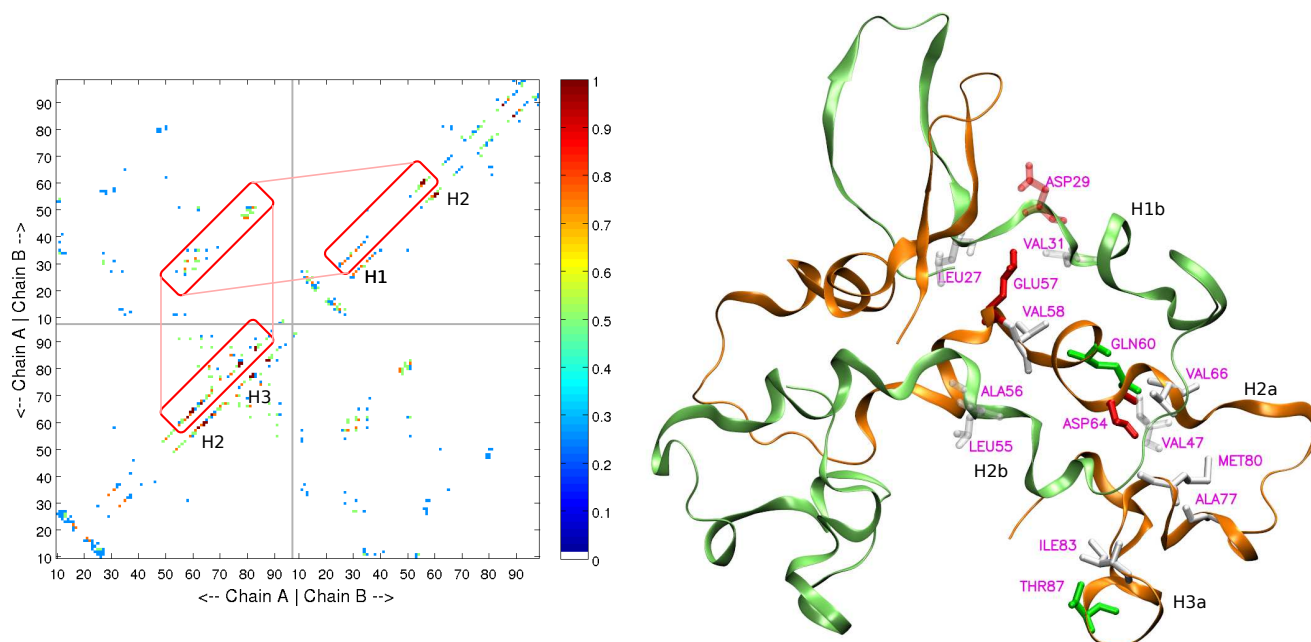


Figure 5.6: Binding interface get ordered; and asymmetric complementary local secondary structure refinement.

Stage (4), Fig. 5.7. A well-coupled refinement process of both secondary and tertiary structures emerges.

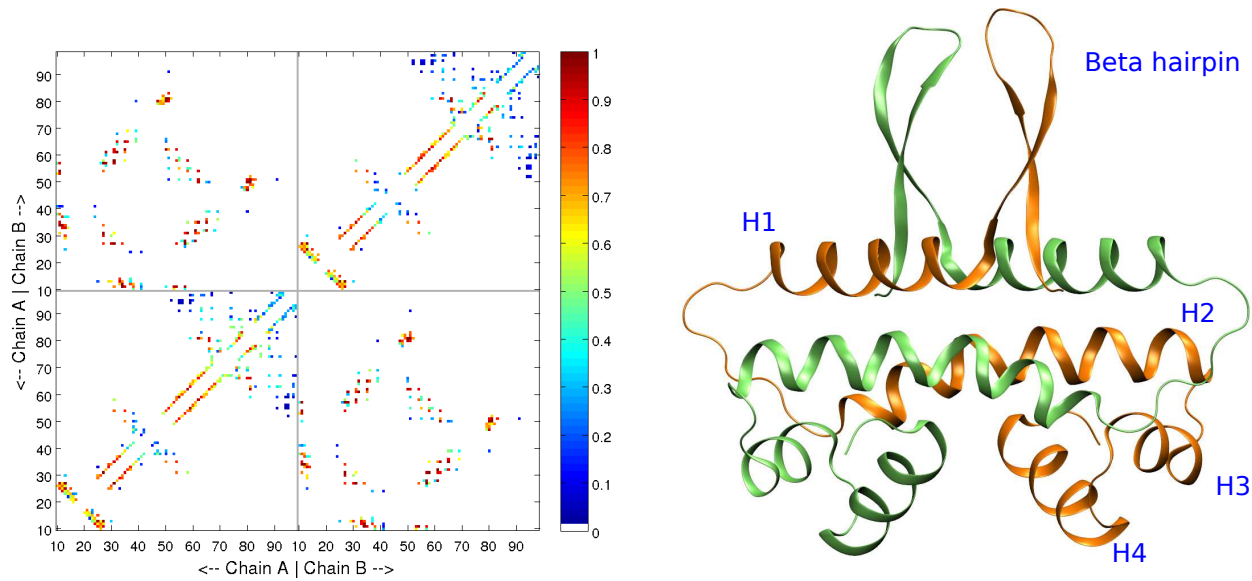


Figure 5.7: Well-coupled refinement of both secondary and tertiary structure.

Phi values: The high phi-value residues are mostly distributed in the entangled dimer core, which is mostly made up by the first helix and second alpha helices, especially surrounding the root of beta hair-pins (Fig. 5.8). These high phi-value residues are labeled in the structure (Residue LEU11, THR12, VAL13, LYS25 and LEU27 are at the foot of the hairpin; VAL31, LEU35 in helix 1; and VAL54, GLU 57, PRO61, LEU62 and MET65 in helix 2). Slight asymmetry between monomer A and monomer B is observed on the phi value map, which is likely due to the the rare sampling of transition state. MET81 is important for the relatively independent structural module, composed by helix 3 and helix 4.

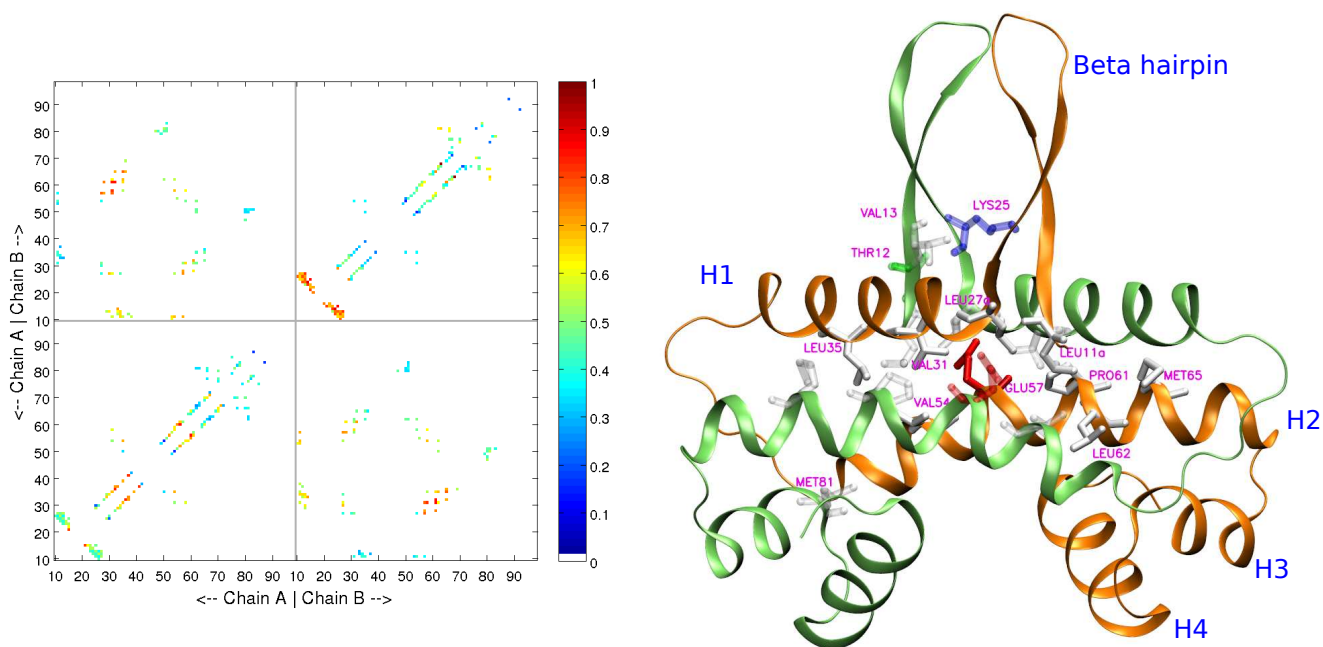


Figure 5.8: Key residues with high phi-values in 1f36

Experimental evidences: Suggestions were made by observing crystal structures, that the activation beta hairpin of inversion stimulation factor (1f36) has its roots stabilized by a group of hydrophobic residues. Its far-end hairpin head-loop is mobile but structured by hydrogen bonding network [86, 87]. We found good agreement with this suggestion in our simulations that the beta hairpins form at the early stage of the transition and stay structured most of the time. The hydrogen bond network of the hairpin loop (VAL16, ASN17, and GLN21 Fig. 5.3) forms very early and sustains itself with relatively good stability. Also, the supporting hydrophobic group around the hairpin feet (LEU11 LEU27, and from the other subunit ALA34 LEU35 TYR38 LEU53 VAL54) locks the hairpin to the protein core, producing high probability spots/bands on the contact map (Fig.5.5); Another modeling study also pointed out the importance of this hydrophobic core around LEU11 [88]. This inter-monomer hydrophobic core, serves to explain the perfect coupled binding-folding behavior of 1f36, as stated above.

Recent theoretical study shows that the fly-casting is more effective when a protein possesses a small folding barrier and relative rigid extended part towards its target [84]. Experiments have demonstrated that the N-terminal rigid hairpin head loop of 1f36 is responsible for contacting the recombinase and activating the DNA invertase [86, 87]. Given the mobility of the N-terminal hairpin and the relative rigidity of its head loop revealed in our simulation, as well as the existence of extended unstructured monomers with a low folding barrier, we hypothesize/suggest that the fly-casting approach is adopted by 1f36 in binding to its targets.

There are two other 2-state dimers which follow similar well coupled binding-folding scenario, Troponin C site III (1cta) and Trp repressor (2oz9). Both of their minimal path are symmetric about two monomers but somewhat smoothly off the diagonal. Trp repressor (2oz9) has its minimal path curve down, indicating a slightly more favored folding process in the beginning of the transition, leaving the binding process slightly more to the second half (as shown in Fig. 5.9a). While Troponin C site III, sharing a similarity with 1arr, has slight curving up in the first half of

the minimal free energy path (as shown in Fig 5.9b).

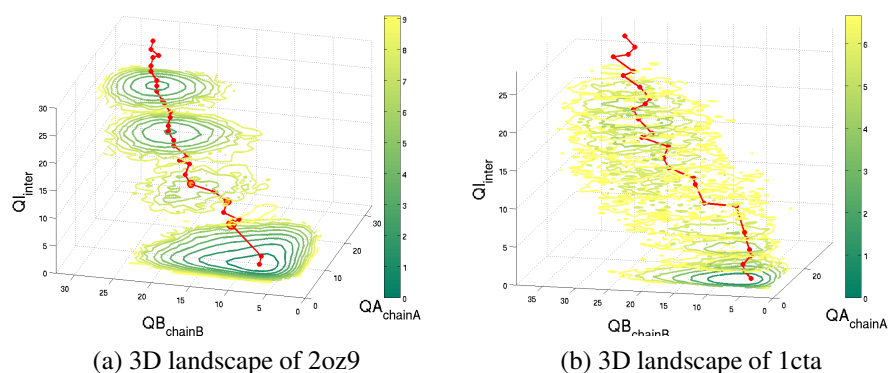


Figure 5.9: Two other closely coupled binding-folding dimers

5.2 Arc repressor (1arr)

From its free energy profile (Fig. 5.10), we notice that there is a decoupled binding process at the beginning of the transition (vertical segment of the minimal free energy path between stage 0 and 1), followed by a decoupled folding process (decoupled from binding, as the red minimal free energy line travels nearly horizontally from stage 1 to stage 2) in the first half of the binding-folding transition. While at the second half of transition (from stage 2 to stage 4), binding and folding are closely coupled. This suggests the whole binding-folding process is actually a twisted, and nonuniform mix of folding and binding. Coupled folding and binding only exits at the second half of the transition process, which is mostly about local structure refinement after the monomers and interface between them are relatively well formed. While the first half is composed of decoupled initiation binding followed by decoupled monomers folding. This step-wise separation of binding and folding shows a strong contrast with the behavior of other two state dimers, like 1f36, in which binding and folding are always closely coupled, as demonstrated above.

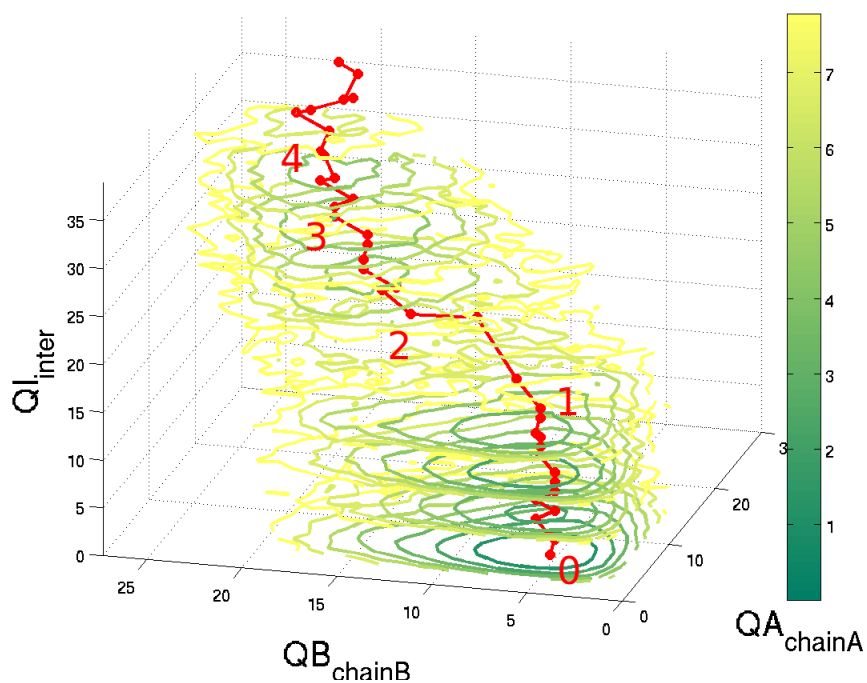


Figure 5.10: 3D landscape of Arc repressor (1arr)

Step-wise structural evolution along the minimal free energy path. Stage (0), Fig. 5.11. Being completely denatured random coils, local contacts of each chain fluctuate on and off, especially around residues ARG31, VAL33, GLU36 and LYS47. Most of these residues form contacts between the 1st and the 2nd alpha helices. Note here, when we decide which residue is the key one to be displayed and labeled, we not only consider its contact with the highest probability, but also how many contacts it makes with other residues. This is the reason that we only label a few residues from the region circled out on the map, and leave out a few isolated high probability spots (singular contact pairs) on the map.

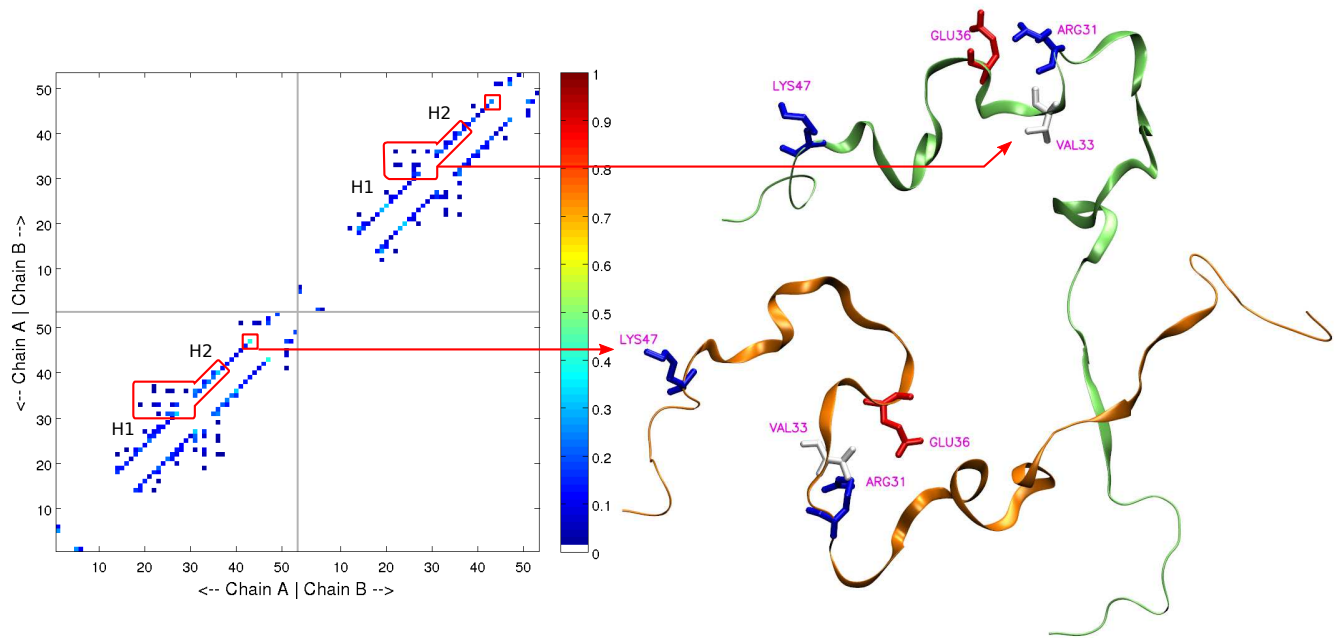


Figure 5.11: Local contacts before binding.

Stage (1), Fig. 5.12. Decoupled binding between the beta strands of the two monomers occurs first. Almost all residues on the anti-parallel beta strand interface have formed contacts. (Residues: LYS6, MET7, PRO8, GLN9, PHE10, ASN11, LEU12, ARG13 and TRP14. Only residues on monomer A is labeled; residues on monomer B can be identified by comparing its imaging residue labeled on monomer A.)

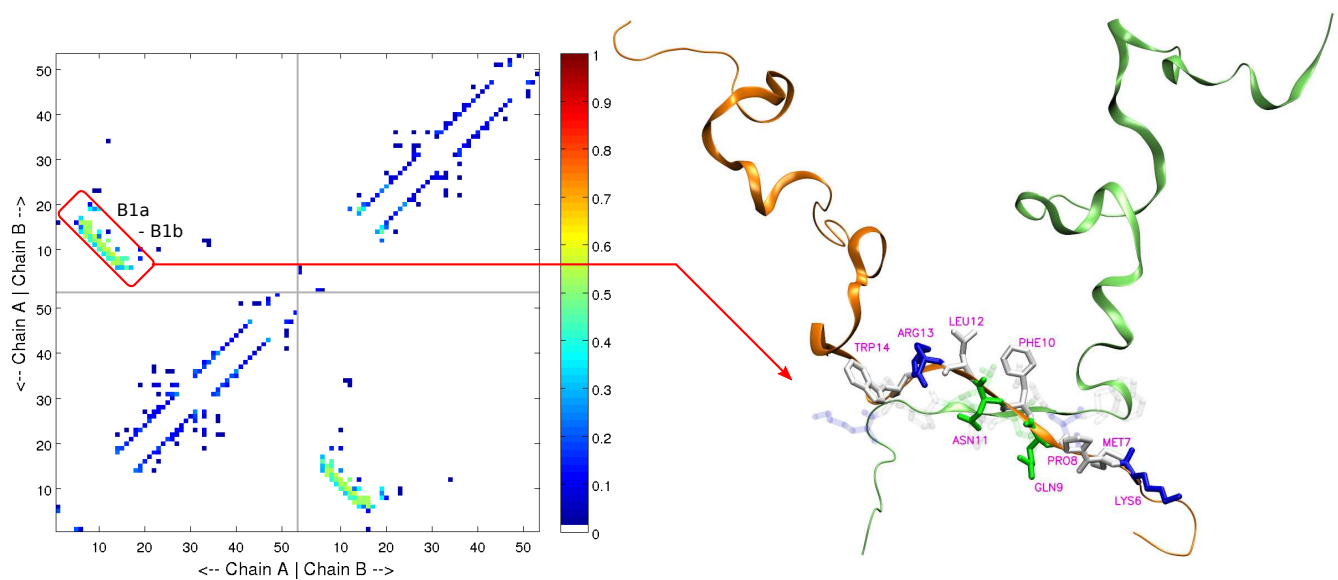


Figure 5.12: Beta strands starts binding process. Only residues in Chain A are labeled. Residues from Chain B are made transparent for clarity.

Stage (2), Fig. 5.13. Following the initial binding of beta strands, a relatively decoupled folding event happens in each individual monomer. The overall intra-contacts in alpha helix 1 and alpha helix 2 get strengthened especially at the first half of helix 1 (key residues: TRP14, LEU19 and ARG23), which is in the vicinity of the already formed beta strand interface. This decoupled folding process may prepare the dimer for the following stage of coupled binding-folding development. At the same time, the binding interface between beta strands become strengthened, as can be seen in the contact map that the color of the binding contacts zone has changed from blue/green (in stage 0) to orange/red, which means higher formation probability.

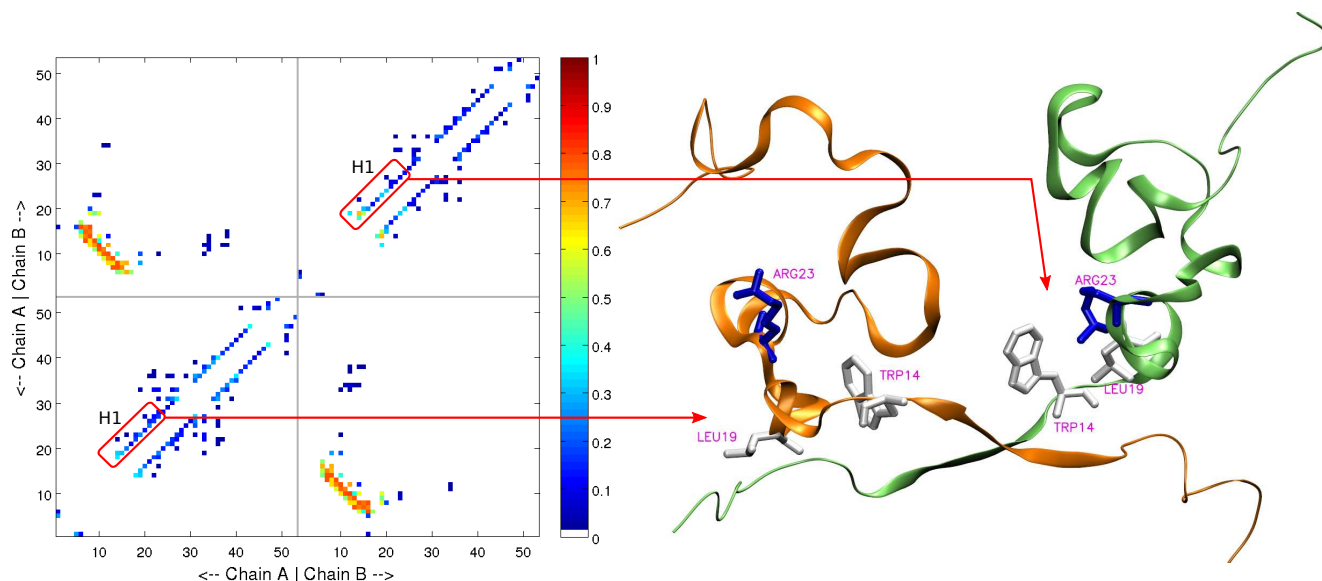


Figure 5.13: Decoupled folding event

Stage (3), Fig. 5.14. In this second half coupled binding-folding proceeds, more contact pairs developed both on the interface and within individual monomers. Their representatives are (VAL41-ARG40b, LEU12-ASN34b, "b" denotes chain B) and (VAL22-VAL33) respectively. The coils become more packed and the overall geometry of the native dimer appears. Thanks to the symmetry between two monomers, we only showed and labeled one copy of all these interactions in the structure to avoid crowdedness. Intra-monomer interaction in B is not shown; and for inter-monomer interactions, residues on monomer B is made transparent.

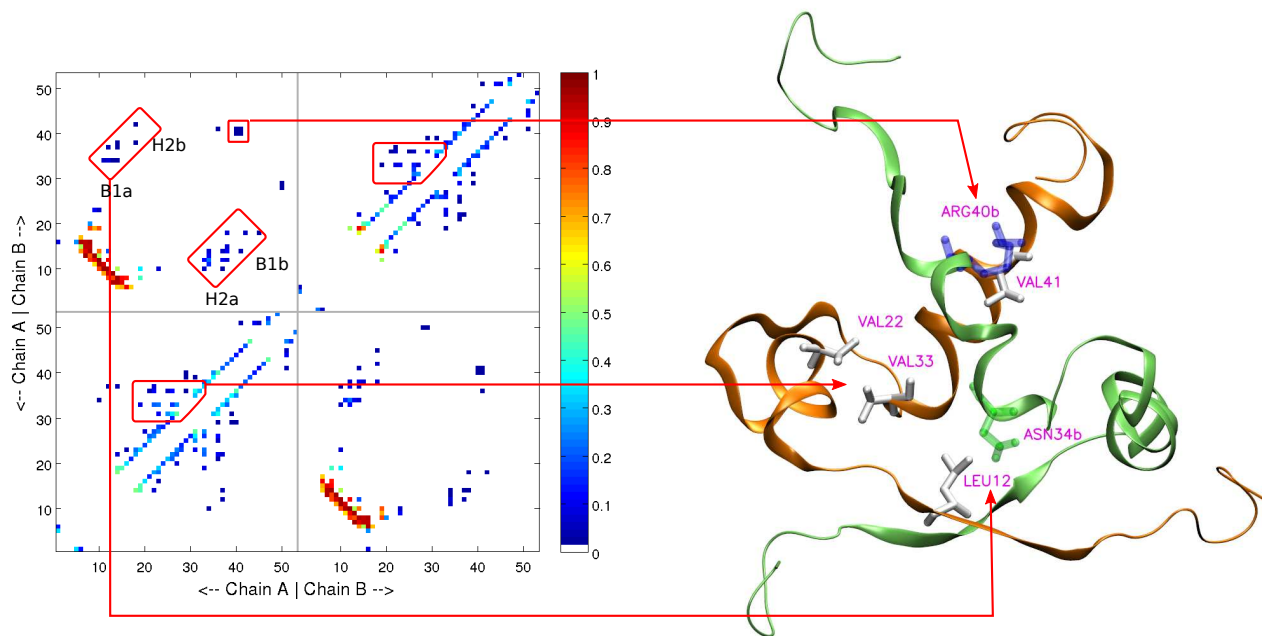


Figure 5.14: Coupled binding/folding in the second half

Stage (4), Fig. 5.15. Refinements continues until most native contacts are formed. The native structure from PDB is shown below with secondary structures labeled.

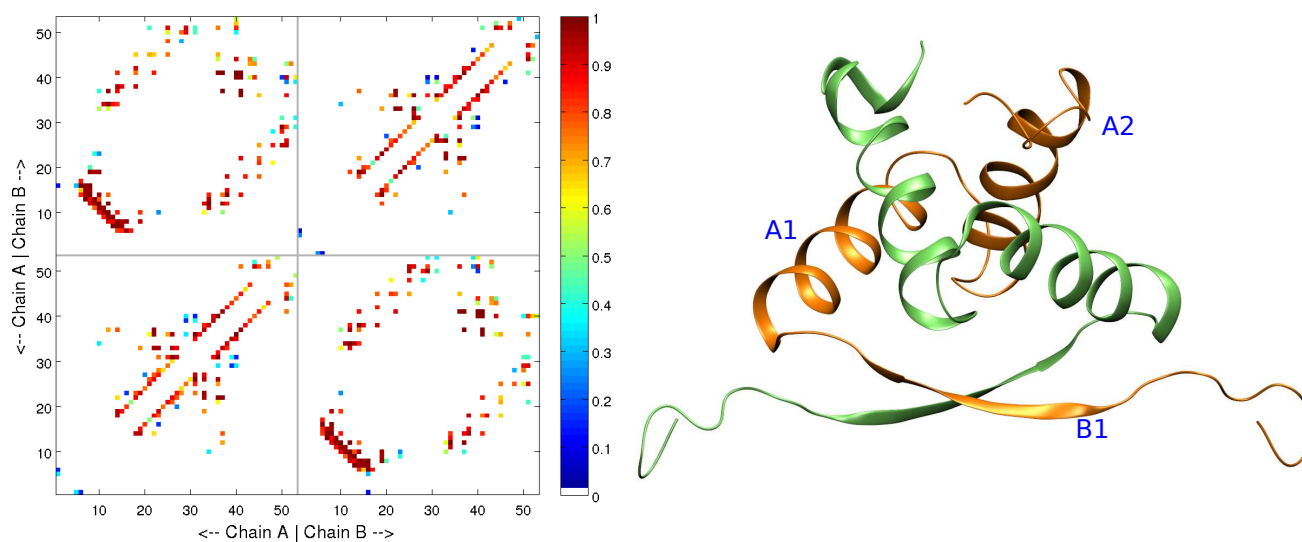


Figure 5.15: Well refined 1arr dimer.

Phi values: The distribution of high phi value residues are almost symmetric between two monomers (Fig. 5.16). From the phi value contact map and its mapping onto the native PDB structure, we can see all the identified high phi-value interactions are mostly non-polar residues. These residues

largely compose and surround the dimer hydrophobic core. This again suggests the hydrophobic interaction plays a similar dominant role in a 2-state coupled binding-folding transition, as in a single globular monomer protein folding transition. Apart from the beta sheet, high phi value residues have a tendency to distribute around the connections (turns) between secondary structure units, i.e. the end of beta strand, the end of helix 1 and the first half of helix 2. Note that almost all residues on the beta strand give high phi values, but only PRO8, PHE10, LEU12 and TRP14 are labeled. Because these four residues have their side chain pointed to the dimer core, they have more contacts formed, but their average phi values per contact are not higher than MET7, GLN9, ASN11 and ARG13, which are not shown due to crowdedness. This suggests the role of key residues on the beta strands is not just about hydrophobic core formation, but more likely, about the initial binding between the beta strands (stage 1 illustrated above).

(To avoid over-crowded labels, we only labeled residues on Chain A (the orange backbone). The unlabeled residues in Chain B (lime colored backbone) can be easily recognized by its mirror image on chain A. Key residues only appear in chain B are labeled separately with a suffix "b".)

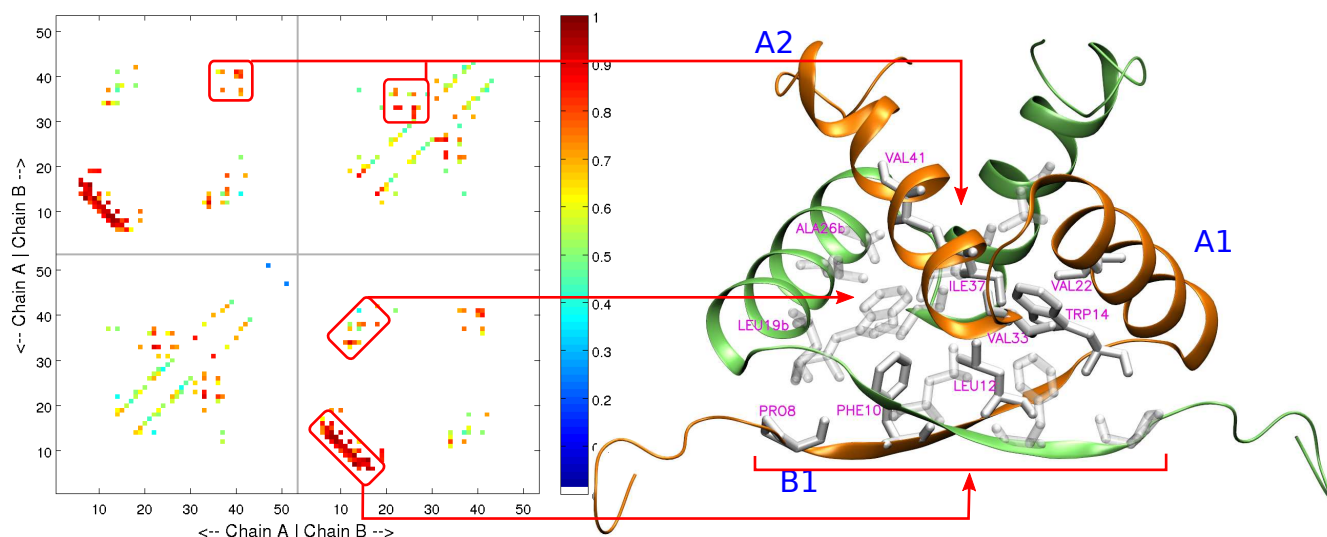


Figure 5.16: 1arr phi values

Experimental evidences and corrections: In agreement with experiments which suggest the dissociation and unfolding of Arc repressor are closely coupled in an equilibrium between folded dimers and unfolded monomers [89, 90, 91], we observe no stable structured monomers in our simulations. The free energy landscape shows its binding and folding processes are mixed in the first half and closely coupled in the second half of the transition. Suggestions have been made that the overall structure of transition state must be somewhat native-like, given the large amount of buried hydrophobic surface (75%), and the response of the folding and unfolding rates to mutations (i.e. Expected key structural residues identified in the native state can affect the rates substantially. Many mutations influence unfolding rate greatly, but very few can influence refolding rate, which suggests the transition state for unfolding keeps a lot of native like interactions.) [90, 92, 91]. In

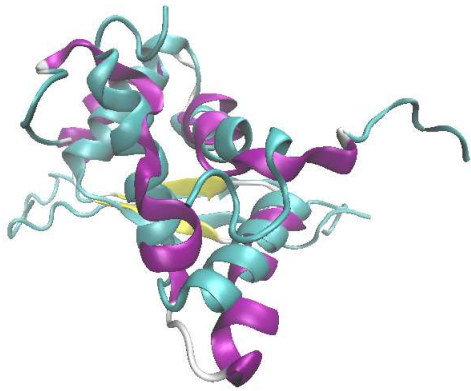


Figure 5.17: 1arr_TS_vs_native state. Cray colored is the native state, and brightly colored is the average TS structure.

our simulation results, the transition state configurations identified at the saddle of the minimal free energy path do possess this overall native-likeness, as can be seen from the structural alignment of them (Fig. 5.17 in supplement 5.17), and the average RMSD between transition state configurations and native state is about 7.5 Å, which is rather close given the size of the protein.

Given the importance and availability of Arc repressor, extensive mutagenesis experiments [90, 91, 93, 92, 94] have been carried out in studying the structural importance of each residue. After every possible alanine-substitution mutant was studied, seven out of 51 mutants (WA14, VA22, EA36, IA37, RA40, VA41, and FA45) were so thermodynamically disturbed that they didn't form dimers stable enough for kinetic study [90, 91]. Among these destructive single site mutants, with only two exceptions (E36 and F45), all other five mutated positions are identified as high phi value residues in our simulation and calculations (Fig. 5.16), suggesting the model used can preserve the importance of key structural residues. Another evidence of this is residue Pro8, which gives the longest strong band in our contact evolution map (Fig. 2.3 in supplement) (Pro8 forms native contacts early compared to other residues, and remains stably contacted through out the transition processes. Note this is not simply due to the rigidity of proline, since Pro15 gives a very ordinary band). Its mutation to alanine or leucine greatly decrease unfolding rate by two orders of magnitude, without affecting the refolding rate much [94, 90]. That is to say, proline 8 here are actually imposing a strong destructive impact on the thermostability of Arc repressor. So, strong contact formation propensity, bright colored bands in the evolution map, may only imply the critical role of corresponding residues in the transition processes, but not necessarily being constructive in stabilizing the protein.

However, we observe a stable partially binded intermediate state with a few inter-monomer contacts and low intra-monomer contacts. Scrutinizing the structures reveals that the participated regions in this initiation binding for Arc repressor is the anti-parallel beta-sheet at the N-terminal (stage 2 in the structural evolution illustrated above). This is against the suggestion [91], which argues that intermediate states, if present, are poorly populated, compared to the native and denatured protein, and that beta-strand is not formed in TS.

5.3 Lambda repressor (1lmb)

From the 3 dimensional free energy space, we can see distinctively that folding and binding of Lambda repressor are decoupled all the time at transition temperature (Fig. 5.18). One monomer completely folds first and serves as the recognition surface for the second unstructured monomer to bind to. The unstructured second monomer binds to its well formed counterpart half way to complete the binding interface, and then is induced to fold itself. Only after the second monomer folds, can the second half of binding proceed and complete the binding interface.

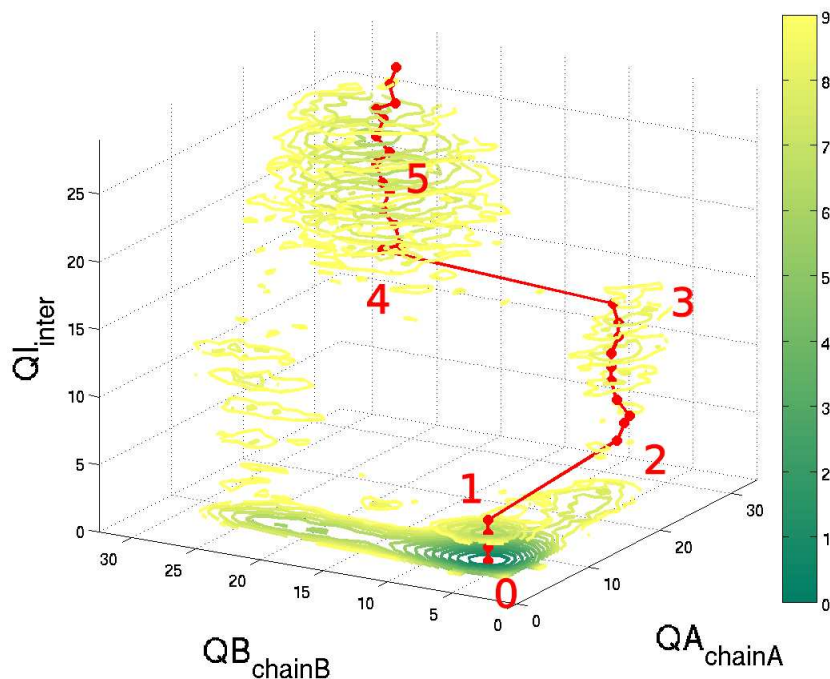


Figure 5.18: 3D landscape of 1lmb

Stage (0), figure 5.19. Before binding-folding transition starts, only temporary formation of some local contacts along the peptides is observed in the contact map, representing secondary structure propensities and random thermal collisions among nearby residues. Residue LEU12, LEU31, VAL36, MET42, VAL47, LEU50, ILE54 and ASN61 give noticeably higher contact formation probability and likely serve as the centers of (secondary structure) nucleation. Notice that these residues, representing places with higher local secondary structure tendency, distribute almost evenly over the first 4 alpha helices, which compose the monomer globular body. While the singled-out fifth alpha helix (from PRO78 to TRP88) is extended out from the monomer core to form a large part of the binding interface of the two monomers. This suggests that the fifth alpha helix may be more flexible than the other four helices, and may explain the fly-casting binding behavior observed below.

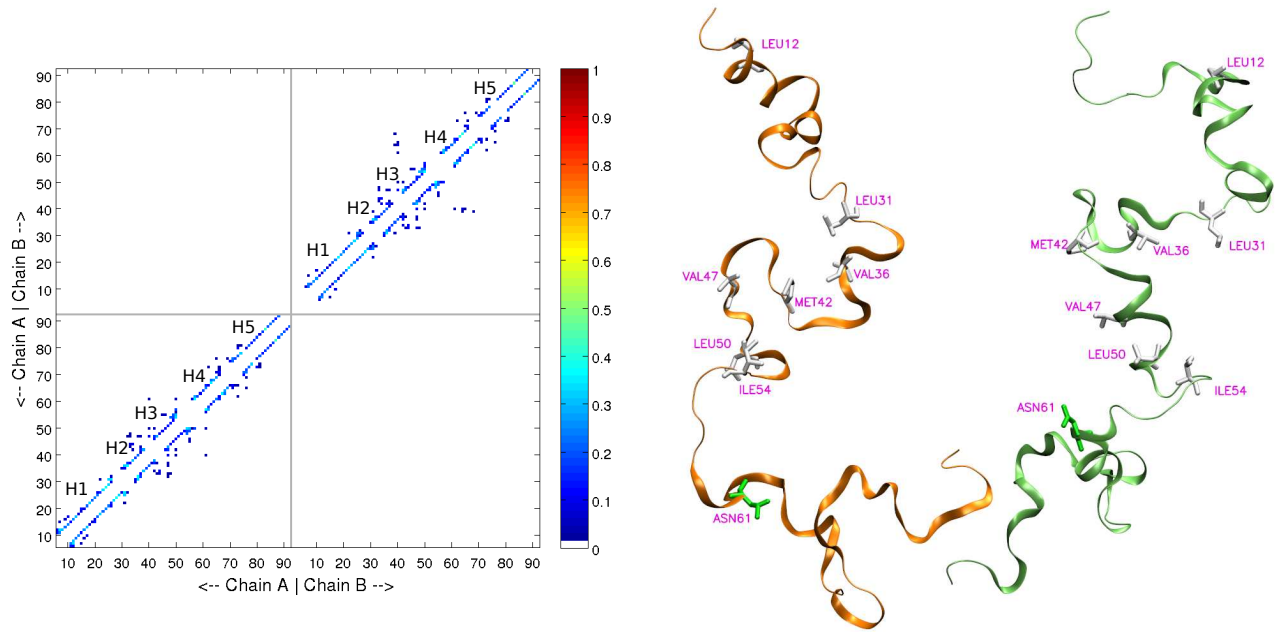


Figure 5.19: Before transition

Stage (1), figure 5.20. The step-wise transition process starts with the binding between the C terminus (including the fifth alpha helix). Residue VAL73, ILE84, TYR88 and VAL91 are involved in this initiation of binding. At this time, these interface contacts are of low formation probability (blue colored), as can be seen from the contact map. The interface between the two C terminus are not clear yet.

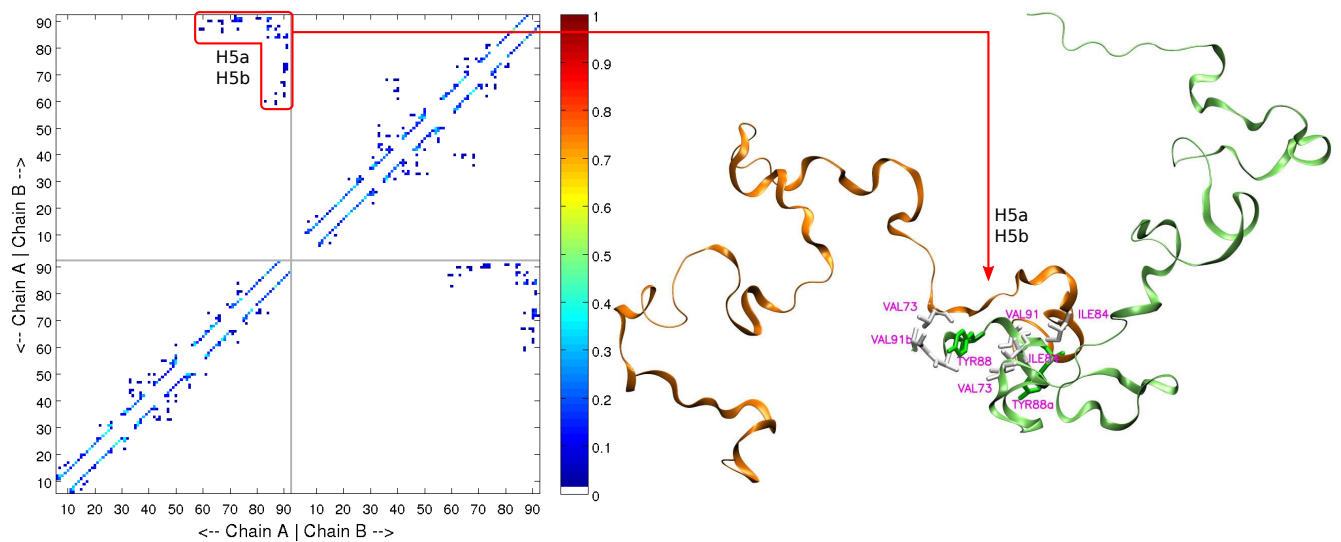


Figure 5.20: Initial binding

Stage (2), figure 5.21. Then one monomer chain condensates and folds, with most of its native contact pairs formed but not very stable yet. The key residue involved in this process are LEU18, VAL36, MET42, LEU50, LEU65 and PHE76. They form the strongest interactions in contact map and pull the first four alpha helices together to form the hydrophobic core. Meanwhile, part of the C-terminus initial binding interface gets sacrificed when the first monomer folds, i.e. contacts such as between VAL91 of Chain A and VAL73 and ILE84 of Chain B are lost.

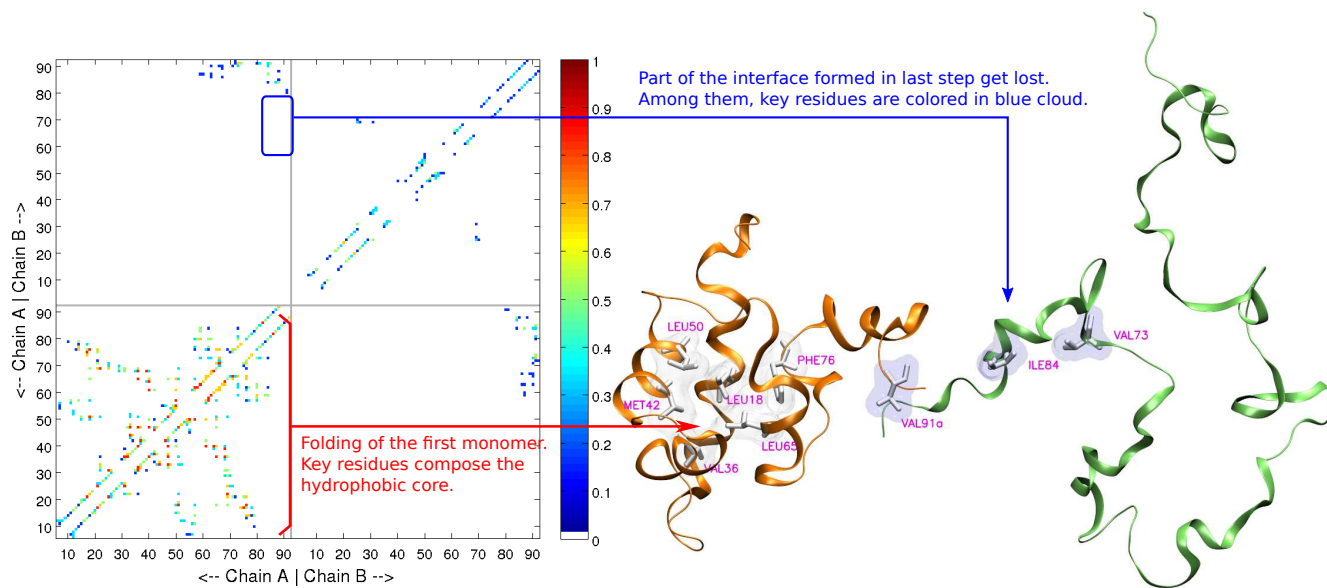


Figure 5.21: First dimer folds; Sacrifice part of binding interface.

Stage (3), figure 5.22. Binding interface grows back and gets consolidated (residues: VAL73, ILE84, TYR88, ALA90 and VAL91, the same group of residues as in stage 1, but in stronger interaction now), with one monomer folded and the other still not. Interface between two monomers become clear, which possesses two relatively well aligned alpha helix 5 from both C terminus.

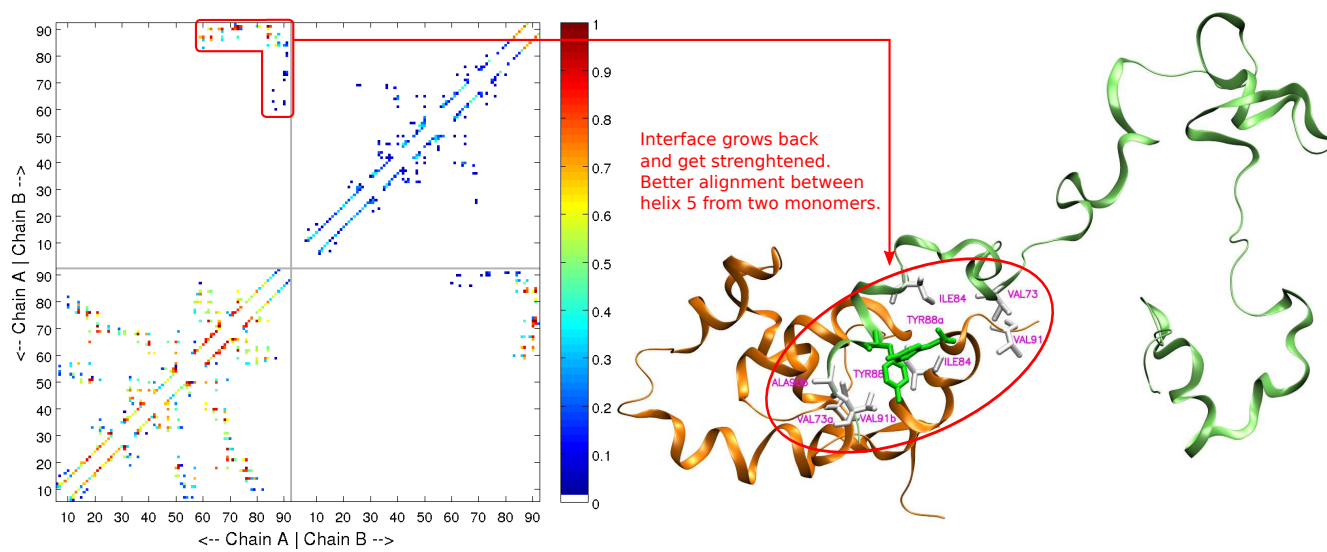


Figure 5.22: Development of interface

Stage (4), figure 5.23. Finally the second monomer folds and the binding interface gets further improved. Key residues contribute to the folding of the second monomer are LEU18, VAL36, VAL47, LEU57, LEU65, LEU69 and PHE76. Only four of them (LEU18, VAL36, LEU65 and PHE76) also contribute significantly in the first monomer folding. This suggests the folding mechanism is changed for the second monomer, since it already has a partial binding interface with its folded partner and thus is under the influence of it.

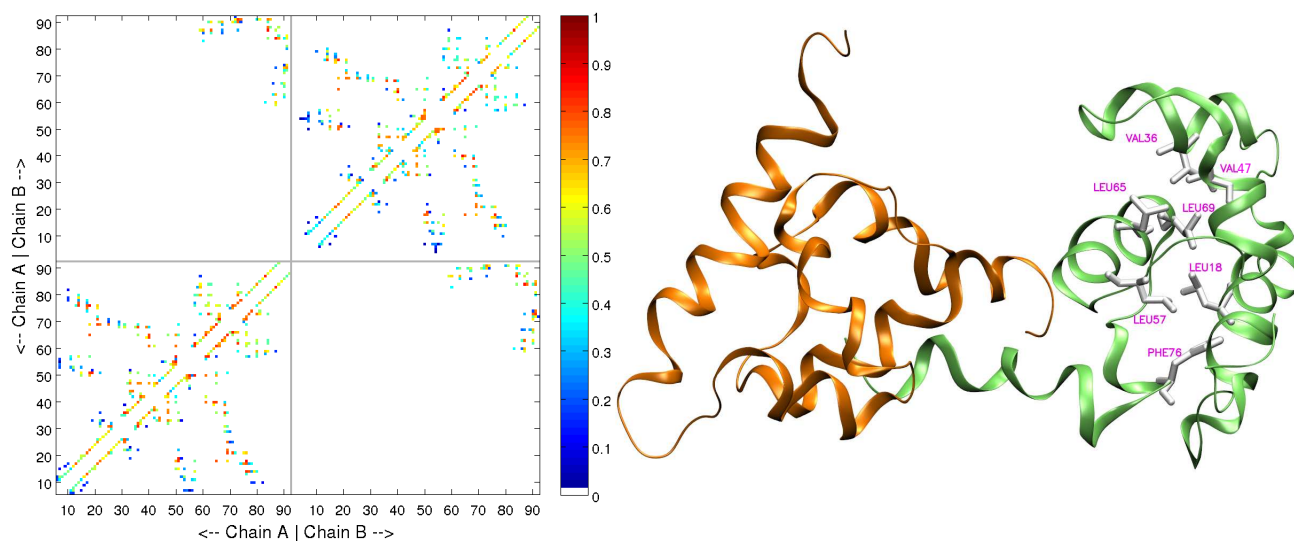


Figure 5.23: Second monomer folds

Stage (5), figure 5.24. Binding proceeds and completes the interface. The number of inter-monomer contacts increases as the probability of contact formation increases in the detail refinement of C-terminus binding interface.

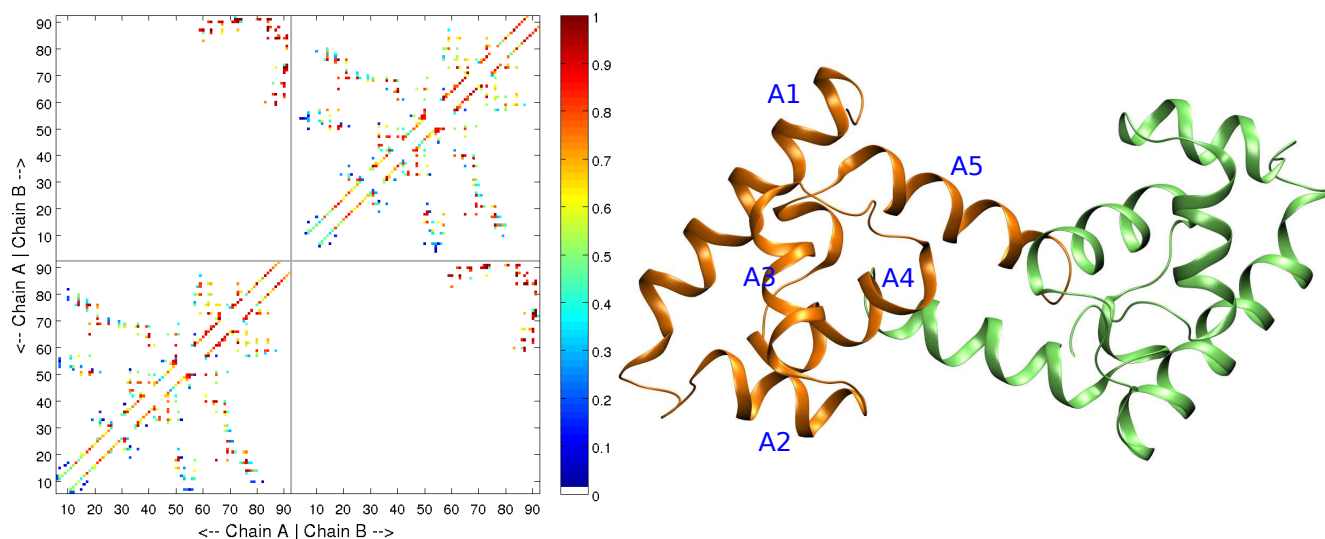


Figure 5.24: Native state

Phi values. Since there are two clearly separate transition processes in the coupled binding-folding of 1lmb, we performed the phi value analysis for each of these two transitions, and identified the high phi-value "key residues", which are more likely to have their native contacts formed in the transition state (Fig. 5.25). We can see the key residues in the first monomer folding transition are located around the hydrophobic monomer core, which indicates that hydrophobic collapse is likely the key feature of this transition (the white cloud in the structure is the surface of these key residues). While, distinctively, the second transition have key residues located on and around the binding interface of the fifth helix, instead of the monomer core. The surface of high phi value residues in monomer A is colored orange; surface of residues in monomer B is colored lime. In this way, the binding interface can be revealed more clearly. In agreement with previous observation in its structure evolution, the second monomer takes a different folding path from the first one, under the influence of the binding interface with its folded partner. This binding-coupled-folding is what fly-casting mechanism suggests.

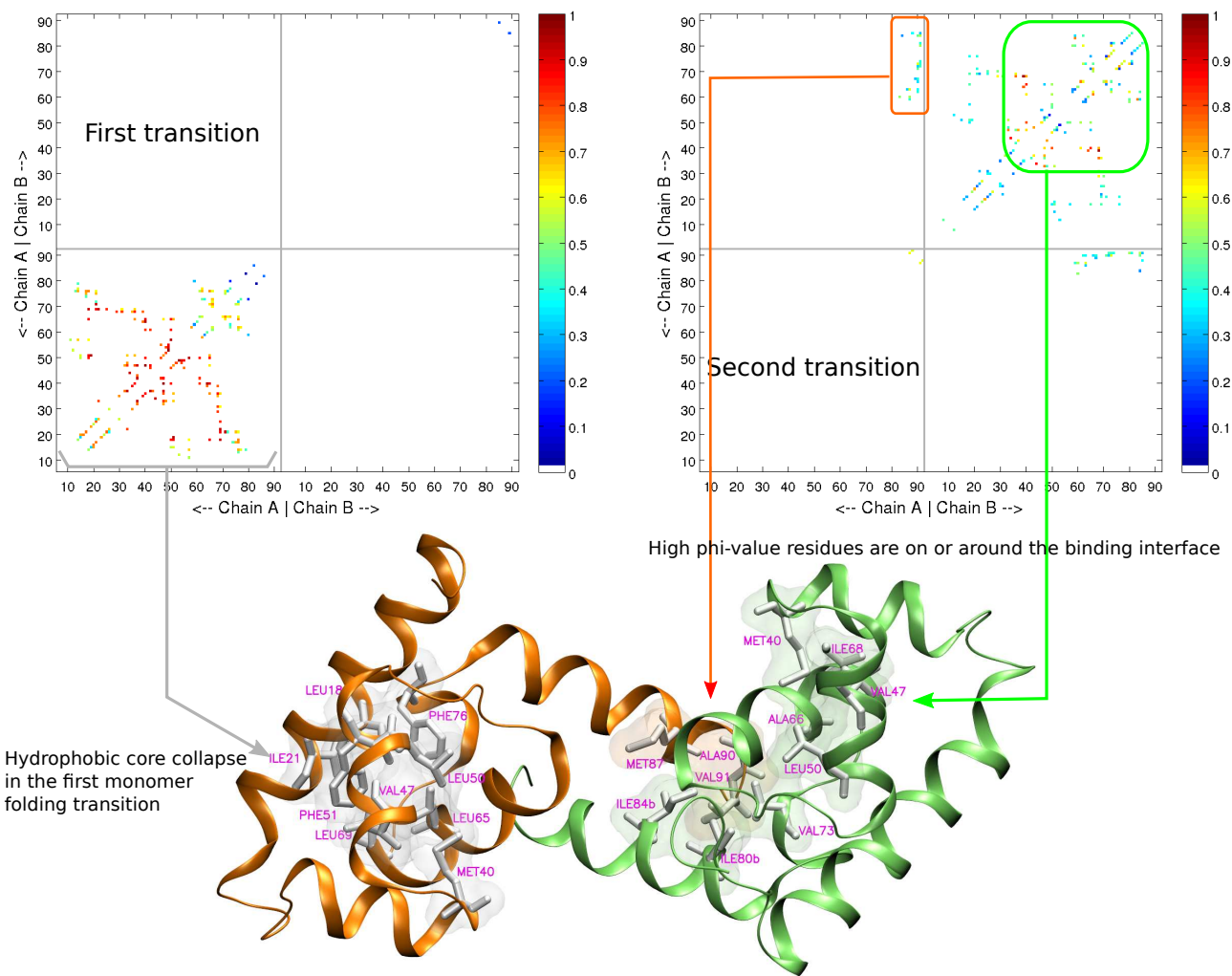


Figure 5.25: 11mb phi values

Experimental evidences: Our simulation reveals an asymmetric behavior of lambda repressor dimer, which could possibly explain the experimental observation that at least two different folding routes exist for lambda repressor monomer [95], and its dimer binds to the two halves of its operator differently [96], and further more its two subunits also have slight difference in crystal structure [97]. Similar results was found from other studies with AWSEM model, in which the intermediate state observed consists an ensemble of a variety of encounter complexes [98].

In reality, our model may have missed certain properties due to the absence of operator DNA in the simulation. It has been shown that the dimerization of lambda repressor and its binding to DNA are coupled equilibria and the intermediate quaternary structure contributes to operator affinity of lambda repressor [5].

5.4 Streptomyces subtilisin inhibitor (3ssi)

For Streptomyces subtilisin inhibitor (3ssi, Fig. 5.26), after slight initial binding, two monomers collapse significantly with only a slight increase in binding contacts. Then binding proceeds to complete binding interface after the two monomers become relatively folded (globular shaped).

Given that no stable intermediate state is found in 3D free energy landscape (Fig. 5.26), we observe that 3ssi is a two-state dimer, which is in agreement with experimental observations [99, 100, 101]. However, its minimal free energy path is not a coupled binding-folding pathway (not following the diagonal closely, like that of 1f36), since it has a horizontal folding segment (from stage 2 to stage 3 in Figure 5.26) followed by a vertical binding segment (from stage 3 to stage 4 in Figure 5.26).

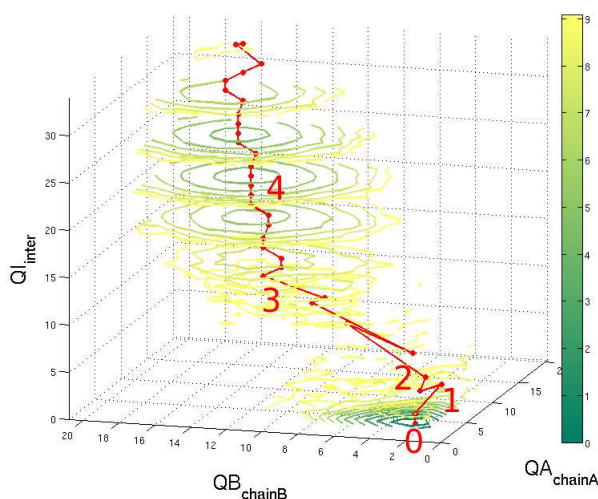


Figure 5.26: 3D landscape of 3ssi

Stage (0), Fig. 5.27. Local contacts weakly form around several centers, preventing the peptide chain from extending too long. Secondary structure propensity, including both alpha helix and beta sheet, is revealed here, especially at the ends of neighboring beta strands (THR34 at the end of beta strand 2; VAL85 at the end of beta strand 3). Other residues, PRO39, LEU53, VAL56, ASN104, GLY107, and ALA112 also form local contacts actively.

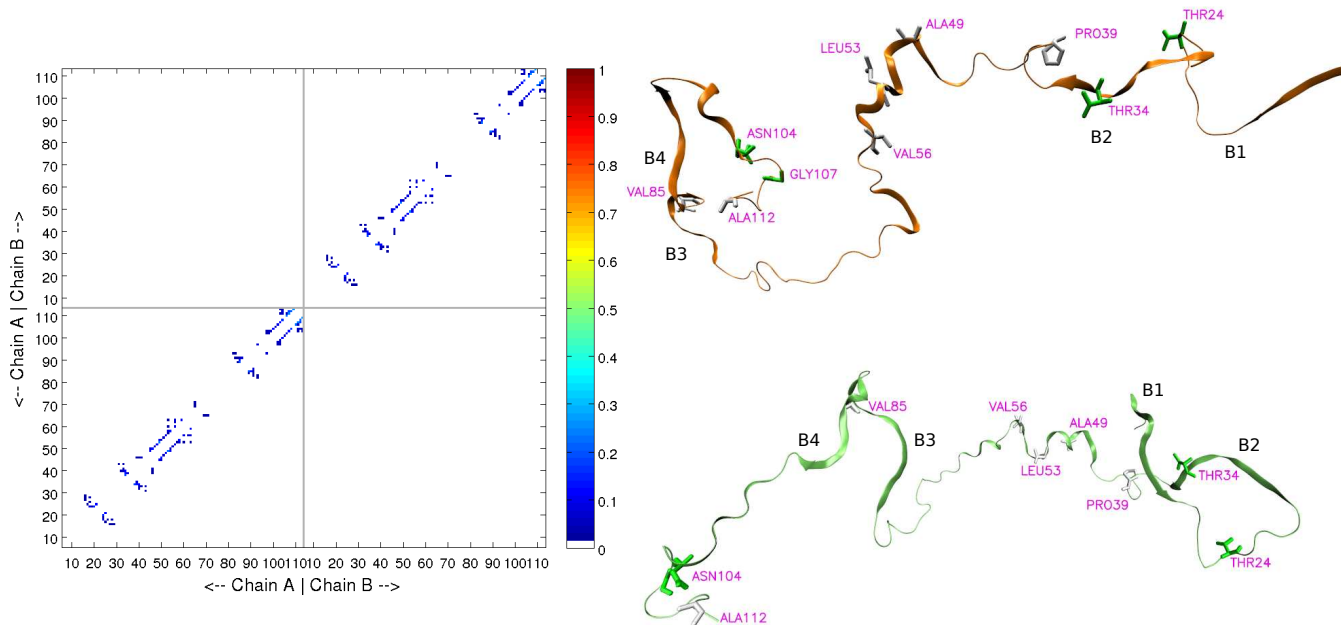


Figure 5.27: Local contacts formation centers. Highlighted residues in the structure are those who give relatively higher formation probability in the contact map on the left.

Stage (1), Fig. 5.28. Intra-monomer contacts between beta strand 3 and 4 likely form first. So does part of beta strand 1 and 2. We don't see the same patterns in the contact map for two monomers, because the way we trace minimal free energy path is restricted to the right side of the diagonal of the 3D landscape, where monomer A always have no less intra-contacts than monomer B. In fact, monomer B is similar to monomer A at this stage, given that the 3D free energy landscape are very symmetric about the diagonal, and structures of monomer A and B are alike. Key residue involved here are: VAL16, GLY17, ALA25 and GLUE28 between beta strand 1 and 2; and ASP76, VAL 78, ASP83, VAL85, LYS89, VAL91, VAL96, SER98, and ASN99 around beta strand 3 and 4).

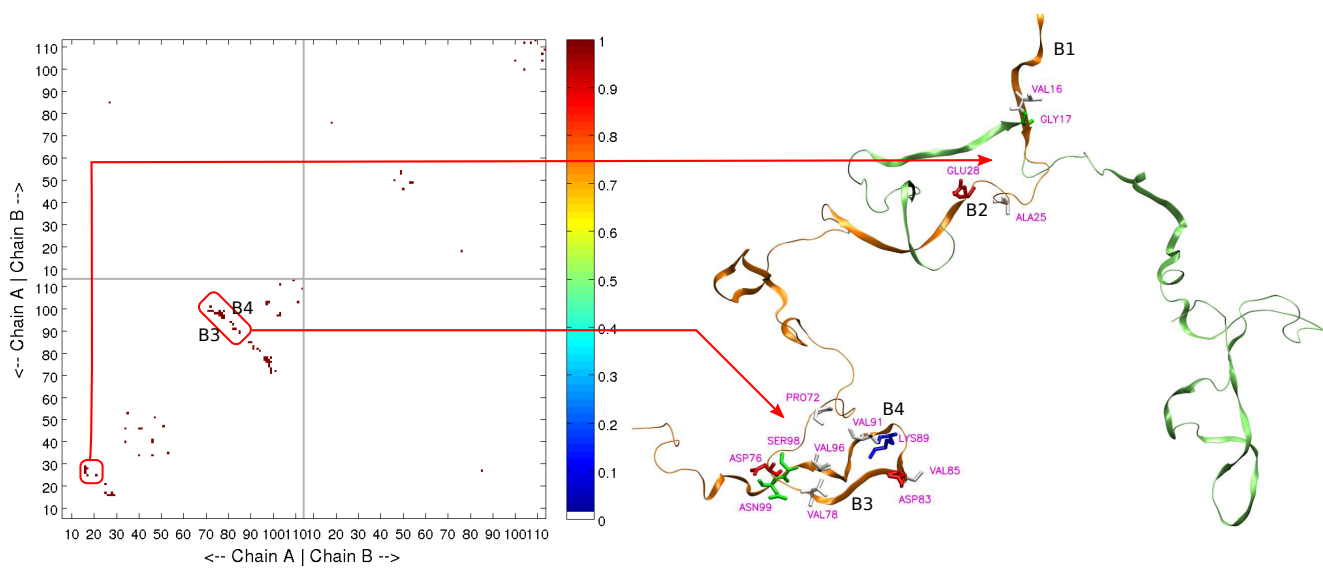


Figure 5.28: Beta strands 3 and 4, and part of strands 1 and 2 appear in one monomer.

Stage (2), Fig. 5.29. Inter-monomer binding starts between beta strands 1 of one monomer and beta strand 2 from the other monomer. (residues: PRO9, ALA11, LEU12, VAL13, GLU28, ARG29, and ALA30 in chain A; and PRO9, VAL13, ALA30, and THR32 in chain B.) At this time, intra-monomer contacts between beta strand 3 and 4 in a monomer are relatively well formed, but strand 1 and 2 still lack significant intra-monomer contacts.

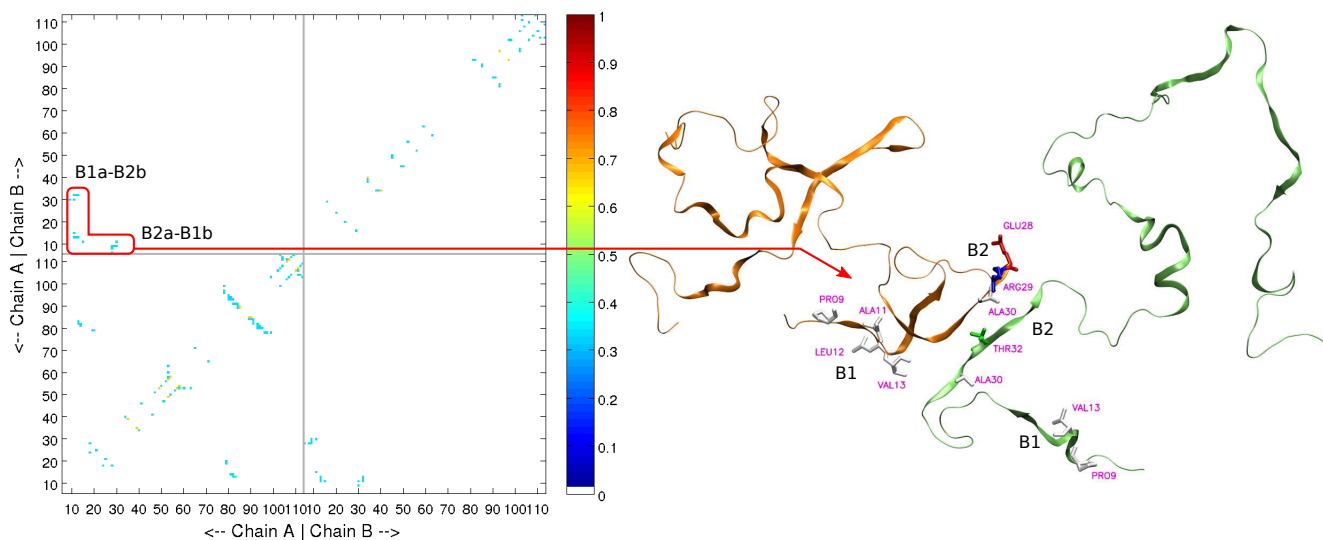


Figure 5.29: Binding starts slightly

Stage (3), Fig. 5.30. Folding moves on, developing intra-contacts between beta strands 1 and 2, 1 and 3, and 3 and 4. Two subunits come together, but the binding interface is still not very well formed, as can be seen from the low inter-monomer contact number in the 3D free energy landscape and the low contact formation probabilities (blue dots) in the 2nd and 4th quadrants of the contact map. (Key residues: LEU12, VAL13, VAL31, LEU33, LEU80, THR81, VAL82, GLY84, VAL91, ARG95, and PHE97 in chain A; and LEU14, VAL31, LEU33, LEU80, THR81, VAL82, VAL91, ARG95, and PHE97 in chain B.)

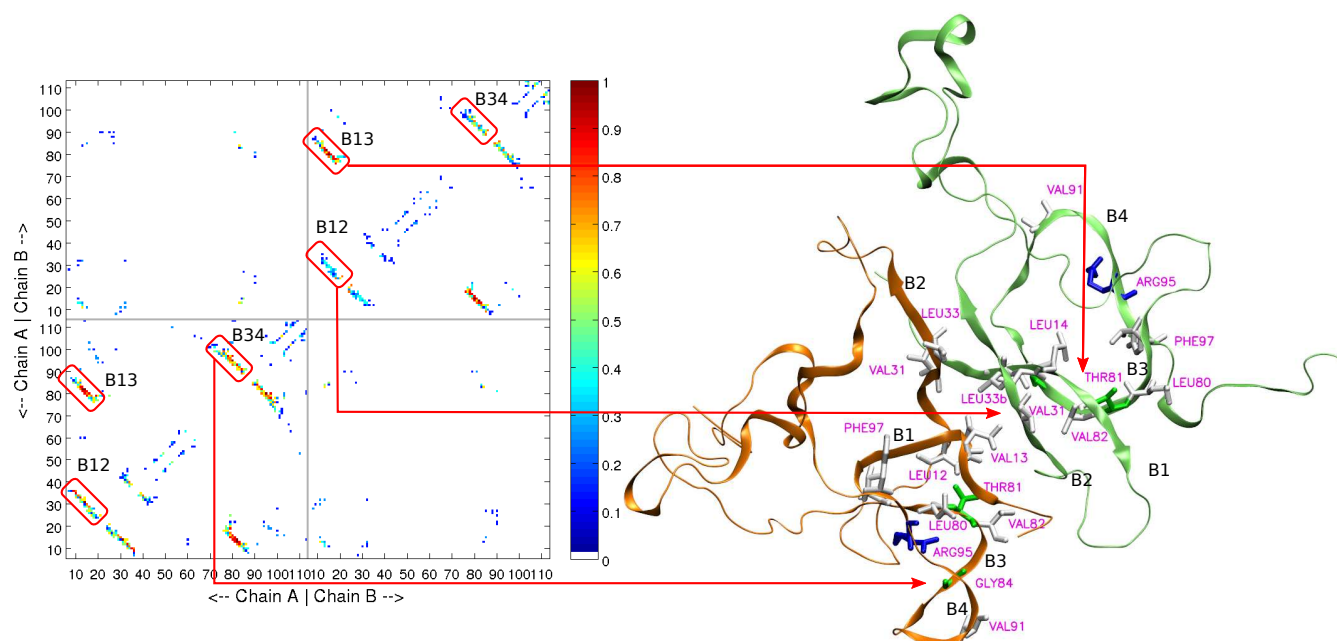


Figure 5.30: All intra-monomer beta strands come into being.

Stage (4), Fig. 5.31. Contacts on binding interface develops and folding contacts get strengthened, especially among those beta strands which make up the binding interface. This concurrence of binding interface formation and folding contacts consolidation, represents the underlying coupled binding-folding mechanism. On one hand, four beta strands in one monomer need to be aligned by folding in order to establish its surface for binding; and on the other hand, the interaction from the other monomer may guide the alignment of these four beta strands. Thus folding and binding depend on and cooperate with each other. Residue involved in the interface at this stage are: THR32 and LEU33 on beta strand 2 (plotted transparently in the figure because they are far in the back); THR81, ASP83, GLY84 and VAL85 on beta strand 3; GLY88 between strand 3 and 4; and ARG90 at the beginning of beta strand 4.

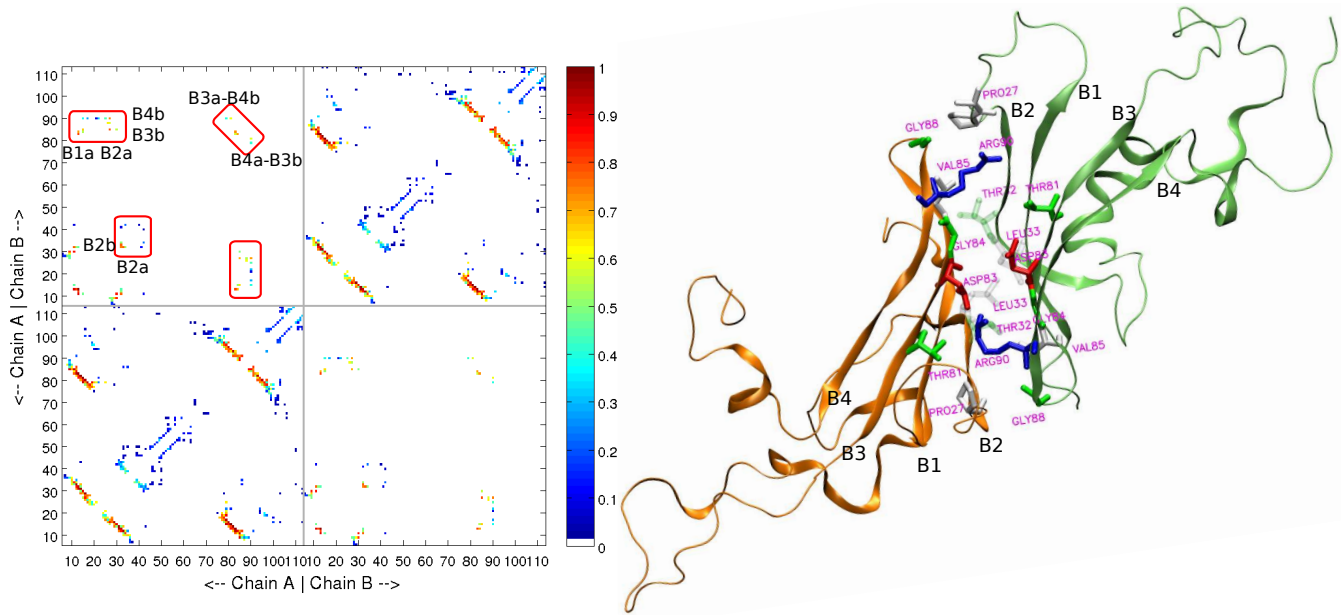


Figure 5.31: Binding interface formed.

Stage (5), figure 5.32. Further strengthening of intra- and inter-monomer beta strands contacts occurs for refining the final binding complex. However, to discuss the later refinement process of the complex structure can be risky for us here, since the monomer has got reasonably well packed and thus very likely that disulfide bonds are created, which was not represented in our model. But the previous association process revealed in our analysis above keeps its value if we assume the disulfide bonds are not formed in the beginning of folding-binding transition.

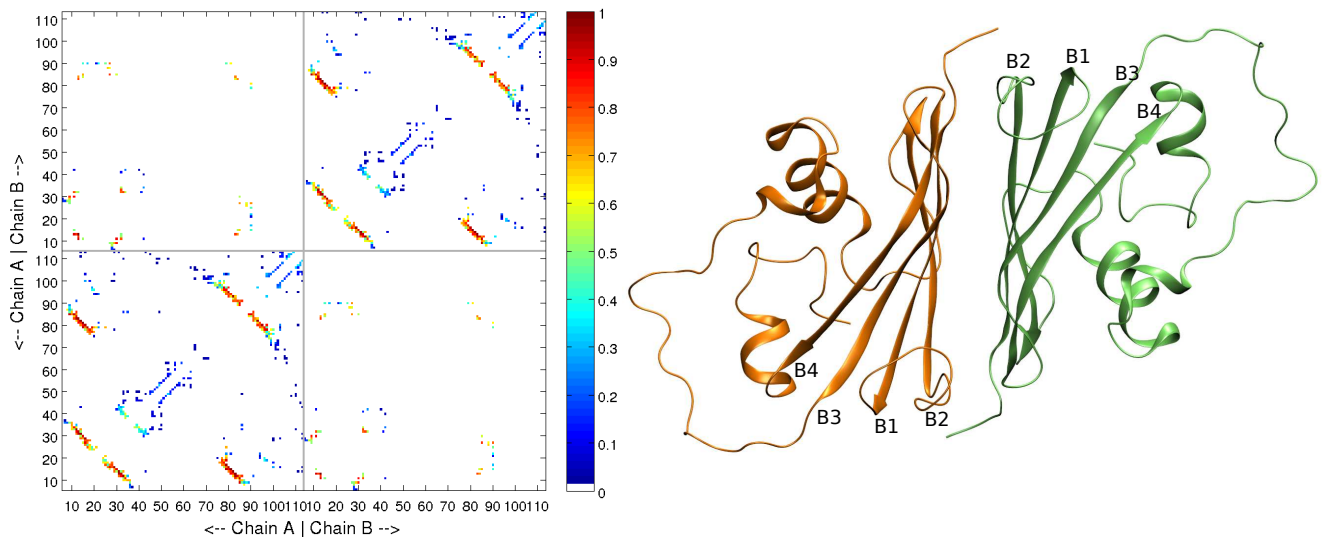


Figure 5.32: Native structure

Phi Values: From the phi value contact map (Fig. 5.33), we can learn that, in the transition state, intra-monomer contacts between beta strand 1 and strand 3, between beta strand 1 and 2, and between beta strand 3 and 4, are relatively well formed, especially the two central beta strands (1 and 2). This agrees with our free energy profile and structure evolution, in the sense that monomer folding is the saddle point on the minimal free energy path.

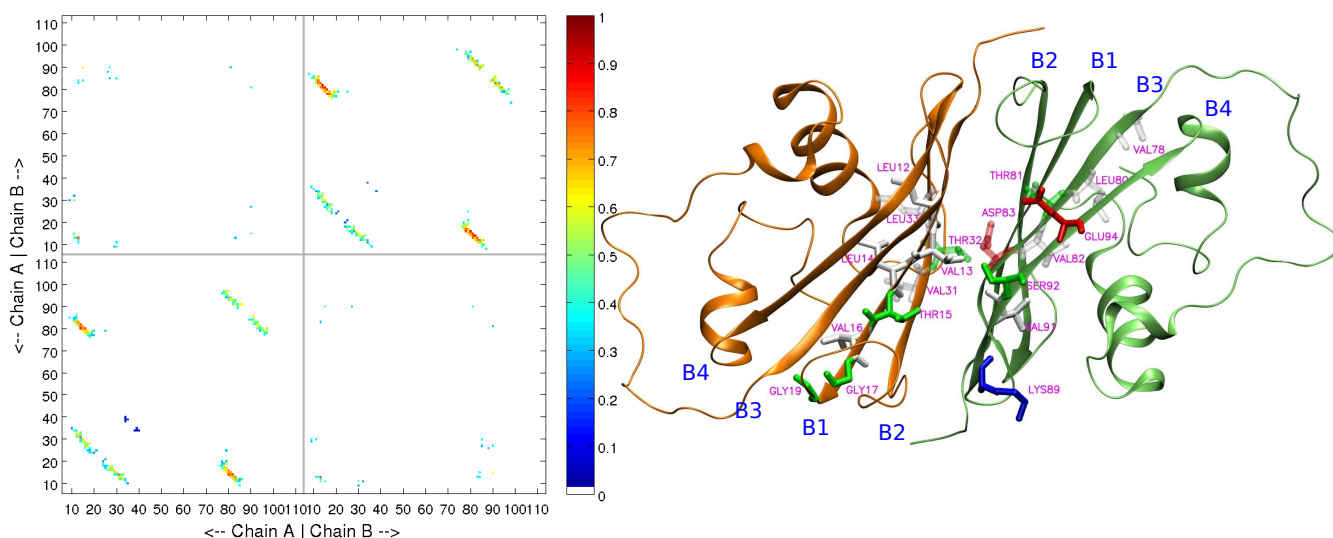


Figure 5.33: 3ssi phi values. The phi value map is almost the same for two monomers, and their structures are symmetric about the center. To avoid crowding, we only labeled key residues on beta strand 1 and 2 for monomer A, and key residues on strand 3 and 4 for monomer B.

Experimental evidences: In experiments, no monomeric native or intermediate forms have been found upon denaturation, and the native dimer dissociates into denatured monomers (i.e. 3ssi is a two-state dimer) [99, 100, 101]. The previous simulation work with course-grained residue level model classified 3ssi into 3-state dimer [21], while we are happy to find that our all-atom structure-based model used here gives only two basins on free energy landscape, which agrees with the two state dimer scenario in the experiment.

Each subunit possesses two disulfide bridges (CYS35-CYS50 and CYS71-CYS101), which plays essential roles in holding the structure integrity of each folded subunit [102, 103]. We did not take into account these disulfide bridges. This resulted in not sufficiently well folded subunits at folding-binding transition temperature (Fig.5.34).

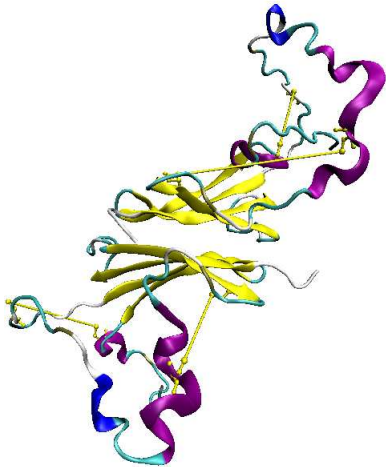


Figure 5.34: The disassembled subunits of 3ssi. Yellow line indicates the elongated disulfide bonds.

Though this approximation keeps us from further interpreting the binding-folding mechanism of 3ssi, it proves, from the other side, that disulfide bond, as a covalent chemical bond, is more important than a regular contact pair in structure-based model, which is intended to model hydrogen bond or hydrophobic packing; and more remarkably, the all-atom model used here can tell the difference between them. (Given that, in its native state, each subunit of 3ssi is a well packed globular protein, possessing 306 defined intra-contacts in each subunit and 79 inter-contacts between them, 3ssi lacks no intra-monomer contacts, but still can not form stable monomer without correct representation of two disulfide bonds.)

Also, hydrogen exchange experiments has observed that amide protons on the beta strand interface exchanges only by global unfolding, while the two helix and C terminal region can exchange by local fluctuation [104]. Many of the residues in helix 1 and helix 2 have no detectable protection factors in H exchange experiments, but beta strand 1 and beta strand 4 on the beta sheet have well-organized secondary structure [99]. These experimental suggestions of the superior stability of beta sheet interface and the lack of stability in helices are clearly observed in our simulation as well.

5.5 LFB1 transcription factor (1lfb)

Due to the clear separation between binding and folding (it even gives different transition temperatures in our simulations, evidenced by two peaks in heat capacity curve Fig. 6.12) in LFB1 transcription factor (1lfb), it is impossible to demonstrate its binding-folding process on free energy profile at only one temperature (T_f). Experiments have found that the homeodomain of LFB1 (the exact part of LFB1 we simulated here) exists as a monomer in solution, and only forms dimer when joined with another domain (domain B) of LFB1 upon binding to DNA [105, 106].

From the contacts evolution map at its binding temperature (Fig. 5.36), we see that the first/left half of map 1, 2 and 4 are completely white (lacking of sample points), indicating no sample points exit with low total/chainA/chainB contact numbers at binding transition temperature, since the monomers are already folded when binding starts.

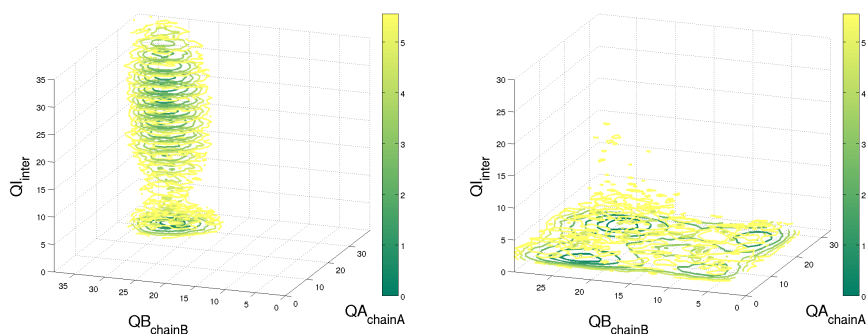


Figure 5.35: 3D free energy landscape for 1fb binding (left, at temperature 103.3) and folding (right, at temperature 111.1) respectively

We observe many almost constant probability strips from or around the interface sequences (Fig. 5.36, map 2, the long yellow color strips go all the way from left to right on the map, especially from or around those interface residues; interface residues are marked yellow on the side bar, and residues not on the interface are colored dark red), suggesting the contact surface on each subunit is well formed, and the monomers dock to each other rather rigidly (lock and key model).

At a higher temperature, unfolding of monomers can be observed. The stable patterns with high probabilities on the map of monomer folding transition (Fig. 5.37. uniform yellow bars in lower part of map 3 and upper part of map 4) suggest that the two subunits are independent from each other in folding. Two subunits follow the same developing patterns just like twins, reflecting each other. (map 4 is the image of map 3 if we place a mirror on the gray line in the middle).

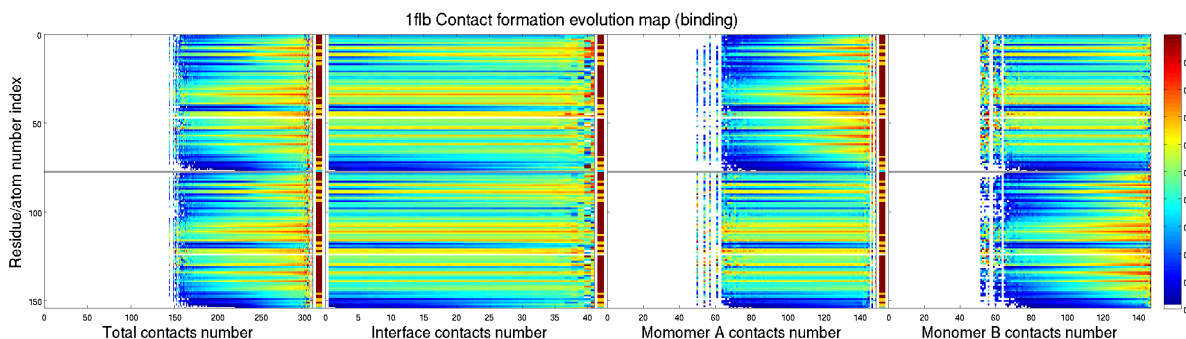


Figure 5.36: Contacts evolution map at binding temperature

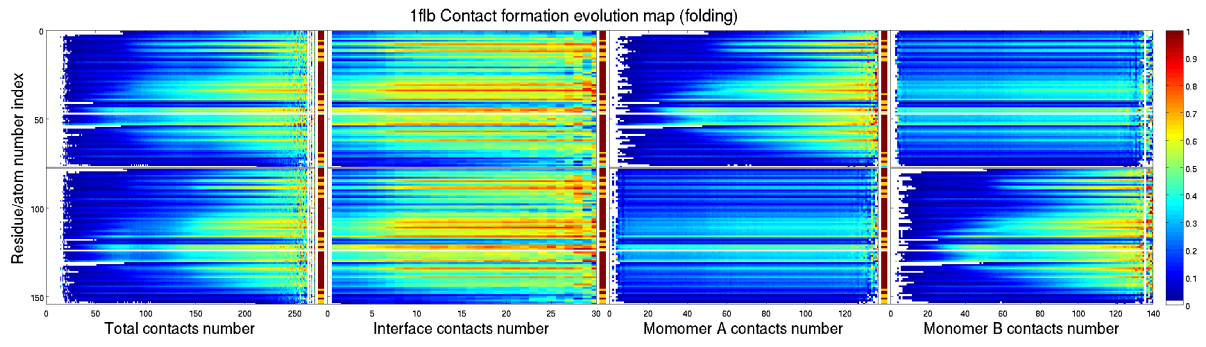


Figure 5.37: Contacts evolution map at folding temperature

6 Landscape topography of dimerization

Free energy profiles of lower dimensions (2 dimensions) are presented in the section. They may not capture the relationship between binding and folding as well as 3D landscapes demonstrated above, because Q_a and Q_b have to be averaged together in describing monomers folding. But they still have other merits over 3D landscapes. (1) Thanks to the reduce of computation resources required in calculating lower dimensional free energy profiles, finer bins can be used in histograms. (2) Due to the removing of one dimension, more sampling frames fall into each single bin, and thus gives less rugged profiles. (3) Visualization is easier and more appealing, since the convention of free energy wells/basins can be directly represented here.

Moreover, different reaction coordinates (root mean square deviation, center of mass distance, native contacts number, and some of their combinations) are tested, and the proper reaction coordinate (native contact number Q), which gives the most clear and interpretable landscapes, was chosen based on the comparison of 2D free energy landscapes.

6.1 Density of state of the funneled landscape

From figure 6.1 and figure 6.2, we can see that configurational entropy, $S = \ln(n(Q))$, decreases monotonically with Q_i and $Q_a + Q_b$, and eventually vanishes when $Q = 1$, which is corresponding to the full native state. This indicates, as folding/binding proceeds, the cross-section area of the funnel, measured by $S(Q)$, keeps decreasing until it reaches a single point, i.e. the unique native state. And, the entropy S decreases with the decreasing of energy E (Fig. 6.3), demonstrating the landscape is indeed funneled.

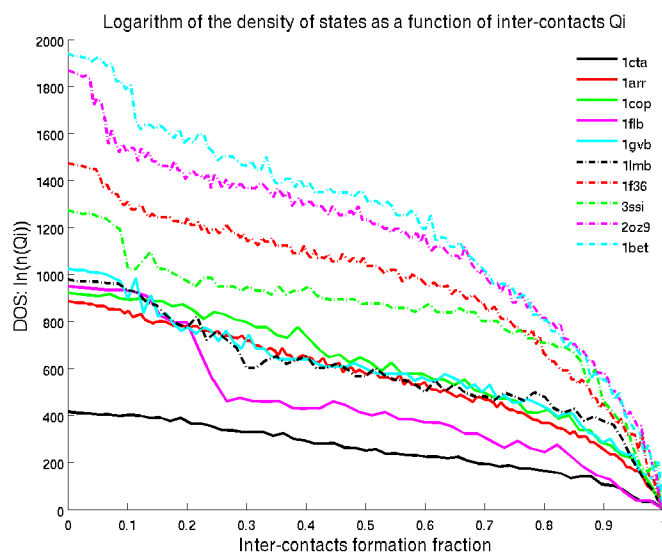


Figure 6.1: Logarithm of the density of states as a function of inter-contacts Q_i (native contacts between two monomers).

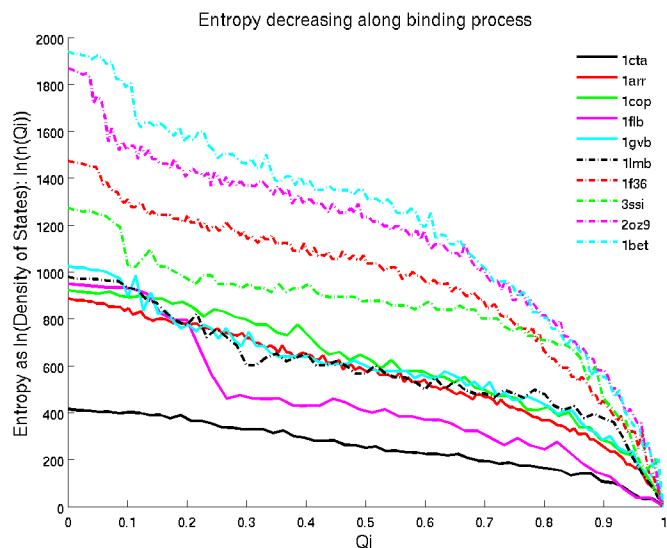


Figure 6.2: Logarithm of the density of states as a function of intra-contacts Q_a+Q_b (native contacts within two monomers).

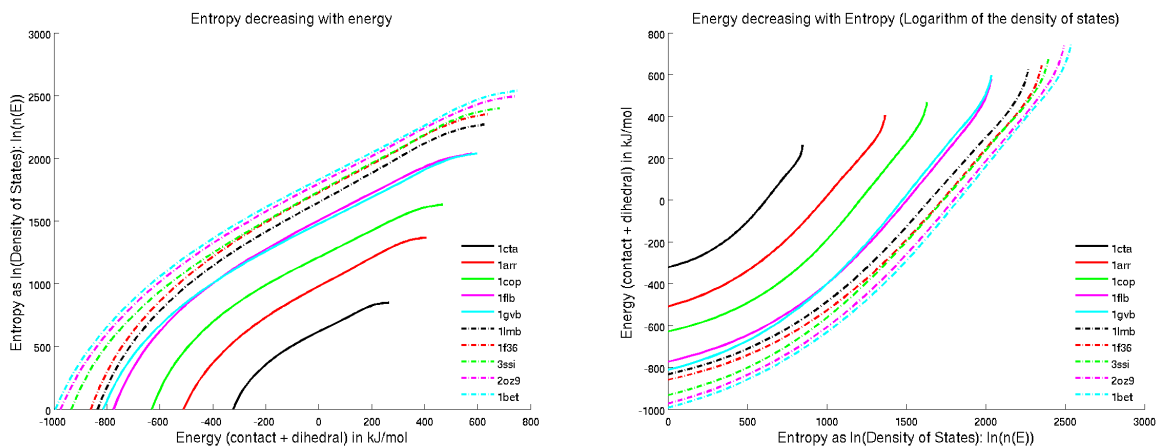


Figure 6.3: Logarithm of the density of states as a function of energy.

6.2 Binding-folding mechanism in Free energy profile

With the micro-canonical DOS of the system, free energy profile, i.e. free energy as a function of reaction coordinates Q_i or Q_a , can be drawn at any temperature. We can gain a first grasp of the binding-folding relationship through the 2D free energy profile (Fig. 6.4. The left hand side landscape shows a 2 state dimer; while the right hand side one is a clear 3 state dimer). Since unbound states usually involve zero interface contact, the unbound state wells on the free energy surface are very narrow.

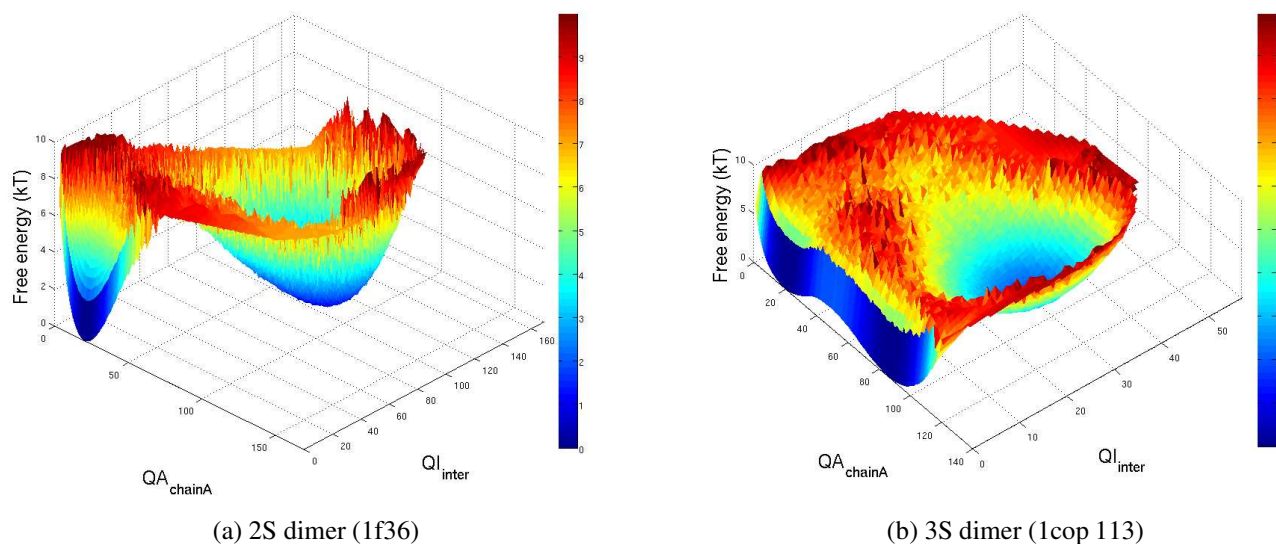


Figure 6.4: 3D free energy profile examples. Left is show for a 2 state dimer; right is for a 3 state dimer.

From the 2D free energy landscapes and heat capacity curves, we can infer the binding-folding mechanism for each dimer. Free energy landscape is constructed at the transition temperature of each dimer, which is identified as the peak of heat capacity curve. For 8 of the 10 dimers, there is only one dominant peak on heat capacity curve (not all shown here, similar to the one in Fig. 6.8), which may suggest coupled folding-binding. The rest two of these dimers have two peaks in heat capacity curve that are expected to represent folding and binding respectively. And these two two-peaks dimers have clear 3 states appearance on their 2D free energy profiles. This part of analysis/interpretation had actually been finished before the construction of 3D landscapes, we can see that the results are consistent with those from 3D free energy landscapes. But, we can also notice the limitation of 2D landscapes.

Six quite different binding-folding patterns are suggested (the name and thumbnail structure of each dimer can be found in Table 2):

1. Coupled binding-folding.

There is one dominant peak in heat capacity curve, and only two basins on the free energy contour. (1f36 Fig. 6.5; 1gvb Fig. 6.7; 2oz9 Fig. 6.6)

2. First fold, then bind. Binding follows folding immediately, without any interval. (1cop Fig. 6.8).

There is only one peak in heat capacity. There are three basins on contour, and they co-exist at the same temperature.

3. First partially bind, then followed by coupled binding-folding. (1arr Fig. 6.9; 1cta Fig. 6.11)

4. One monomer first fold, then the second monomer bind, and then the binding interface formed, and finally the second monomer fold. (1lmb Fig. 6.10)

5. First fold; then bind, with an interval exists in between. (1fb Fig. 6.12).
There is clearly two peaks in heat capacity; three basins exist on contour, but only two of them can co-exist at the same temperature.
6. Decoupled binding-folding, with binding interface complete first (3ssi Fig. 6.13, dominant heat capacity peak) and then folding compete (3ssi Fig. 6.14, lower heat capacity peak).

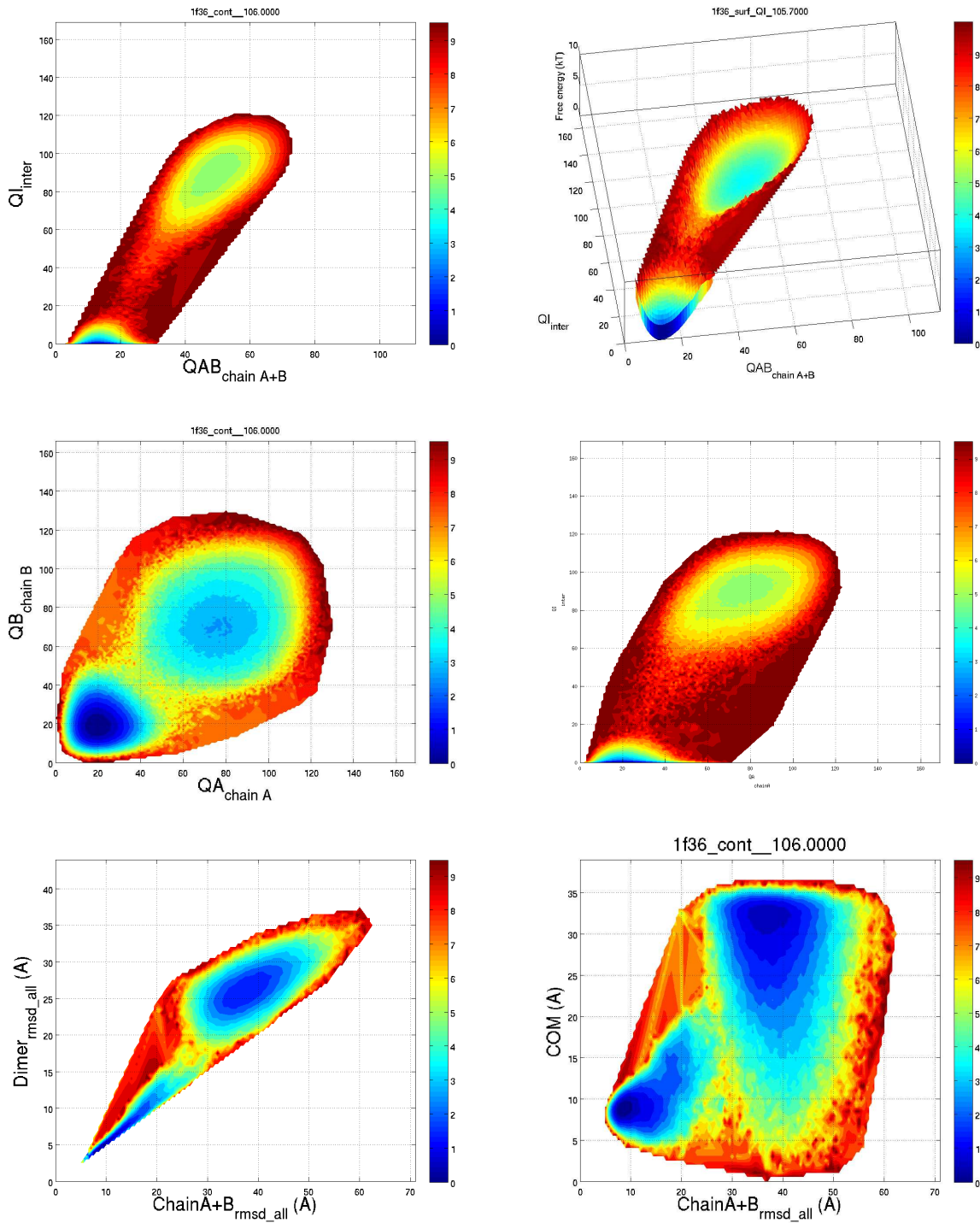


Figure 6.5: 1f36; Factor of inversion stimulation. 2 state dimer.

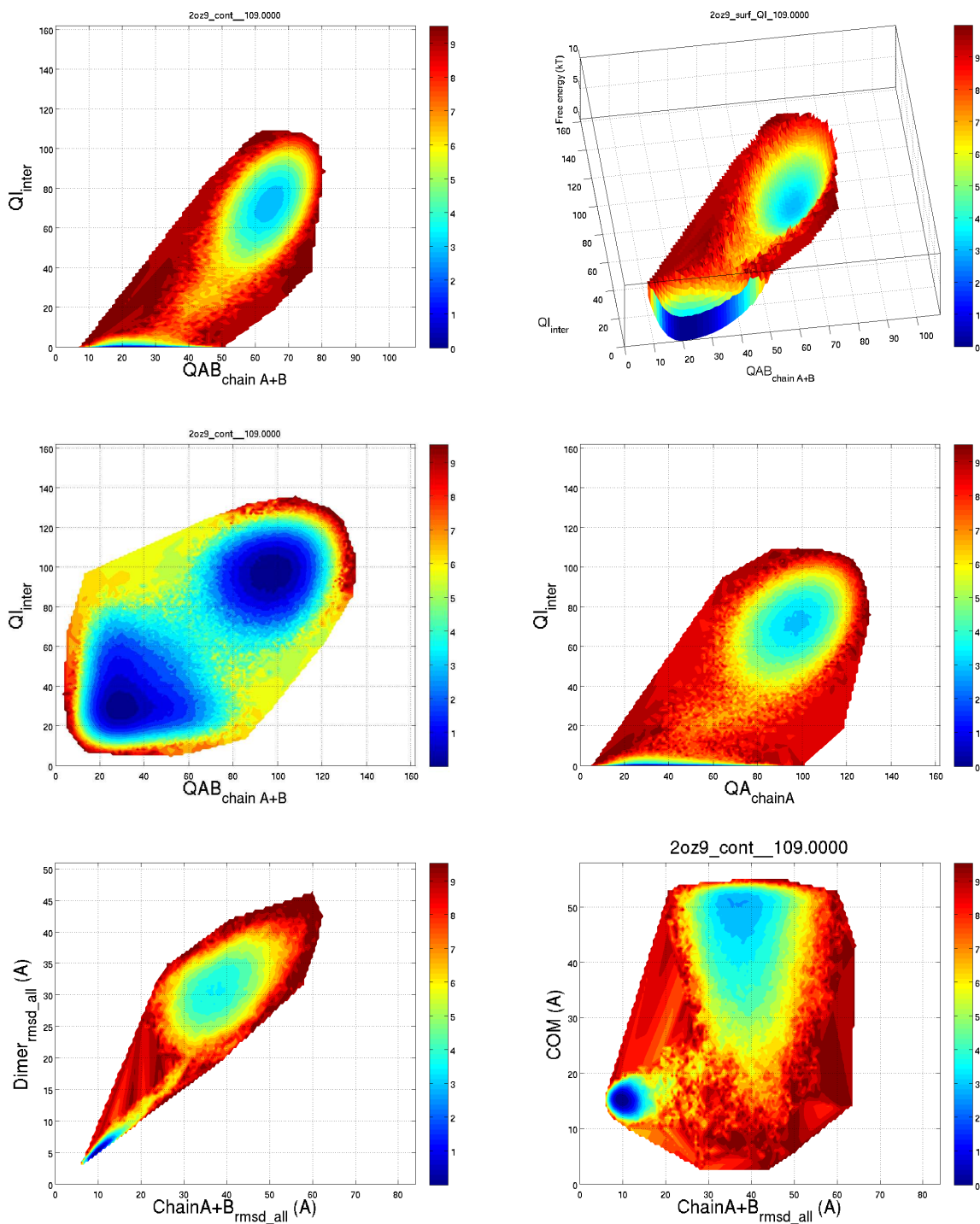


Figure 6.6: 2oz9; Trp repressor. 2 state dimer.

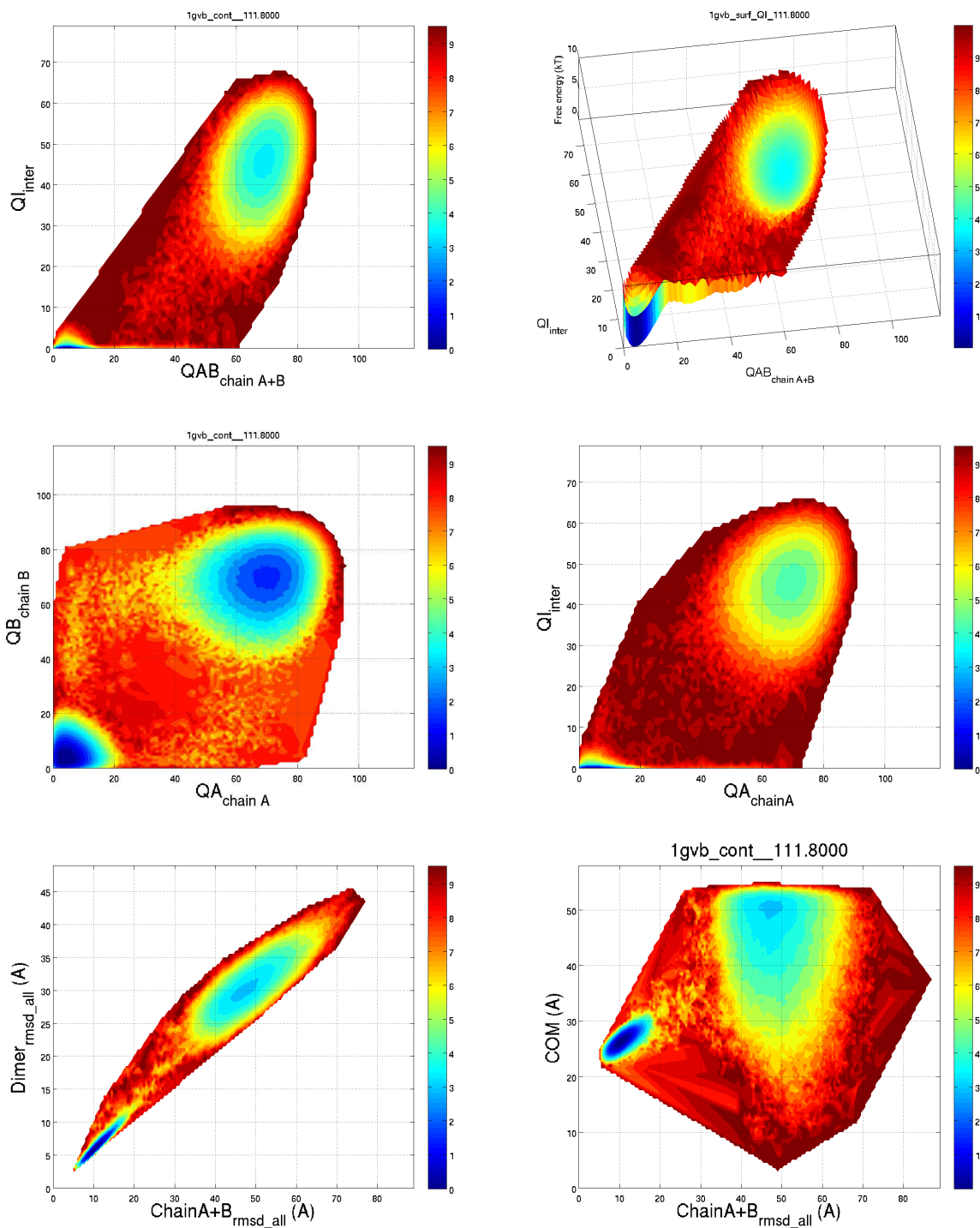


Figure 6.7: 1gvb; Gene V protein. 2 state dimer.

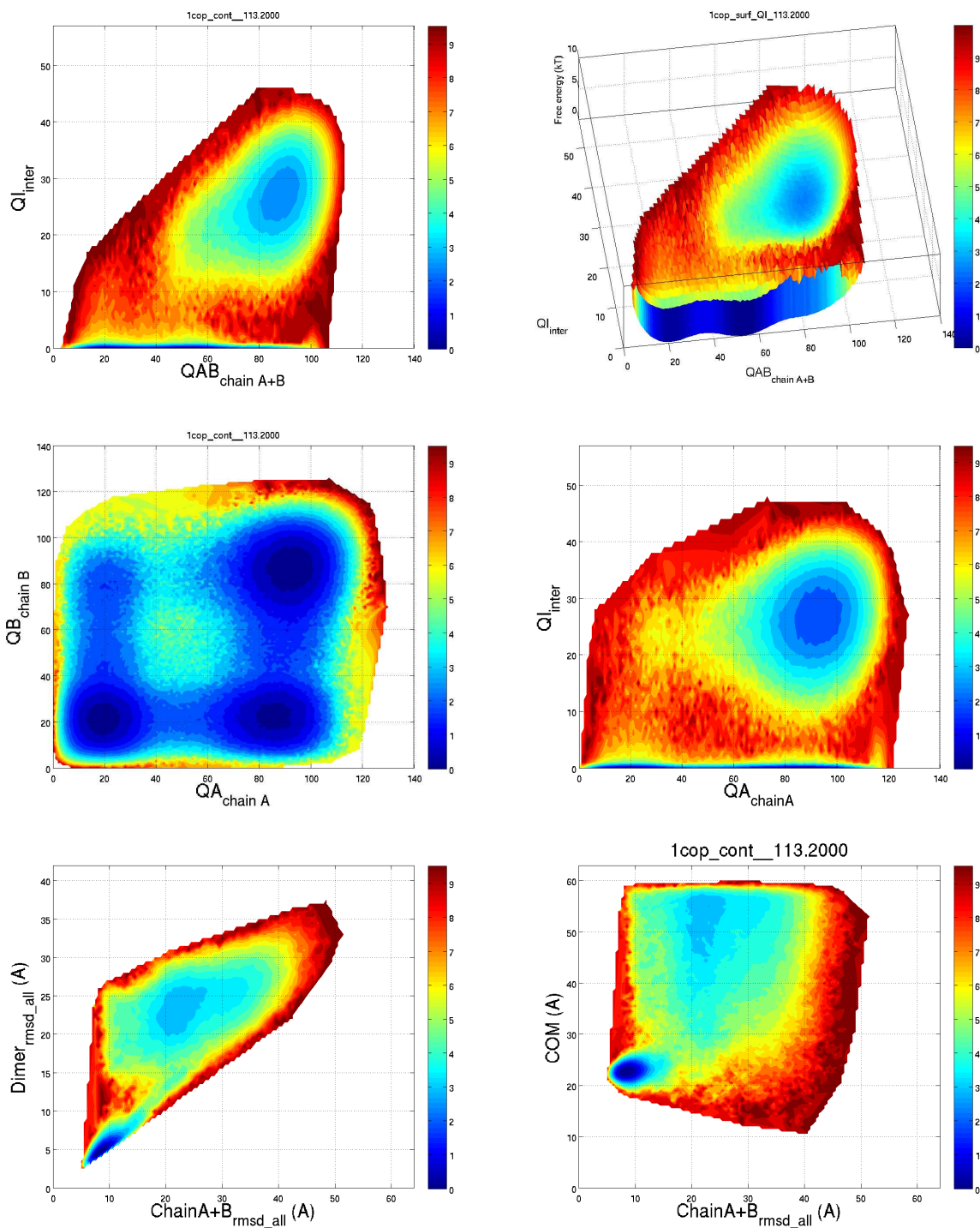


Figure 6.8: 1cop; Lambda Cro repressor. Although this dimer clearly possesses at least 3 states in free energy contours, the heat capacity curve only have one dominant peak. This may suggest binding and folding happen successively in time.

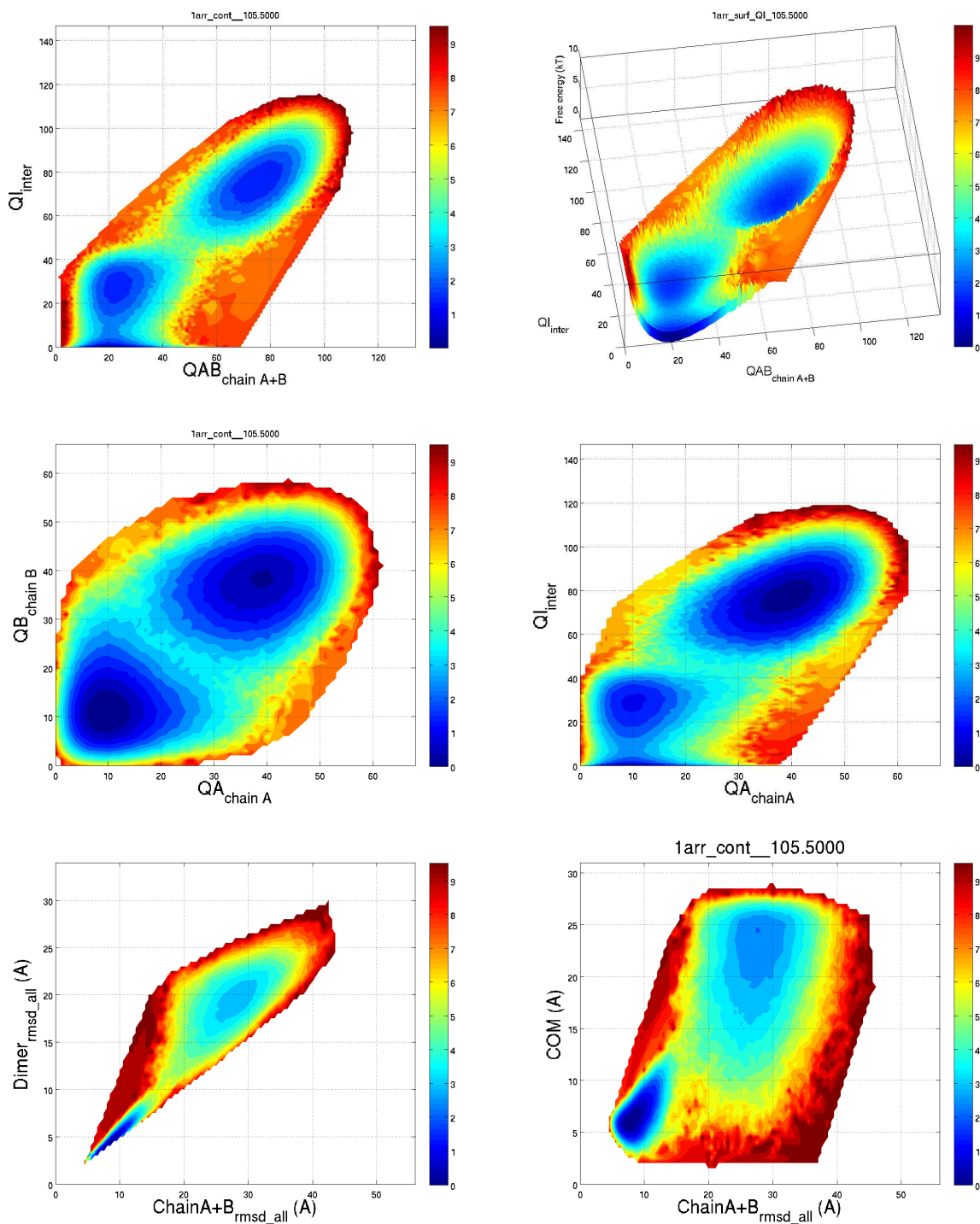


Figure 6.9: 1arr; Arc repressor. First partially bind, then coupled binding-folding.

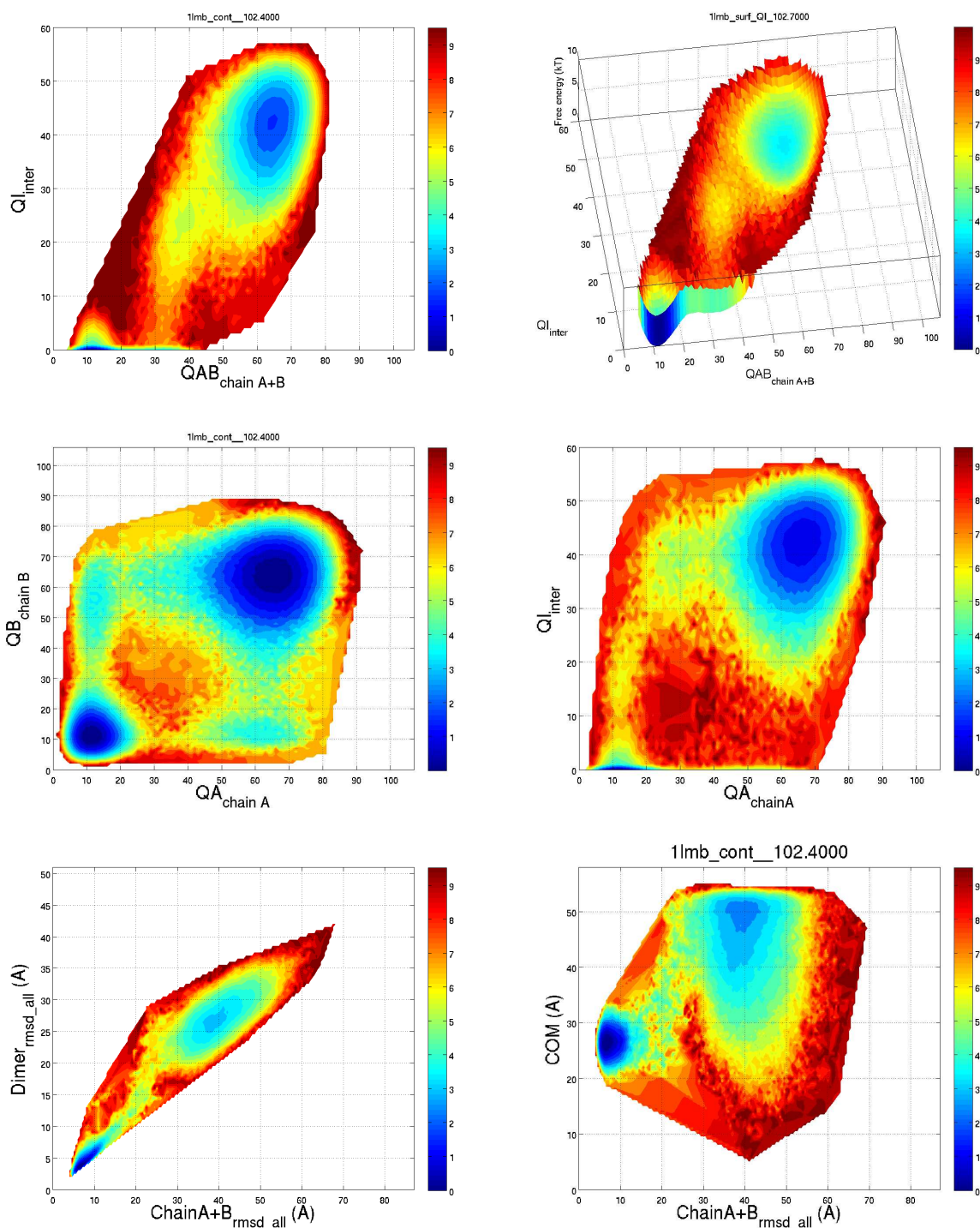


Figure 6.10: 1lmb; Lambda repressor. The 3rd contour may suggest a “fold-then-bind” route; the 4th contour may suggest a “bind-then-fold” route; the 1st and 2nd contours seem to give a “partial fold - bind - total fold” pattern. Combining them together, I would like to hypothesize a scenario: one monomer first fold; cause the second one to bind and formation of binding interface; then the second monomer fold. Note that contour with x-axis as Q_a+Q_b , and is two times scaled down than the one with axis Q_a or Q_b , so 35 on (Q_a+Q_b) axis equivalent to 70 on Q_a or Q_b axis.

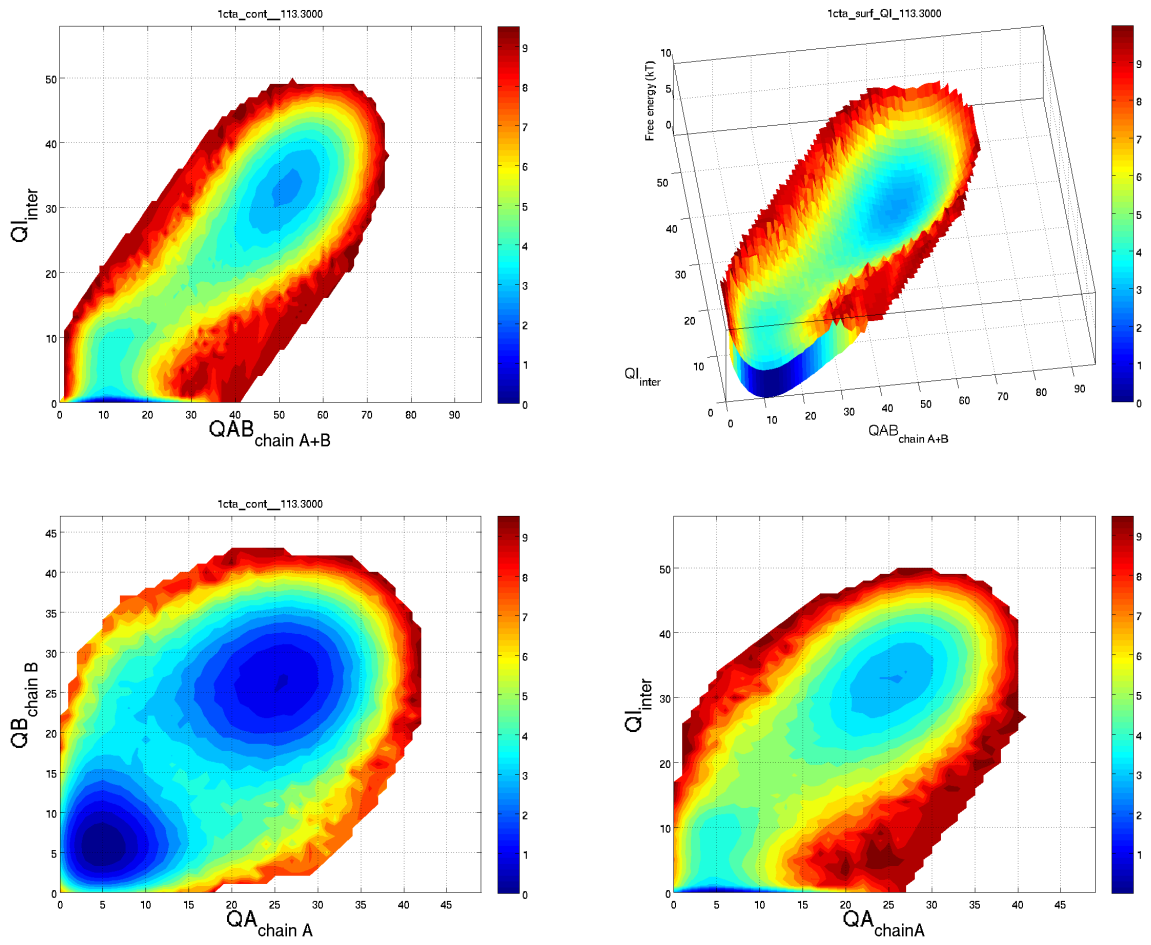


Figure 6.11: 1cta; Troponin C site III. First partially bind, then coupled binding-folding.

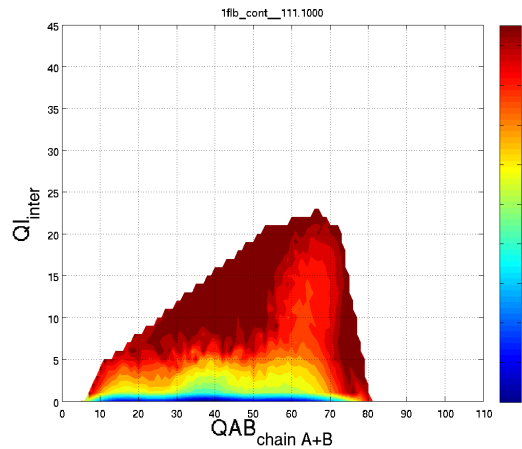
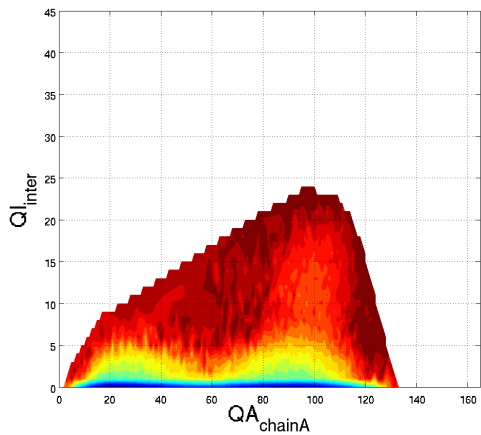
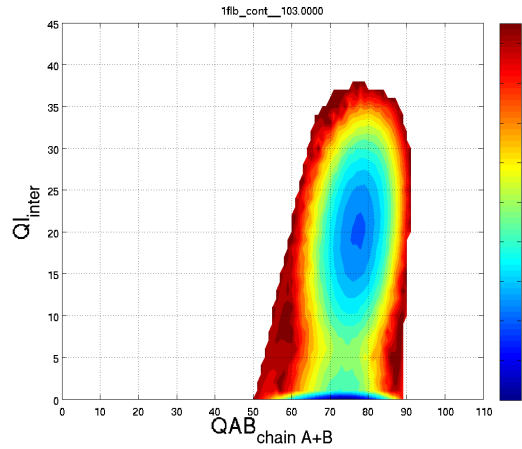
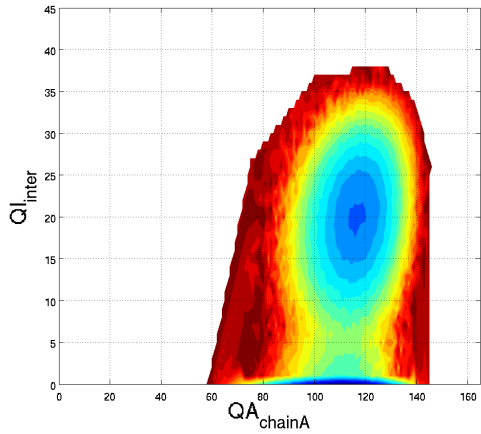
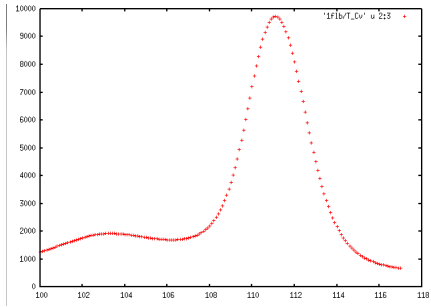


Figure 6.12: 1fb; LFBI transcription factor. The one column top figure is heat capacity curve. The middle two free energy contours show binding process at the first heat capacity peak. The lower two contours, showing folding process, is the cause of the second peak. Notice that the lower right contour gives three narrow basins, indicating the two monomers unfold at different time.

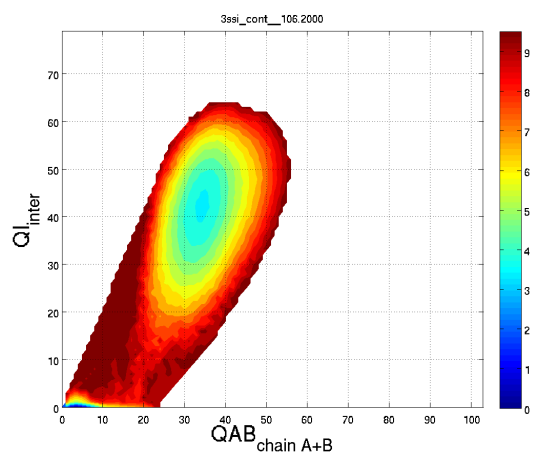
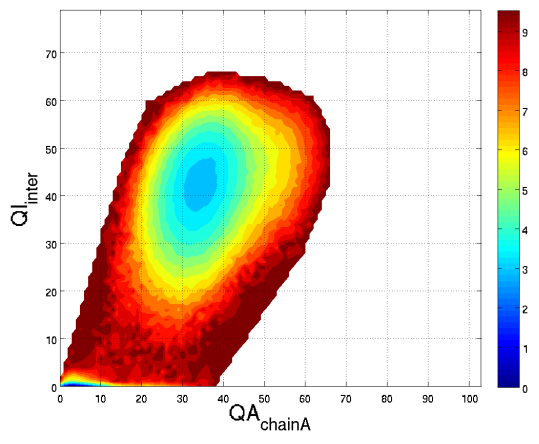
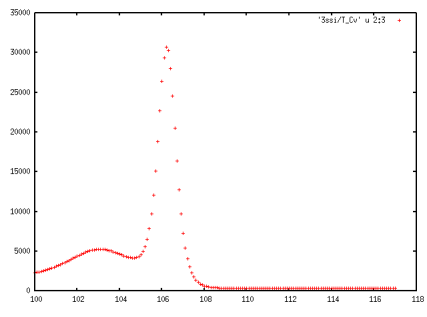


Figure 6.13: 3ssi; *Streptomyces subtilisin*. The one column top figure is heat capacity curve. At the temperature of the second (dominant) peak, the two free energy contours here reveal coupled binding-folding process, is the cause of the second peak.

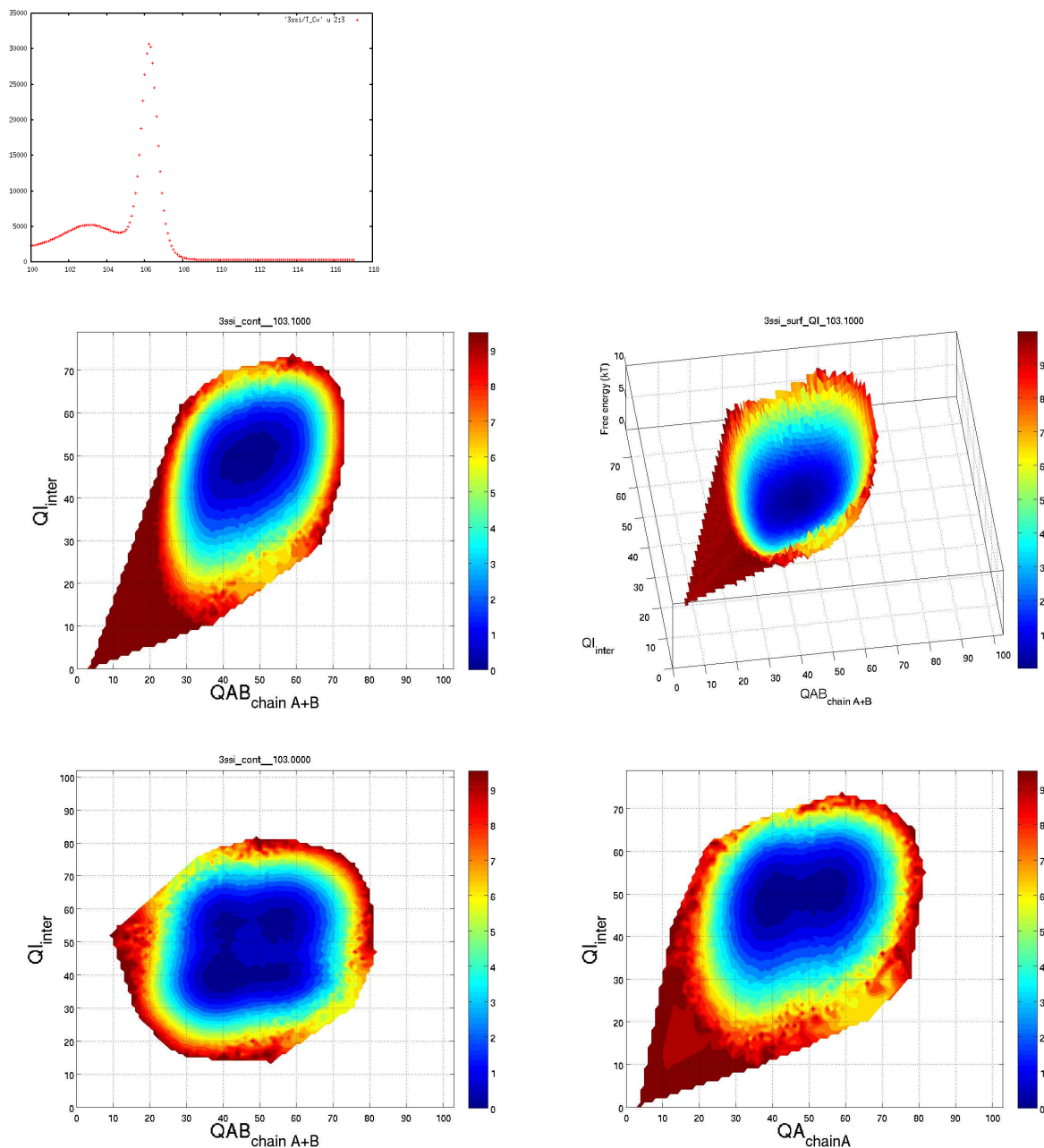


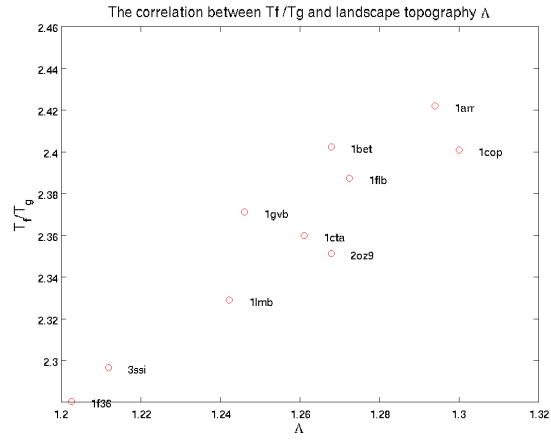
Figure 6.14: 3ssi; *Streptomyces subtilisin*. The one column top figure is heat capacity curve. The lower left free energy contour indicates a slightly improved folding of chain A and chain B at first heat capacity peak (the lower peak); but this slight folding improvement is not captured by using Q_a+Q_b as reaction coordinate in the middle contours.

6.3 Correlation between λ and thermodynamic stability

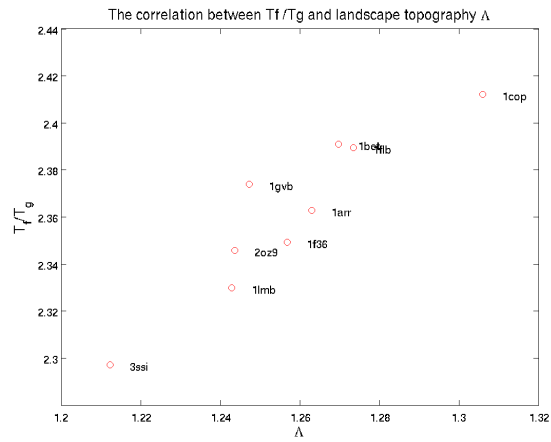
Three 2 dimensional sets of reaction coordinate (ordered parameter) are used in calculating Λ values, so that comparisons can be made among them. These three reaction coordinate sets are: (a) Q_i, Q_a+Q_b ; (b) Q_i, Q_a ; (c) $\text{RMSD}_a+\text{RMSD}_b$, RMSD of the whole dimer. With any one of

these three choices of reaction coordinate sets, good positive correlation is observed between Λ and thermodynamic stability, measured by (T_f/T_g) (Fig. 6.15). This agrees with expectation, since a large λ value may indicate a deeper native minimum, a smoother funnel surface and/or a smaller funnel mouth, and therefore a more stable and robust folding process.

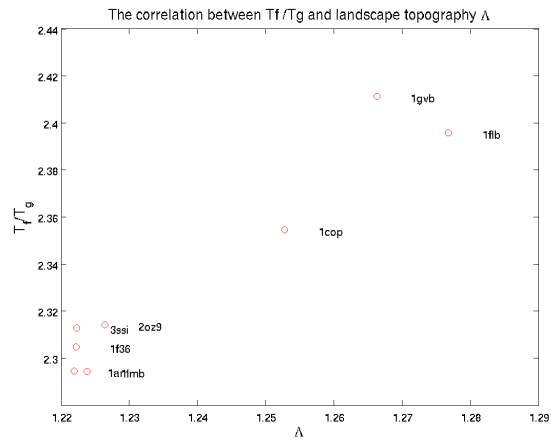
The first two order parameter sets, using Q values, have overall agreement with each other but differs a lot from the third set which uses RMSD values. This gives rise to the question of what is the most appropriate order parameter to capture topography ratio λ . We later decided to use only Q because it can distinguish binding (Q_i) from folding (Q_a and Q_b) most efficiently.



(a) Q_i, Q_a+Q_b



(b) Q_i, Q_a



(c) RMSDa+RMSDb, RMSD all.

Figure 6.15: Positive correlation between Δ and thermodynamics stability (T_f/T_g) is observed with all three sets of order parameters. (a) Q_i, Q_a+Q_b ; (b) Q_i, Q_a ; (c) RMSDa+RMSDb, RMSD of the whole dimer.

7 Kinetics about binding-folding

The binding-folding speed of our dimers as a function of temperature generally follow a skewed/asymmetric U shape. It starts with very slow speed at low temperature end, especially when it is below the glass trapping transition temperature, and gradually gets faster and faster as temperature increases. This increase of binding-folding speed along temperature stops when it is approaching to or going beyond binding-folding transition temperature, where folded-binded native state has become significantly less favored thermodynamically. These two high ends of the U shaped curves can be expected since in the extreme cases, i.e. far above denature temperature or far below glass trapping temperature, folding/binding can never occur.

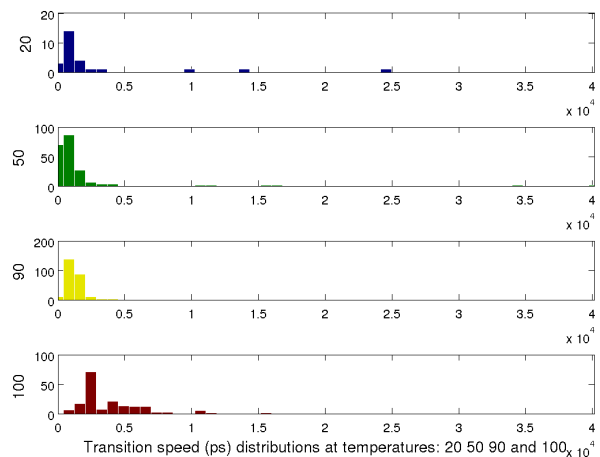
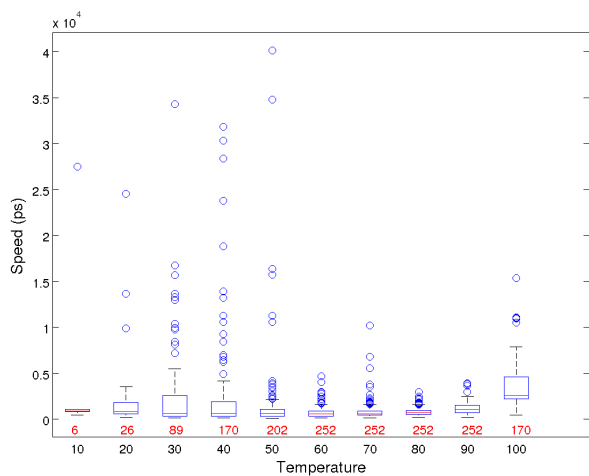
The cause of slow binding-folding speed at low temperature end is the reduced thermal fluctuation/flexibility. Small local free energy minima traps which are easy to break out at a warmer temperature become more retarded when the thermal fluctuations are reduced. When it is below the glass trapping temperature, the trapping from local minima becomes even more serious that some traps are just impossible to escape from. We can see from the box-whisker plots on the left column that even at temperature 10, which is well below glass trapping temperatures of all dimers (around 30 to 50), few folding events are still observed. These rare folding events may result from the existing chances that the initial configuration of a dimer happens to be around its native basin, and thus has no necessity to cross any barrier or escape from any local trap. This also helps to rationalize that once these rare folding events occurs from very lucky initial configurations, they can happen very fast (e.g. 1arr, 1f36 and 1cta at temperature 10). But given the small counts of these events (only a very small percent (<5%) of initial structures get folded at temperature 10), we can see the folding is extremely hard to happen below glass trapping temperature. With these low count numbers (i.e. from 2 to 11), even survival statistic analysis is impossible to perform; but we can expect very slow speed for most initial configurations and a huge variance/fluctuation here.

The glass trapping temperature for most dimers are around 30 to 50. We can see that all five monomers have a huge increase in the the number of folding events observed between temperature 30 and 40, or 40 and 50.

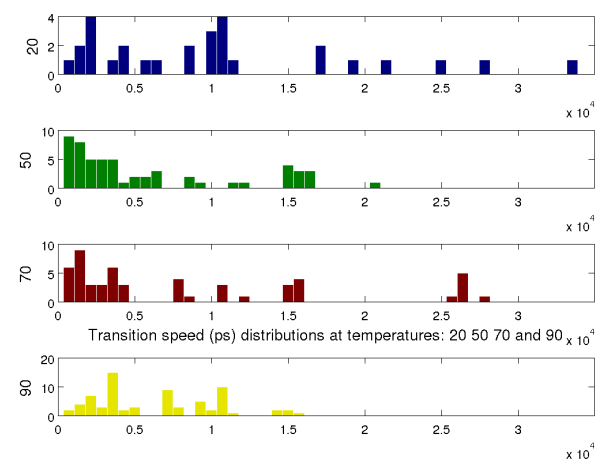
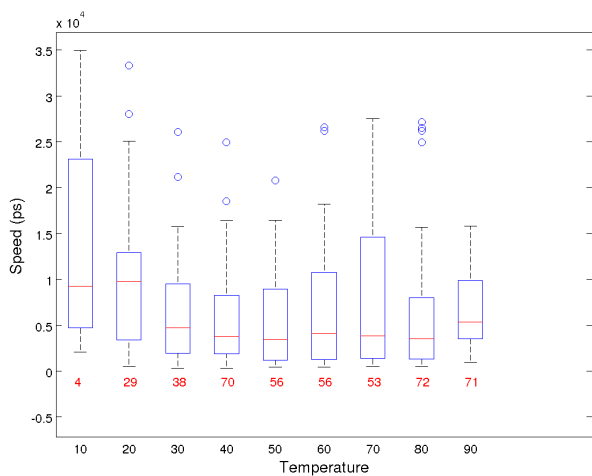
At the high temperature end, close to binding-folding transition temperature, accessible configuration space needed to be searched suddenly increases greatly. The searching speed, and the flexibility of the peptide chains, also increase. The increase of configuration space dominates and lead to eventually slower folding speed. A larger variance and smaller number of folding events are also observed, likely caused by the same reason, i.e. the expansion of accessible configurational space.

Another interesting feature we can observe by comparing the speed distribution histograms among dimers is that, 3 state dimers (1lmb and 1cop) have bigger variances (more fluctuations) than two state dimers (1arr, 1f36 and 1cta). Histograms of 3 state dimers have a much broader distribution, which infers bigger variance in binding-folding transition speed. This is not a surprise because there are two transitions involved in 3 state dimers, which requires crossing two barriers successively. 3 state dimers can pass the first transition, reach its intermediate state and then turn back to its denatured state, instead of proceeding to the second transition; while 2 state dimers can't have this behavior.

1arr



1mb



1cop

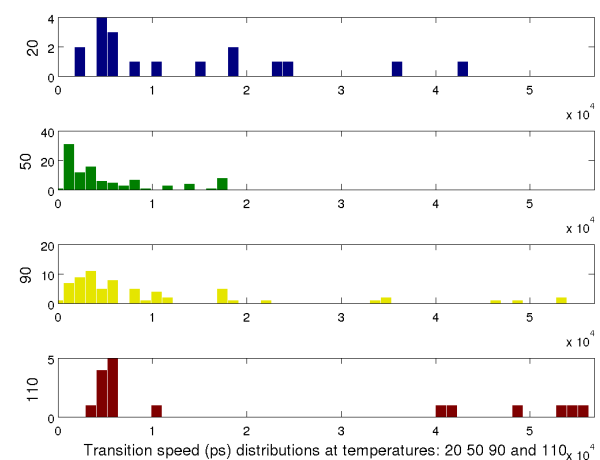
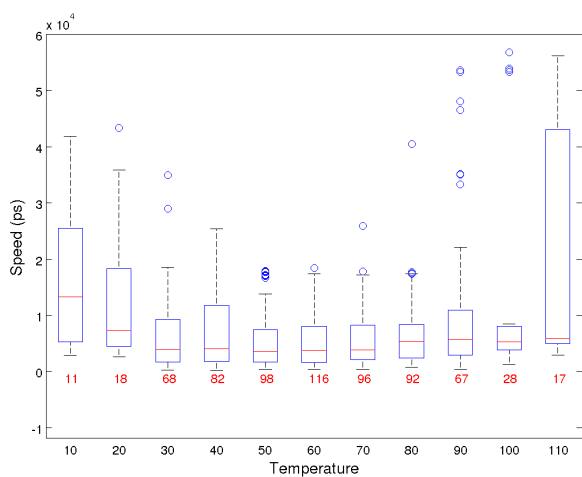
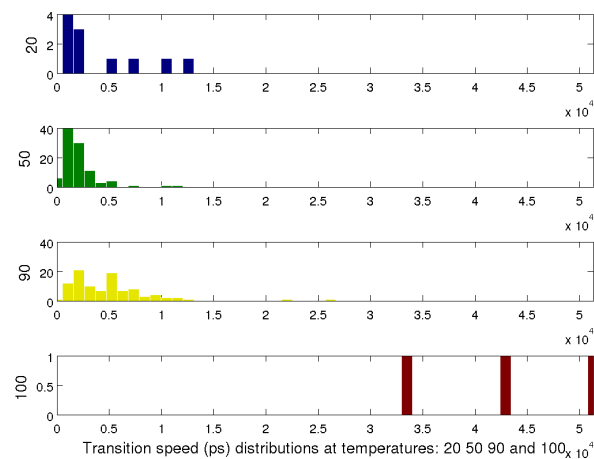
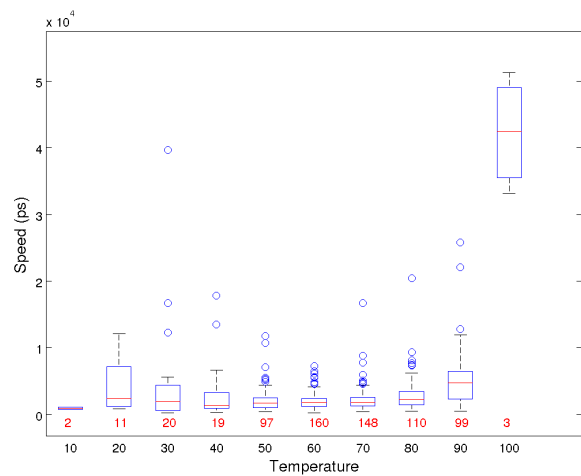


Figure 7.1: Binding-folding Speed (part I)

1f36



1cta

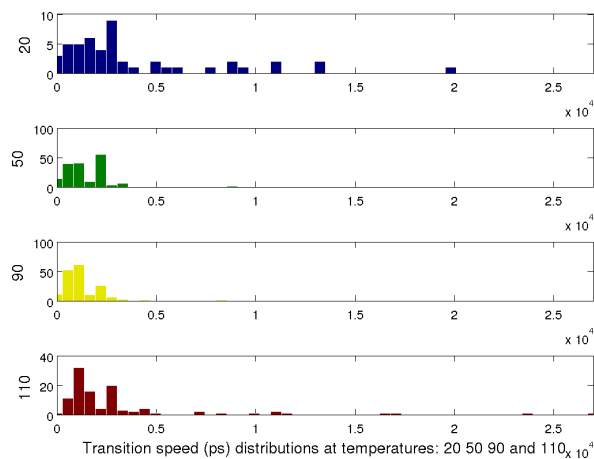
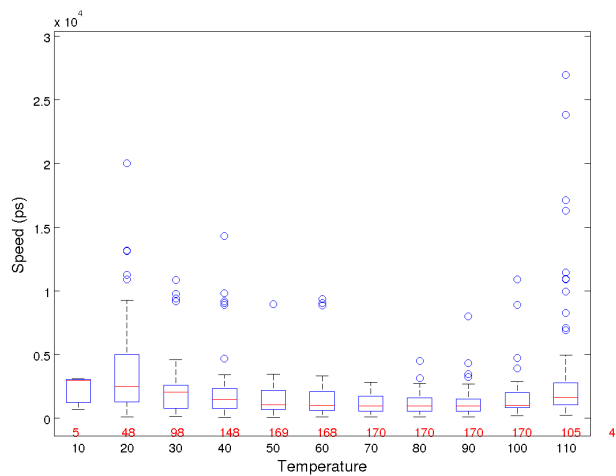


Figure 7.2: Binding-folding Speed (part II)

8 Discussion

8.1 Structure-based model with minimal frustration

The artificially defined native interaction between two spatially close atoms seems to render the structure-based model at the risk of being unrealistic. But the underlying supporting and ratification for this model is the minimally frustrated intrinsic funneled landscape of folding and binding. The assumption behind is that evolution need to eliminate the destructive energetic frustrations that could possibly delay or trap folding/binding processes. The native contacts interaction defined in structure-based model may not represent the real interaction between two atoms, like charge/dipole interaction or hydrophobic packing, but they incarnate the overall driving force/bias toward folded/bound state.

Also, apart from energetic terms, topology of a protein dimer has been demonstrated to have a bigger influence on its dimerization mechanism [22, 21]. Structure-based model keep the topology of a protein structure. And in our atomic-level, the representation of topology is much better than the residue level model, including the topological frustrations from amino acid side chains.

8.2 Other factors influencing protein dimerization

8.2.1 Involvement of DNA molecule

Half of the dimers we studied here are transcription factors. Their natural functionality involves DNA binding, and their dimerization processes may be possibly concurrent with DNA binding equilibrium. In order to bind to DNA, these dimers possess extra positive charges to help them bind to negatively charged phosphate groups on DNA molecules. These positive charges can prevent two subunits from coming together without the presence of DNA, but this effect is not modeled here in this work. Our atomic level structure-based model does not consider charges at this time. However, most of the dimers (9 out of 10) studied here exist and have been determined by X-ray crystallography or NMR without the presence of DNA. In this sense, we can study the binding-folding mechanism of these dimeric proteins alone without the complexities of introducing DNA motifs.

The only exception is lambda repressor (PDB ID 1lmb), which is from a crystal structure with bound DNA in it. We just grab the protein part from it. However, it has been shown that the dimerization of lambda repressor and its binding to DNA are coupled equilibria and the intermediate quaternary structure contributes to operator affinity of lambda repressor [5]. So, cautions should be applied when interpreting the mechanisms demonstrated in our study, especially when their natural binding partners, like DNA sequences, are involved.

8.2.2 Co-translational folding/binding

Translation rate in protein synthesis is about 50 amino acids per second in prokaryotes and 5 amino acids per second in eukaryotes. The folding time scale of the small protein dimers studied here is likely at milliseconds scale. So co-translational folding should be expected. For 3 state dimers, which have a stable folded monomeric form, i.e. non-obligatory dimers, the folding of a monomer may happen right during the translation process. While for 2 state (obligatory) dimers,

which are basically IDPs, the polyribosome transcription process can increase the local concentrations of the unfolded monomer, and facilitate coupled binding-folding. Of course, the ribosome surface and crowded cellular environment can also affect the binding-folding process of protein dimers, which are not included in this study.

8.2.3 Chaperone

Protein folding events may need assistance in the crowded in vivo cellular environment, especially for heat/stress-induced unfolded proteins or multi-domain big proteins. For the relatively small protein dimers studied here, chaperones may not be necessary for assisting folding. But since the monomers from a 2 state dimer are unstructured before binding, chaperone may be required to prevent the aggregation of them. Also, since chaperones are also involved in assisting the proper assembly of macromolecular complexes, including both proteins and DNA, they could be possibly involved in the dimerization of transcription factor dimers, and their association with DNA operators.

8.2.4 Slower dimerization $F + U \rightleftharpoons D$

The dissociation constant of Lambda Cro repressor (1cop) is much higher than the dissociation constant of the complex of Cro dimer and its operator DNA sequence. And its dimerization process is relatively "weak and slow" [76, 75, 73]. The model we proposed can partially explain this behavior. At a fixed total Cro protein concentration ($[F] + [U] = \text{const.}$), reaction rate following a second order model is $k[F][U]$, which reaches its peak at the concentration where $[F] = [U]$, which can be hardly satisfied at physiological conditions. Though the transition barrier height between states F and U is only around 1~2 kT at transition temperature, at temperatures below transition temperature, the folded state F will become dominantly populated. More rigorous reaction kinetics about this set of two transitions can be modeled in the future.

9 Summary

In this study, new simulation methods are applied to explore the dimerization mechanisms of regulatory protein dimers. We focus on the relationship between two important aspects of dimerization: monomer folding and interface binding. Through our multiple characterizing and analyzing approaches, new dimerization mechanisms are revealed. Meanwhile, we realize that the conventional terminology used in the description of dimerization mechanisms are not specific/clear enough, and to some extent even misleading. Through an example with detailed illustrations, we discussed the subtle distinctions among these terms, like 2-state dimer, 3-state dimer, obligatory or non-obligatory dimer, and coupled/decoupled binding-folding process.

Equipped with all-atom structure-based model and funneled landscape theory, we can possibly bridge the timescale gap in simulating protein folding/binding. With 3 dimensional free energy profiles derived from efficient replica exchange sampling, diverse dimerization mechanisms are revealed. The binding-folding relationship of protein homodimers is actually more complicated than we previous thought. Current classification method for dimerization mechanisms can not handle/include the new scenarios discovered in this work. And these newly emerging scenarios from our simulation have good agreement with existing experimental observations, and can even reconcile some conflicting experimental results.

The advantages of all-atom structure-based model are demonstrated in this work. Atomic level granulation has been shown to possess a better representation of hydrophobic packing, and hydrogen bonding. Since the atomic level granulation can realize/materialize these interactions, given the underlying shadow map contact definition method has successfully identified these interaction pairs as atomic-level native contacts. The geometric shape of each residue, as well as the topological information are preserved in this model. Though there is no energy roughness in structure-based model, the topological roughness are better represented here than in course-grained (residue level) models. For two dimers, our model gives better agreement with experiments, and corrects the their classification from residue-level simulations. However, we still can not distinguish different kinds of native interactions in our model. New development of this model will be to include electrostatic interactions and introduce non-native energetic frustrations.

References

- [1] D. S. Goodsell and A. J. Olson. Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct*, 29:105–153, 2000.
- [2] K. E. Neet and D. E. Timm. Conformational stability of dimeric proteins: quantitative studies by equilibrium denaturation. *Protein Sci*, 3(12):2167–2174, Dec 1994.
- [3] Jessica A O. Rumfeldt, Celine Galvagnion, Kenrick A. Vassall, and Elizabeth M. Meiering. Conformational stability and folding mechanisms of dimeric proteins. *Prog Biophys Mol Biol*, 98(1):61–84, Sep 2008.
- [4] J. D. Klemm, S. L. Schreiber, and G. R. Crabtree. Dimerization as a regulatory mechanism in signal transduction. *Annu Rev Immunol*, 16:569–592, 1998.

- [5] M. A. Weiss, C. O. Pabo, M. Karplus, and R. T. Sauer. Dimerization of the operator binding domain of phage lambda repressor. *Biochemistry*, 26(3):897–904, Feb 1987.
- [6] K. A. Lee. Dimeric transcription factor families: it takes two to tango but who decides on partners and the venue? *J Cell Sci*, 103 (Pt 1):9–14, Sep 1992.
- [7] Edwin Li and Kalina Hristova. Receptor tyrosine kinase transmembrane domains: Function, dimer structure and dimerization energetics. *Cell Adh Migr*, 4(2):249–254, 2010.
- [8] Fischer E. Einfluss der configuration auf die wirkung der enzyme. *Ber Dtsch Chem Ges*, 27(3):2984–2993, 1894.
- [9] H. R. Bosshard. Molecular recognition by induced fit: how fit is the concept? *News Physiol Sci*, 16:171–173, Aug 2001.
- [10] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci U S A*, 44(2):98–104, Feb 1958.
- [11] Peter Csermely, Robin Palotai, and Ruth Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem Sci*, 35(10):539–546, Oct 2010.
- [12] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol*, 293(2):321–331, Oct 1999.
- [13] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic. Intrinsically disordered protein. *J Mol Graph Model*, 19(1):26–59, 2001.
- [14] Vladimir N. Uversky. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci*, 11(4):739–756, Apr 2002.
- [15] B.A. Shoemaker, J.J. Portman, and P.G. Wolynes. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proceedings of the National Academy of Sciences*, 97(16):8868–8873, 2000.
- [16] Xiakun Chu, Yong Wang, Linfeng Gan, Yawen Bai, Wei Han, Erkang Wang, and Jin Wang. Importance of electrostatic interactions in the association of intrinsically disordered histone chaperone chz1 and histone h2a.z-h2b. *PLoS Comput Biol*, 8(7):e1002608, 2012.
- [17] J. Wang and G.M. Verkhivker. Energy landscape theory, funnels, specificity, and optimal criterion of biomolecular binding. *Physical review letters*, 90(18):188101, 2003.
- [18] J. Wang, L. Xu, and E. Wang. Optimal specificity and function for flexible biomolecular recognition. *Biophysical journal*, 92(12):L109–L111, 2007.

- [19] Yong Wang, Xiakun Chu, Sonia Longhi, Philippe Roche, Wei Han, Erkang Wang, and Jin Wang. Multiscaled exploration of coupled folding and binding of an intrinsically disordered molecular recognition element in measles virus nucleoprotein. *Proc Natl Acad Sci U S A*, 110(40):E3743–E3752, Oct 2013.
- [20] S. L. Shammass, J. M. Rogers, S. A. Hill, and J. Clarke. Slow, reversible, coupled folding and binding of the spectrin tetramerization domain. *Biophys J*, 103(10):2203–2214, Nov 2012.
- [21] Y. Levy, P.G. Wolynes, and J.N. Onuchic. Protein topology determines binding mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 101(2):511–516, 2004.
- [22] Y. Levy, S.S. Cho, J.N. Onuchic, P.G. Wolynes, et al. A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes. *Journal of molecular biology*, 346(4):1121–1146, 2005.
- [23] Y. Levy, G.A. Papoian, J.N. Onuchic, and P.G. Wolynes. Energy landscape analysis of protein dimers. *Israel journal of chemistry*, 44(1-3):281–297, 2004.
- [24] Xiakun Chu, Linfeng Gan, Erkang Wang, and Jin Wang. Quantifying the topography of the intrinsic energy landscape of flexible biomolecular recognition. *Proc Natl Acad Sci U S A*, 110(26):E2342–E2351, Jun 2013.
- [25] Jin Wang, Yong Wang, Xiakun Chu, Stephen J. Hagen, Wei Han, and Erkang Wang. Multi-scaled explorations of binding-induced folding of intrinsically disordered protein inhibitor ia3 to its target enzyme. *PLoS Comput Biol*, 7(4):e1001118, Apr 2011.
- [26] J.D. Bryngelson and P.G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences*, 84(21):7524–7528, 1987.
- [27] P.E. Leopold, M. Montal, and J.N. Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences*, 89(18):8721–8725, 1992.
- [28] H.S. Chan and K.A. Dill. Transition states and folding dynamics of proteins and heteropolymers. *The Journal of chemical physics*, 100:9238, 1994.
- [29] J. Wang, J. Onuchic, and P. Wolynes. Statistics of kinetic pathways on biased rough energy landscapes with applications to protein folding. *Physical review letters*, 76(25):4861–4864, 1996.
- [30] Jin; Wolynes Peter G. Plotkin, Steven S.; Wang. Statistical mechanics of a correlated energy landscape model for protein folding funnels. *Journal of Chemical Physics*, 106:2932, 1997.
- [31] K.A. Dill, H.S. Chan, et al. From levinthal to pathways to funnels. *Nature structural biology*, 4(1):10–19, 1997.

- [32] J.S. Milne, Y. Xu, L.C. Mayne, and S.W. Englander. Experimental study of the protein folding landscape: unfolding reactions in cytochrome c. *Journal of molecular biology*, 290(3):811–822, 1999.
- [33] J.N. Onuchic and P.G. Wolynes. Theory of protein folding. *Current opinion in structural biology*, 14(1):70–75, 2004.
- [34] Joseph D Bryngelson and Peter G Wolynes. Intermediates and barrier crossing in a random energy model (with applications to protein folding). *The Journal of Physical Chemistry*, 93(19):6902–6915, 1989.
- [35] Peter G Wolynes, Jose N Onuchic, and D Thirumalai. Navigating the folding routes. *SCIENCE-NEW YORK THEN WASHINGTON-*, pages 1619–1619, 1995.
- [36] CL Brooks, José N Onuchic, and David J Wales. Statistical thermodynamics: taking a walk on a landscape. *SCIENCE-NEW YORK THEN WASHINGTON-*, pages 612–613, 2001.
- [37] C. Levinthal. Mossbauer spectroscopy in biological systems. In *Proceedings of a meeting held at Allerton House. P. Debrunner, JCM Tsibris, and E. Munck, editors. University of Illinois Press, Urbana, IL, 1969.*
- [38] P. G. Wolynes. Recent successes of the energy landscape theory of protein folding and function. *Q Rev Biophys*, 38(4):405–410, Nov 2005.
- [39] Alexander Schug and José N. Onuchic. From protein folding to protein function and biomolecular binding by energy landscape theory. *Curr Opin Pharmacol*, 10(6):709–714, Dec 2010.
- [40] N. Go. Theoretical studies of protein folding. *Annual review of biophysics and bioengineering*, 12(1):183–210, 1983.
- [41] P.C. Whitford, J.K. Noel, S. Gosavi, A. Schug, K.Y. Sanbonmatsu, and J.N. Onuchic. An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. *Proteins: Structure, Function, and Bioinformatics*, 75(2):430–441, 2008.
- [42] Yaakov Levy and Amedeo Caflisch. Flexibility of monomeric and dimeric hiv-1 protease. *The Journal of Physical Chemistry B*, 107(13):3068–3079, 2003.
- [43] S. Yang, S.S. Cho, Y. Levy, M.S. Cheung, H. Levine, P.G. Wolynes, and J.N. Onuchic. Domain swapping is a consequence of minimal frustration. *Proceedings of the National Academy of Sciences of the United States of America*, 101(38):13786–13791, 2004.
- [44] Yong Wang, Xiakun Chu, Zucui Suo, Erkang Wang, and Jin Wang. Multidomain protein solves the folding problem by multifunnel combined landscape: theoretical investigation of a y-family dna polymerase. *Journal of the American Chemical Society*, 134(33):13755–13764, 2012.

- [45] J. Wang, R.J. Oliveira, X. Chu, P.C. Whitford, J. Chahine, W. Han, E. Wang, J.N. Onuchic, and V.B.P. Leite. Topography of funneled landscapes determines the thermodynamics and kinetics of protein folding. *Proceedings of the National Academy of Sciences*, 109(39):15763–15768, 2012.
- [46] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, Aug 1990.
- [47] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, 93(1):13–20, Jan 1996.
- [48] L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285(5):2177–2198, Feb 1999.
- [49] Yaakov Levy and José N. Onuchic. Mechanisms of protein assembly: lessons from minimalist models. *Acc Chem Res*, 39(2):135–142, Feb 2006.
- [50] Yaakov Levy, José N Onuchic, and Peter G Wolynes. Fly-casting in protein-dna binding: frustration between protein folding and electrostatics facilitates target recognition. *Journal of the American Chemical Society*, 129(4):738–739, 2007.
- [51] J.K. Noel, P.C. Whitford, K.Y. Sanbonmatsu, and J.N. Onuchic. Smog@ ctbp: simplified deployment of structure-based models in gromacs. *Nucleic acids research*, 38(suppl 2):W657–W661, 2010.
- [52] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1):141–151, 1999.
- [53] S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, and P.A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.
- [54] Ferrenberg and Swendsen. Optimized monte carlo data analysis. *Phys Rev Lett*, 63(12):1195–1198, Sep 1989.
- [55] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J C. Berendsen. Gromacs: fast, flexible, and free. *J Comput Chem*, 26(16):1701–1718, Dec 2005.
- [56] G. Bussi C. Camilloni D. Provasi P. Raiteri D. Donadio F. Marinelli F. Pietrucci R.A. Broglia M. Bonomi, D. Branduardi and M. Parrinello. Plumed: a portable plugin for free energy calculations with molecular dynamics. *Comp. Phys. Comm.*, 180:1961, 2009.
- [57] J.K. Noel, P.C. Whitford, and J.N. Onuchic. The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function. *The Journal of Physical Chemistry B*, 116(29):8692–8702, 2012.
- [58] ND Socci, J.N. Onuchic, and P.G. Wolynes. Diffusive dynamics of the reaction coordinate for protein folding funnels. *arXiv preprint cond-mat/9601091*, 1996.

- [59] J. Sabelko, J. Ervin, and M. Gruebele. Observation of strange kinetics in protein folding. *Proceedings of the National Academy of Sciences*, 96(11):6031–6036, 1999.
- [60] DK Klimov and D. Thirumalai. Criterion that determines the foldability of proteins. *Physical review letters*, 76(21):4070–4073, 1996.
- [61] Garegin A. Papoian and Peter G. Wolynes. The physics and bioinformatics of binding and folding—an energy landscape perspective. *Biopolymers*, 68(3):333–349, Mar 2003.
- [62] S. Kumar, B. Ma, C. J. Tsai, N. Sinha, and R. Nussinov. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci*, 9(1):10–19, Jan 2000.
- [63] C. J. Tsai, S. Kumar, B. Ma, and R. Nussinov. Folding funnels, binding funnels, and protein function. *Protein Sci*, 8(6):1181–1190, Jun 1999.
- [64] A. D. Johnson, C. O. Pabo, and R. T. Sauer. Bacteriophage lambda repressor and cro protein: interactions with operator dna. *Methods Enzymol*, 65(1):839–856, 1980.
- [65] A. D. Johnson, A. R. Poteete, G. Lauer, R. T. Sauer, G. K. Ackers, and M. Ptashne. lambda repressor and cro—components of an efficient molecular switch. *Nature*, 294(5838):217–223, Nov 1981.
- [66] D. H. Ohlendorf, W. F. Anderson, Y. Takeda, and B. W. Matthews. High resolution structural studies of cro repressor protein and implications for dna recognition. *J Biomol Struct Dyn*, 1(2):553–563, Oct 1983.
- [67] W. F. Anderson, D. H. Ohlendorf, Y. Takeda, and B. W. Matthews. Structure of the cro repressor from bacteriophage lambda and its interaction with dna. *Nature*, 290(5809):754–758, Apr 1981.
- [68] P. Leighton and P. Lu. Lambda cro repressor complex with or3 dna: 15n nmr observations. *Biochemistry*, 26(23):7262–7271, Nov 1987.
- [69] H. Matsuo, M. Shirakawa, and Y. Kyogoku. Three-dimensional dimer structure of the lambda-cro repressor in solution as determined by heteronuclear multidimensional nmr. *J Mol Biol*, 254(4):668–680, Dec 1995.
- [70] I. A. Bolotina, A. V. Kurochkin, and M. P. Kirpichnikov. Studies of the structure of bacteriophage lambda cro protein in solution. analysis of the circular dichroism data. *FEBS Lett*, 155(2):291–294, May 1983.
- [71] A. A. Pakula and R. T. Sauer. Amino acid substitutions that increase the thermal stability of the lambda cro protein. *Proteins*, 5(3):202–210, 1989.
- [72] G. I. Gitelson, V. Griko Yu, A. V. Kurochkin, V. V. Rogov, V. P. Kutysenko, M. P. Kirpichnikov, and P. L. Privalov. Two-stage thermal unfolding of cys55-substituted cro repressor of bacteriophage lambda. *FEBS Lett*, 289(2):201–204, Sep 1991.

- [73] Rinku Jana, Tony R Hazbun, AKMM Mollah, and Michael C Mossing. A folded monomeric intermediate in the formation of lambda cro dimer-dna complexes. *Journal of molecular biology*, 273(2):402–416, 1997.
- [74] W John Satumba and Michael C. Mossing. Folding and assembly of lambda cro repressor dimers are kinetically limited by proline isomerization. *Biochemistry*, 41(48):14216–14224, Dec 2002.
- [75] P. J. Darling, J. M. Holt, and G. K. Ackers. Coupled energetics of lambda cro repressor self-assembly and site-specific dna operator binding ii cooperative interactions of cro dimers. *J Mol Biol*, 302(3):625–638, Sep 2000.
- [76] Haifeng Jia, W John Satumba, Gene L Bidwell, 3rd, and Michael C. Mossing. Slow assembly and disassembly of lambda cro repressor dimers. *J Mol Biol*, 350(5):919–929, Jul 2005.
- [77] R. Jana, T. R. Hazbun, J. D. Fields, and M. C. Mossing. Single-chain lambda cro repressors confirm high intrinsic dimer-dna affinity. *Biochemistry*, 37(18):6446–6455, May 1998.
- [78] A. J. Hubbard, L. P. Bracco, S. J. Eisenbeis, R. B. Gayle, G. Beaton, and M. H. Caruthers. Role of the cro repressor carboxy-terminal domain and flexible dimer linkage in operator and nonspecific dna binding. *Biochemistry*, 29(39):9241–9249, Oct 1990.
- [79] Jeffrey K Noel, Paul C Whitford, and Jose N Onuchic. The shadow map: A general contact definition for capturing the dynamics of biomolecular folding and function. *The Journal of Physical Chemistry B*, 116(29):8692–8702, 2012.
- [80] Lee S. J. Yamamoto K. Takimoto M. Akutsu H. & Kyogoku Y. Shirakawa, M. Interaction of lambda cro repressor with operator dna and induced conformational change. *In Structure and Expression*, 1,:167?179, 1987.
- [81] A. K. Mollah, M. A. Aleman, R. A. Albright, and M. C. Mossing. Core packing defects in an engineered cro monomer corrected by combinatorial mutagenesis. *Biochemistry*, 35(3):743–748, Jan 1996.
- [82] R. A. Albright, M. C. Mossing, and B. W. Matthews. High-resolution structure of an engineered cro monomer shows changes in conformation relative to the native dimer. *Biochemistry*, 35(3):735–742, Jan 1996.
- [83] M. C. Mossing and R. T. Sauer. Stable, monomeric variants of lambda cro obtained by insertion of a designed beta-hairpin sequence. *Science*, 250(4988):1712–1715, Dec 1990.
- [84] Emmanuel Trizac, Yaakov Levy, and Peter G Wolynes. Capillarity theory for the fly-casting mechanism. *Proceedings of the National Academy of Sciences*, 107(7):2746–2750, 2010.
- [85] Sarah A. Hobart, Sergey Ilin, Daniel F. Moriarty, Robert Osuna, and Wilfredo Colón. Equilibrium denaturation studies of the escherichia coli factor for inversion stimulation: implications for in vivo function. *Protein Sci*, 11(7):1671–1680, Jul 2002.

- [86] M. K. Safo, W. Z. Yang, L. Corselli, S. E. Cramton, H. S. Yuan, and R. C. Johnson. The transactivation region of the fis protein that controls site-specific dna inversion contains extended mobile beta-hairpin arms. *EMBO J*, 16(22):6860–6873, Nov 1997.
- [87] D. Kostrewa, J. Granzin, D. Stock, H. W. Choe, J. Labahn, and W. Saenger. Crystal structure of the factor for inversion stimulation fis at 2.0 a resolution. *J Mol Biol*, 226(1):209–226, Jul 1992.
- [88] W. S. Tzou and M. J. Hwang. A model for fis n-terminus and fis-invertase recognition. *FEBS Lett*, 401(1):1–5, Jan 1997.
- [89] James U Bowie and Robert T Sauer. Equilibrium dissociation and unfolding of the arc repressor dimer. *Biochemistry*, 28(18):7139–7143, 1989.
- [90] Marcos E Milla, Bronwen M Brown, Carey D Waldburger, and Robert T Sauer. P22 arc repressor: transition state properties inferred from mutational effects on the rates of protein unfolding and refolding. *Biochemistry*, 34(42):13914–13919, 1995.
- [91] RT Sauer, ME Milla, CD Waldburger, BM Brown, and JF Schildbach. Sequence determinants of folding and stability for the p22 arc repressor dimer. *The FASEB journal*, 10(1):42–48, 1996.
- [92] Marcos E Milla and Robert T Sauer. P22 arc repressor: folding kinetics of a single-domain, dimeric protein. *Biochemistry*, 33(5):1125–1133, 1994.
- [93] Matthew HJ Cordes, Nathan P Walsh, C James McKnight, and Robert T Sauer. Solution structure of switch arc, a mutant with 3< sub> 10</sub> helices replacing a wild-type β -ribbon. *Journal of molecular biology*, 326(3):899–909, 2003.
- [94] J. F. Schildbach, M. E. Milla, P. D. Jeffrey, B. E. Raumann, and R. T. Sauer. Crystal structure, folding, and operator binding of the hyperstable arc repressor mutant pl8. *Biochemistry*, 34(4):1405–1412, Jan 1995.
- [95] Wei Yuan Yang and Martin Gruebele. Rate-temperature relationships in lambda-repressor fragment lambda 6-85 folding. *Biochemistry*, 43(41):13018–13025, Oct 2004.
- [96] A. Sarai and Y. Takeda. Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. *Proc Natl Acad Sci U S A*, 86(17):6513–6517, Sep 1989.
- [97] S. R. Jordan and C. O. Pabo. Structure of the lambda complex at 2.5 a resolution: details of the repressor-operator interactions. *Science*, 242(4880):893–899, Nov 1988.
- [98] Weihua Zheng, Nicholas P. Schafer, Aram Davtyan, Garegin A. Papoian, and Peter G. Wolynes. Predictive energy landscapes for protein-protein association. *Proc Natl Acad Sci U S A*, 109(47):19244–19249, Nov 2012.
- [99] H. Sasakawa, A. Tamura, S. Fujimaki, S. Taguchi, and K. Akasaka. Secondary structures and structural fluctuation in a dimeric protein, streptomyces subtilisin inhibitor. *J Biochem*, 126(5):859–865, Nov 1999.

- [100] K. Takahashi and J. M. Sturtevant. Thermal denaturation of streptomyces subtilisin inhibitor, subtilisin bpn', and the inhibitor-subtilisin complex. *Biochemistry*, 20(21):6185–6190, Oct 1981.
- [101] T. Konno, M. Kataoka, Y. Kamatari, K. Kanaori, A. Nosaka, and K. Akasaka. Solution x-ray scattering analysis of cold- heat-, and urea-denatured states in a protein, streptomyces subtilisin inhibitor. *J Mol Biol*, 251(1):95–103, Aug 1995.
- [102] K. Uchida, Y. Miyake, and M. Kainosho. Reductive cleavage and regeneration of the disulfide bonds in streptomyces subtilisin inhibitor (ssi) as studied by the carbonyl ^{13}C nmr resonances of cysteinyl residues. *J Biomol NMR*, 1(1):49–64, May 1991.
- [103] Y. Mitsui, Y. Satow, Y. Watanabe, and Y. Iitaka. Crystal structure of a bacterial protein proteinase inhibitor (streptomyces subtilisin inhibitor) at 2.6 a resolution. *J Mol Biol*, 131(4):697–724, Jul 1979.
- [104] K. Akasaka, T. Inoue, H. Hatano, and C. K. Woodward. Hydrogen exchange kinetics of core peptide protons in streptomyces subtilisin inhibitor. *Biochemistry*, 24(12):2973–2979, Jun 1985.
- [105] L. Tomei, R. Cortese, and R. De Francesco. A pou-a related region dictates dna binding specificity of lfb1/hnf1 by orienting the two xl-homeodomains in the dimer. *EMBO J*, 11(11):4119–4129, Nov 1992.
- [106] Barbara Leiting, Raffaele De Francesco, Licia Tomei, Riccardo Cortese, G Otting, and K Wüthrich. The three-dimensional nmr-solution structure of the polypeptide fragment 195-286 of the lfb1/hnf1 transcription factor from rat liver comprises a nonclassical homeodomain. *The EMBO journal*, 12(5):1797, 1993.

Appendix A

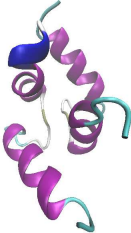
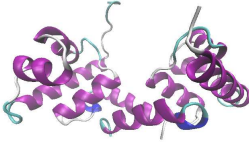
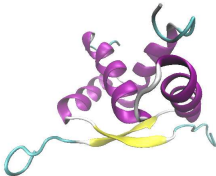
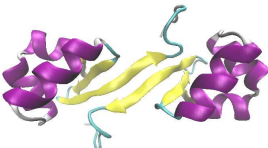
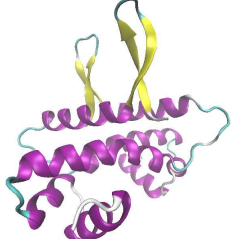
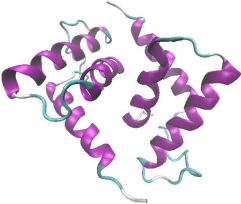
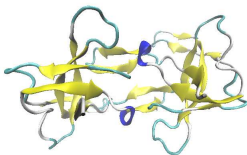
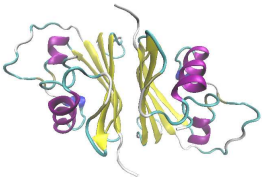


2 state dimers	3 state dimers
	
1cta. Troponin C site III	1lmb. Lambda repressor
	
1arr. Arc repressor	1cop. Lambda Cro repressor
	
1f36. Factor for inversion stimulation	1flb. LFB1 transcription factor
	
1gvb. Gene V protein	3ssi. Streptomyces subtilisin inhibitor
	
1bet. β nerve growth factor	2oz9. Trp repressor

Table 2: PDB ID with protein name and thumbnail of structure.

Appendix B

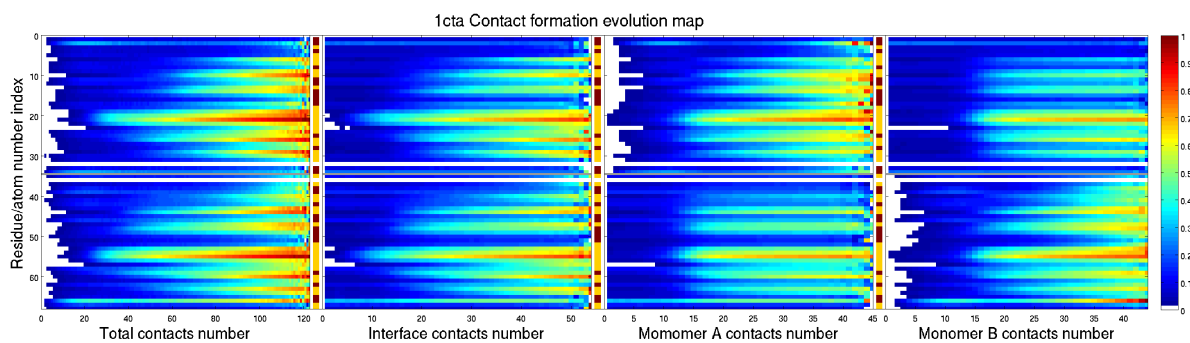


Figure 9.1: Contact evolution of 1cta

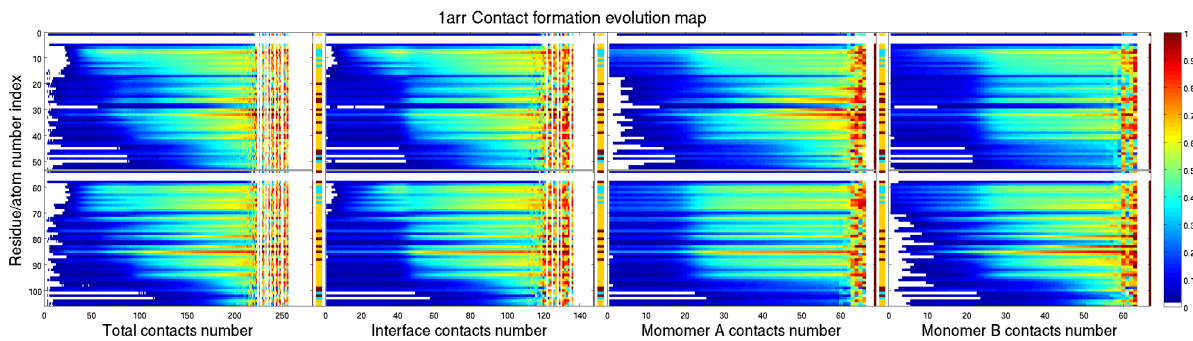


Figure 9.2: Contact evolution of 1arr

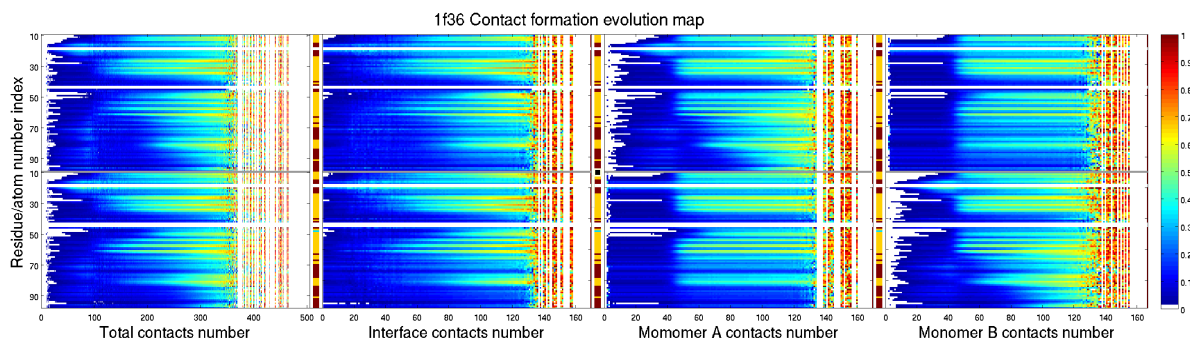


Figure 9.3: Contact evolution of 1f36

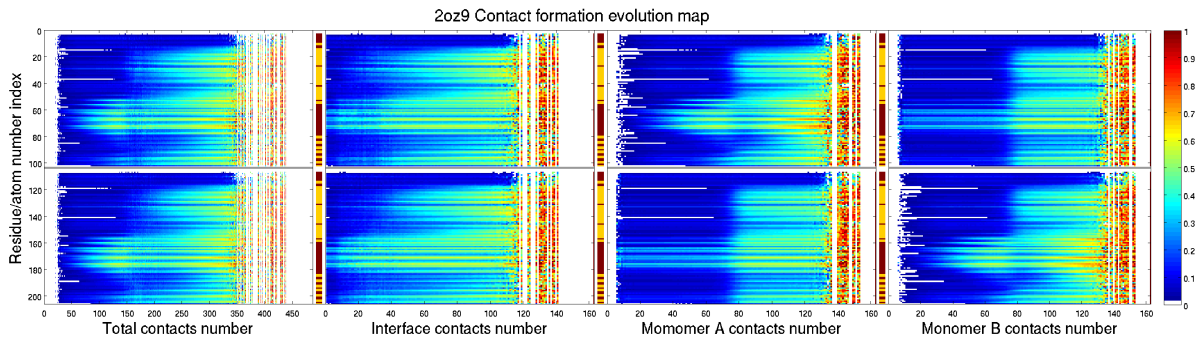


Figure 9.4: Contact evolution of 2oz9

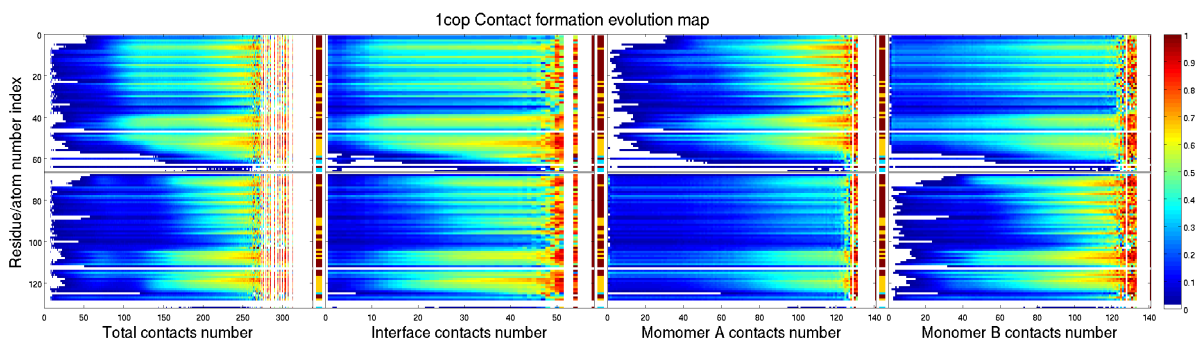


Figure 9.5: Contact evolution of 1cop

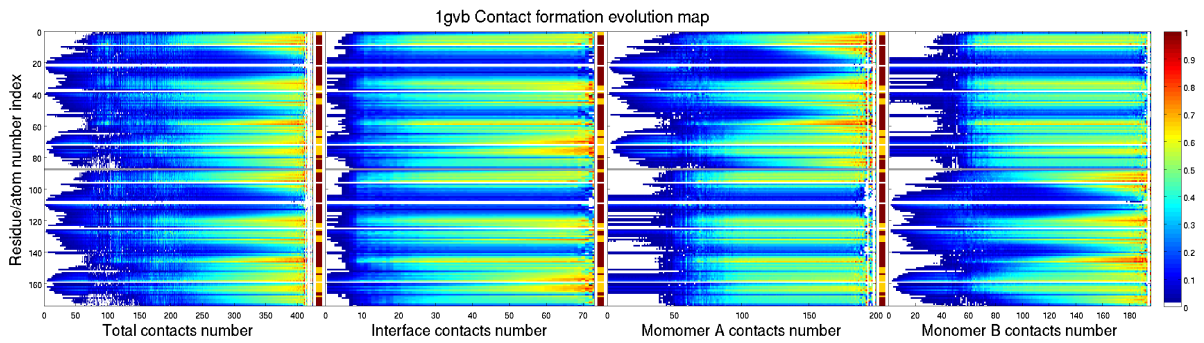


Figure 9.6: Contact evolution of 1gvb

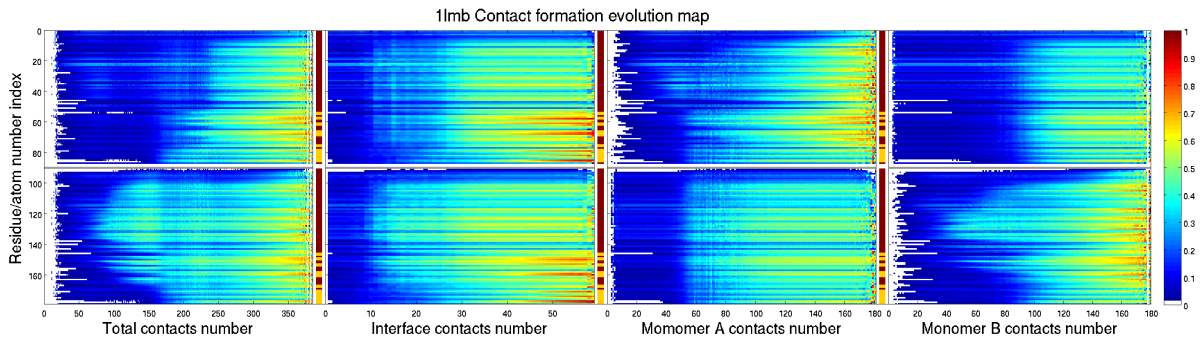


Figure 9.7: Contact evolution of 1lmb

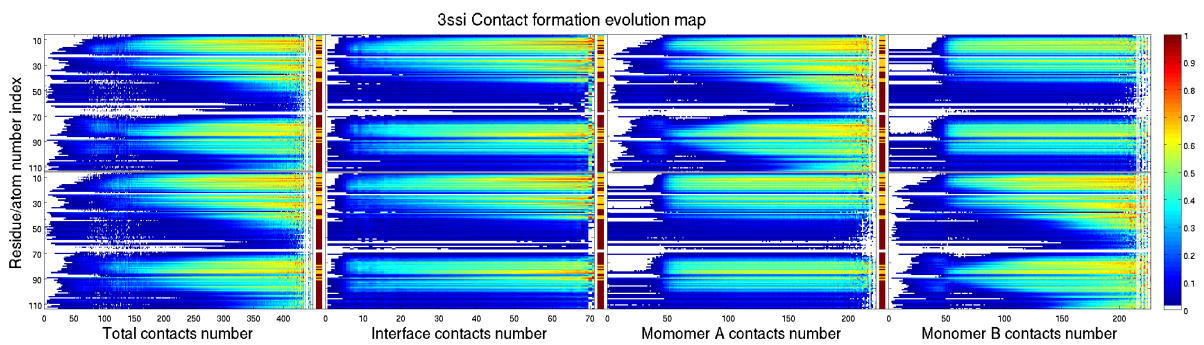


Figure 9.8: Contact evolution of 3ssi

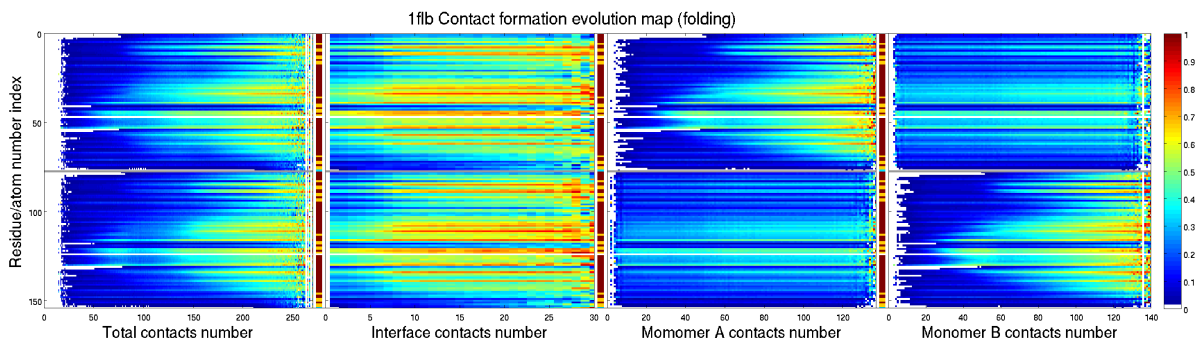
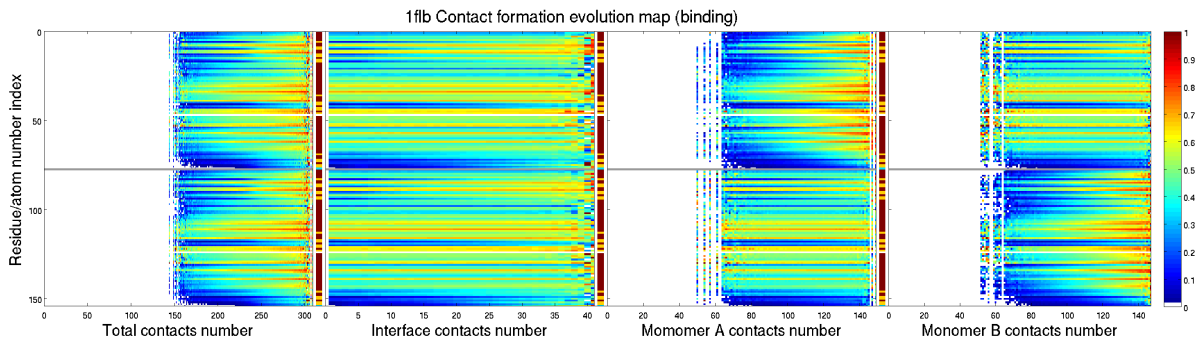


Figure 9.9: Contact evolution of