doi:10.1016/j.jmb.2004.12.021 *J. Mol. Biol.* (2005) **346**, 1121–1145

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

A Survey of Flexible Protein Binding Mechanisms and their Transition States Using Native Topology Based Energy Landscapes

Yaakov Levy[1,2]*, Samuel S. Cho[1,3], José N. Onuchic[1,2] and Peter G. Wolynes[1,2,3]

[1]*Center for Theoretical Biological Physics, University of California at San Diego, 9500 Gilman Drive, La Jolla CA 92093, USA*

[2]*Department of Physics, University of California at San Diego, 9500 Gilman Drive La Jolla, CA 92093, USA*

[3]*Department of Chemistry and Biochemistry, University of California at San Diego, 9500 Gilman Drive, La Jolla CA 92093, USA*

**Corresponding author*

Many cellular functions rely on interactions between protein pairs and higher oligomers. We have recently shown that binding mechanisms are robust and owing to the minimal frustration principle, just as for protein folding, are governed primarily by the protein's native topology, which is characterized by the network of non-covalent residue–residue interactions. The detailed binding mechanisms of nine dimers, a trimer, and a tetramer, each involving different degrees of flexibility and plasticity during assembly, are surveyed here using a model that is based solely on the protein topology, having a perfectly funneled energy landscape. The importance of flexibility in binding reactions is manifested by the fly-casting effect, which is diminished in magnitude when protein flexibility is removed. Many of the grosser and finer structural aspects of the various binding mechanisms (including binding of pre-folded monomers, binding of intrinsically unfolded monomers, and binding by domain-swapping) predicted by the native topology based landscape model are consistent with the mechanisms found in the laboratory. An asymmetric binding mechanism is often observed for the formation of the symmetric homodimers where one monomer is more structured at the binding transition state and serves as a template for the folding of the other monomer. Φ values were calculated to show how the structure of the binding transition state ensemble would be manifested in protein engineering studies. For most systems, the simulated Φ values are reasonably correlated with the available experimental values. This agreement suggests that the overall binding mechanism and the nature of the binding transition state ensemble can be understood from the network of interactions that stabilize the native fold. The Φ values for the formation of an antibody–antigen complex indicate a possible role for solvation of the interface in biomolecular association of large rigid proteins.

Keywords: protein folding; protein binding; induced fit mechanism; fly-casting mechanism; funnel energy landscape

Introduction

Understanding the detailed dynamics of protein interactions with partners that are small molecules or, very often, other biological macromolecules, proteins, nucleic acids, or polysaccharides, must ultimately underpin understanding most protein functions. Transient interactions between both proteins and nucleic acids are ubiquitous and fundamental to many subcellular processes. Recent studies in yeast have demonstrated that most proteins exist in the cell as parts of multicomponent assemblies.[1,2] Most of these complexes have components in common with at least one other multiprotein complex, reflecting a complex high-order network of interacting proteins. Predicting interactions on the proteomic scale[3,4] requires

Abbreviations used: TSE, transition state ensemble; BS-RNase, bovine seminal ribonuclease; HEWL, hen egg-white lysozyme; WHAM, weighted histogram analysis method.

E-mail address of the corresponding author: klevy@physics.ucsd.edu quantitative prediction of the dynamics and specificity of protein recognition and assembly. Such understanding may lead to the ability to design partners that form more stable complexes, which can then act as "network" drugs. Understanding interactions will also help us to find ways of inhibiting pathogenic association processes such as irreversible aggregation.

Biomolecular recognition processes are often described as the association of folded proteins that dock as rigid bodies (the "lock and key" mechanism[5]), or as formation of an encounter complex that undergoes small, local conformational changes to optimize the initial interactions (the "induced fit"[6] mechanism). These models have been proposed and are in modern days discussed by comparing the crystal structures of the bound and unbound proteins. It is becoming increasingly appreciated that flexibility is often more fundamental in the mechanisms of protein association than these pictures lead us to believe. Protein function is not strictly related to the three-dimensional structure of the folded state but to its four-dimensional dynamics governed by an energy landscape.[7] The degree of plasticity involved in protein binding has been recently investigated by molecular dynamics simulations.[8–10] Protein flexibility has also been invoked in the model of "conformational selection",[11–14] which suggests that binding entails choice of the correct conformer for binding out of a rapidly interconverting ensemble. Selection of specific protein conformers and the presence of conformational isomerism prior to association has been shown by kinetic and equilibrium studies,[15,16] as well as by NMR.[17] Support for the functional role of conformational diversity is provided by the observation that a single protein can bind multiple unrelated ligands at the same binding site (promiscuity) or at different sites (moonlighting).[18–20] Catalytic promiscuity and moonlighting are intriguing consequences of pre-existing protein isomers, which may differ in both side-chain and backbone conformations.

Greater flexibility during protein binding processes is also envisioned in the domain-swapping mechanism.[21,22] In this mechanism, proteins assemble by exchanging a secondary structure element or an entire globular domain with the symmetrically identical part of the other subunit, an interaction consistent with the principle of minimal frustration.[23–27] Proteins that form domain-swapped oligomers are found to have at least two well-defined native states: a monomeric and an intertwined structure. Domain-swapping requires a large conformational change as at least partial unfolding of the monomeric structure is needed to allow the interchange to take place.[28] The mechanism of assembly by domain-swapping and its biological significance in both functional protein assemblies and in pathologic aggregates are of much interest.

A still more extreme manifestation of conformational change upon binding occurs when monomer folding is directly coupled to binding events.[29,30] Many cellular proteins appear to be partially, or even completely, unstructured under native conditions, but form a perfectly ordered structure in the presence of the appropriate ligand.[31,32] Intrinsically disordered proteins are thought to be common in the genome rather than rare exceptions.[33] A sequence-based bioinformatics approach has predicted that more than 30% of the genome of 29 eukaryotes have proteins with disordered regions of 40 or more consecutive residues.[34] An energy landscape survey of a large database of protein complexes has suggested that ∼15% of monomers may not fold in the absence of partner proteins.[35] Several physiological advantages have been suggested for the use of disordered proteins that only fold upon reaching their targets. One functional advantage is that natively unfolded proteins are more adaptive, giving them the capability to bind to several different targets,[31,36] overcome steric clashes, and thus achieve high specificity with low affinity.[37] Another advantage of being unfolded is the capability of an extended fragment to form complexes with large interfaces. These large interfaces may therefore contain more information than the smaller ones tolerable in a completely folded protein. For a protein to be stable as a monomer while having extensive interfaces, the size of the protein needs to be more than twice as large as one with a small interface, resulting in increased cellular crowding.[38] Accordingly, disordered proteins provide a simple solution for having large inter-molecular interfaces while maintaining a small genome. A kinetic advantage for being initially unfolded before binding has been postulated through the fly-casting mechanism.[39] A partially structured or unstructured protein has a greater capture radius than a folded protein with its limited flexibility for a specific binding site, thereby enhancing the speed of association. Fly-casting also presents a route for kinetic specificity, even when absolute speed is not essential.

Analyzing the transition state ensemble (TSE) for a binding reaction, in principle, provides microscopic insight into the degree of flexibility involved in protein association and leads to predictions that are testable in the laboratory. The structures of oligomeric proteins at the binding transition state may localize those parts of the internal structure of the complex subunits necessary for association, as well as the crucial interfacial contacts needed for a productive association. Transition state ensembles for folding and binding share similar characteristics, in that for both processes non-bonded interactions are formed, either intra or inter-molecularly. When the binding process is coupled to monomer folding, the search problem is similar to that of protein folding and a single transition state reflects both monomer folding and binding. Binding between already folded subunits will exhibit distinct transition states for monomer folding and binding. Thus, binding can be viewed as analogous to folding processes that often have multiple intermediate states, reflecting a partially ordered ensemble, with the difference of the sequence discontinuity. In general, the search space involved in the binding of fully structured proteins is smaller than that for folding, yet it is still large enough that predicting the structure of a complex formed between two interacting proteins is currently a challenge.[40–42] The search involved in binding is even more extensive when we take into account all the possible complexes a protein might form in a cell with inappropriate partners.

The degree of structure at the transition state is usually quantified by the Φ values, which are essential for understanding protein folding pathways on a microscopic basis.[43–45] Experimentally, a Φ value for a given residue is calculated by the ratio of the effect that a mutation at that position has on the stability of the TSE over its effect on the stability of the folded state, both relative to the denatured ensemble. A Φ value close to 1 means that the mutation similarly affects the TSE and the folded state, suggesting that the mutated residue is analogously structured in the TSE as it is in the folded state. Conversely, a Φ value close to 0 means that the mutation does not affect the stability of the TSE (relative to its unfolded state), indicating that the mutated residue is unstructured at the TSE. While Φ value analysis has been widely used to decipher folding mechanisms, there are only few cases where Φ value analysis has been applied to characterize the TSE of binding.[46–50]

To study the nature of the binding transition state ensemble and its dependence on the degree of flexibility, we have simulated the association of various monomeric proteins into dimers, trimers, and tetramers. The protein complexes we have selected differ in their size, topology, and detailed association mechanism. Specifically, our survey includes nine dimers, a single homotrimer, and a single homotetramer. The dimer selected includes eight homodimers and a single hetrodimer. The folding of three of the homodimers (Arc-repressor, troponin C site III, and FIS dimer) is coupled to their binding ("two-state" homodimers, also called obligatory complexes) and for the other three homodimers (λ repressor, LFB1 transcription factor, and λ Cro repressor) monomer folding is a prerequisite for their association ("three-state" homodimers or non-obligatory complexes). In addition, one homodimer (Trp repressor) is formed *via* a dimeric but incompletely folded intermediate. Our set of homodimeric proteins also includes the dimeric bovine seminal ribonuclease (BS-RNase), which is stable as a monomer but can also form two different quaternary structures: a dimer with a small interface and a domain swapped dimer. A comparison between the different binding modes of BS-RNase provides an instructive example of the peculiarities of domain-swapping. The heterodimer included in our study is the complex between hen egg-white lysozyme (HEWL) and its Fab antibody. These partners associate as already folded subunits (this complex can be treated as a trimer, since the antibody is composed of distinct light and heavy chains). The homotrimer we have studied is the HIV-1 gp41 envelope protein, which shows two-state thermodynamics. The homotetramer chosen for study is the tetramerization domain of the tumor suppressor p53, often termed as a "dimer of dimers". The tetrameric domain of p53 is an especially important system, not only because of its key role in cancer, but also because the same units at different stoichiometry show two distinct binding mechanisms. The tetramer is formed by dimerization of prefolded dimers, while these are in turn formed by association of unstructured monomers.[48] Thus, p53tet serves as an elegant system for a theoretical study of the general principles of protein association.

The microscopic analysis of the Φ values for the transition state of binding was carried out based on native topology-based (Gō) model simulations that include only those contacts that exist in the native complex structure. Such models, which lack non-native interactions, correspond to perfectly funneled energy landscapes. These simulations for all eight of the above-mentioned homodimeric proteins have successfully reproduced the gross features of their experimental association mechanism. This success indicates that binding, like folding, is governed by a funneled energy landscape.[8] The existence of a funneled landscape for both folding and binding processes suggests that proteins are evolutionarily designed to follow the principle of minimal frustration,[23,51,52] which results in a faster search through the many alternatives in the cell and affords considerable robustness of binding capability against possible mutations. The funneled landscape leading toward the native bound configuration guarantees that binding will also be stable against environmental and evolutionary fluctuations. Minimal frustration and the funnel concept has been previously used to explain different binding mechanisms,[11,35,53–55] enzyme pathways and allostery,[56,57] aspects of binding selectivity and specificity,[58] and the role that water-mediated interactions have in enhancing recognition.[59,60] The availability of experimental Φ values for three of the simulated protein complexes (Arc repressor, lysozyme-Fab complex, and the tetrameric domain of p53) enables an evaluation of the accuracy of Φ values calculated from the Gō-model. For the other systems, our study provides predictions of the transition state ensembles for different protein associations.

Results and Discussion

Quantifying topological characteristics of protein complexes

The first step to elucidating how monomers form oligomers is to explore the structural properties of the complexes and to look for a direct correlation of structure with the association mechanism found by experiments. Previous studies have characterized protein–protein interfaces using measures of accessible surface area, interface polarity and planarity,[61] buried water molecules,[62] amino acid composition, and residue–residue preferences.[63] In this study, however, we quantify also the properties of the interface topology, as reflected by the connectivity of the network of residue–residue interactions. We have examined the structures of 25 homodimers (ten are two-state homodimers and 15 are three-state homodimers), as well as a non-redundant data set of 122 homodimers,[64] whose kinetic behavior has not been classified experimentally. The topological analyses of the complex interfaces and monomers of this large set are shown in Figure 1. Characteristics of those protein complexes that were selected for simulation study are summarized in Table 1. Figure 1(a) shows a classification of the dimeric structures based on the number of intra and inter-molecular contacts per residue. This plot illustrates that monomers that fold only upon binding have fewer monomeric contacts per residue than do monomers that fold independently from binding. The interfaces of the resulting dimers are larger in those cases when there is coupling between folding and binding. These dimers also have a higher number of average interfacial contacts per residue. In addition, the interfaces for two-state dimers, when folding and binding obligately couple, are more hydrophobic (Figure 1(b)). This indicates the possible role of electrostatic interactions,[65,66] as well as wet interfaces in the case of binding between folded subunits.

To study further the structural features of the selected homodimers, we analyzed the properties of the network of contacts. The topology of each dimer is characterized by calculating the average clustering coefficient (equation (1)) and the average shortest path-length (equation (2)). For two-state dimers, we find the average clustering coefficient for residues at the interface between the proteins is larger than that for other residues. Thus, interfaces that are formed in a coupled folding/binding reactions, are more packed and have a network of contacts that is denser than that of the monomer itself (Figure 1(c)). This trend is not found for three-state homodimers, where the clustering coefficient for interfacial residues is similar to the average clustering coefficient of all the residues. There are some three-state homodimers, however, with a larger interfacial clustering coefficient than monomeric clustering coefficient. The dimer topologies were additionally quantified by the mean shortest path-length (L). Two-state homodimers with up to 150 residues per monomer have larger L values than the corresponding three-state complexes do (Figure 1(d)), indicating again their imperfect packing within the monomer. This trend is also reflected by smaller C values in comparison to three-state homodimers of similar lengths. The relatively dilute

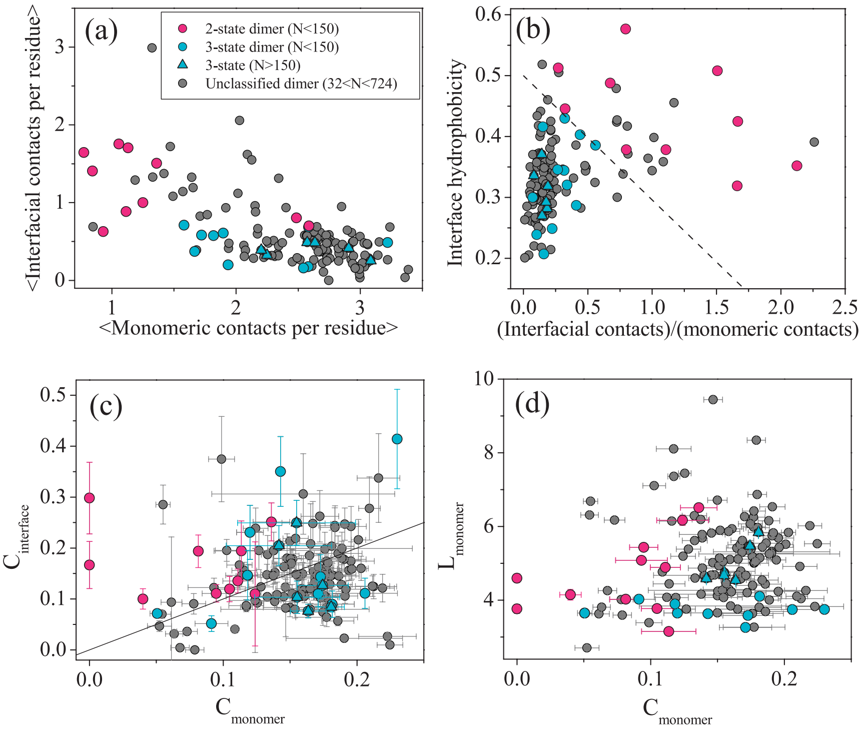

Figure 1. Topological analysis of protein complexes. The analysis includes ten homodimers that exhibit coupling between folding and binding (red circles), 15 homodimers that fold prior to their binding (blue circles and triangles) and a non-redundant data set of 122 homodimers which are not classified experimentally (grey circles). Triangles denote dimers with more than 150 residues per monomer. The complexes are characterized by (a) the average number of monomeric and interfacial native contacts per residue, (b) the interface hydrophobicity, (c) the average clustering coefficient of monomeric and interfacial residues, and (d) the mean shortest path length for a monomer (d). The broken line in (b) ($y=-0.204x+0.5$) is a tentative separation line between two-state and three-state association mechanisms. For the other graphs, separation lines were not drawn due to larger uncertainties. The continuous line in (c) ($y=x$) is plotted to aid the comparison between the clustering coefficient of monomeric and interfacial residues.

Table 1. Structural and topological properties of the studied protein complexes

Name	PDB code	Monomeric NC/No. of residues	Interfacial NC/No. of residues	Interfacial NC/Monomeric NC	Interfacial hydrophobicity	C_{monomer}	$C_{\mathrm{Interface}}$	L_{monomer}
A. *Association of unstructured monomers*								
Troponin C site	1cta	1.11	0.89	0.79	0.58	0.11	0.19	3.15
Arc repressor	1arr	1.06	1.75	1.66	0.32	0.12	0.11	6.17
Factor for inversion stimulation	1fis	1.36	1.51	1.11	0.38	0.09	0.11	5.44
Trp repressor	2wrp	1.25	1.00	0.80	0.38	0.14	0.25	6.51
HIV gp41	1i5x	1.13	1.71	1.51	0.51	0.08	0.19	4.03
Dimeric p53[a]	1sak	0.77	1.65	2.13	0.35	0.04	0.10	4.15
		1.13±0.18	1.42±0.34	1.33±0.48	0.42±0.09	0.10±0.03	0.16±0.06	4.90±1.22
B. *Association of structured monomers*								
λ Repressor	1lmb	1.90	0.61	0.32	0.43	0.17	0.14	3.59
λ Cro repressor	1cop	1.82	0.58	0.32	0.34	0.09	0.05	4.03
LFBI transcription factor	1lfb	1.67	0.37	0.22	0.25	0.21	0.11	3.75
BS-RNase M=M	1bsr	2.58	0.18	0.07	0.3	0.02	0.15	3.9
BS-RNase M×M[b]	1bsr	2.17	0.90	0.41	0.29	0.09	0.12	5.08
Tetrameric p53[a]	1sak	1.58	0.71	0.44	0.4	–	–	–
Lysozyme–Fab complex[c]	3hfm	2.54	0.16	0.16	0.21	0.23	0.41	3.75
		2.04±0.37	0.50±0.26	0.27±0.12	0.32±0.07	0.14±0.08	0.16±0.11	3.80±0.15

The criterion for the existence of native contacts (NC), which is elementary to the topological analysis of the complexes, is a distance between C^{α} atoms of two residues i and j, which satisfy $|i-j|>3$, of less than 8 Å, or a distance between any side-chain heavy atoms in the two residues is less than 4 Å.

[a] For the dimeric p53 the interface between two unfolded monomer was analyzed with respect to a single monomer. For the tetrameric p53 the interface between the two dimers was analyzed with respect to a single dimer. Due to sequence discontinuity, the *C* and *L* parameters cannot be calculated for a tetrameric p53.

[b] Due to the lack of structure for the domain-swapped structure of BS-RNase (M×M), it was modeled based on the non-swapped structure (M=M) (see Models and Method section for details). As the formation domain swapped structure involves association of at least partially unfolded monomer it was not included in the averaging of the complexes that are supposed to be formed by binding of folded subunits.

[c] The complex between the lysozyme and its antibody (Fab) is the only hetro-oligomeric protein in this Table. The monomer in this case refers to the lysozyme only.

contact networks in the monomers of two-state dimers are, in general, compensated by having more contacts per residue at the interface (Figure 1(a)) and a there being denser network of contacts at the interface (Figure 1(c)), as reflected by a higher clustering coefficient for the interfacial residues. This fact is illustrated by the extensive interfaces for two-state homodimers (shown in Figure 2 by the red lines), which are often characterized as intertwined structures, in comparison to the simple topology of the interfaces of three-state homodimers (Figure 5). Additionally, the quantitative structural analysis illustrates that the majority of the homodimers from the non-redundant data set share similar structural properties with three-state homodimers. Accordingly, most homodimers, as reflected by the PDB entries, are formed by the binding of folded monomers. This is in agreement with the previous prediction that 6–17% of proteins encoded by various genomes are fully disordered.[34]

Binding transition state ensemble for dimerization

Coupled folding–binding reactions

The free energy surfaces for the topology based models of binding of Arc-repressor,[67,68] troponin C site III,[69] FIS dimer,[70] and Trp-repressor[71] are shown in Figure 2 as a function of Q_{Total} and the separation distance between the two chains. For Arc-repressor, troponin C site III, and FIS dimer, two states exist at equilibrium: unfolded monomers and folded dimers, i.e. no thermodynamic intermediate is detected in the simulations during their association starting from unfolded chains, in agreement with experiments. We would like to point out that a more detailed analysis has to be performed to address the existence of an on-pathway kinetic dimeric intermediate that might be populated during the folding of these homodimers. To examine the coupling between folding and binding, the free energy surfaces were also projected on the reaction coordinates for folding of the two monomers (Q_{A} and Q_{B}) and the reaction coordinate for binding (i.e. interface formation, $Q_{\mathrm{Interface}}$). These four-dimensional energy surfaces clearly show that the folding processes of monomeric Arc repressor, troponin C site III, and FIS dimer are all coupled to binding. At the binding transition state, the monomers are only partially folded and the interface is partially formed. Interestingly, the structure of the transition state is asymmetric, in that one monomer is significantly more structured than the other monomer. This may indicate a binding between the unfolded chain and a more fully

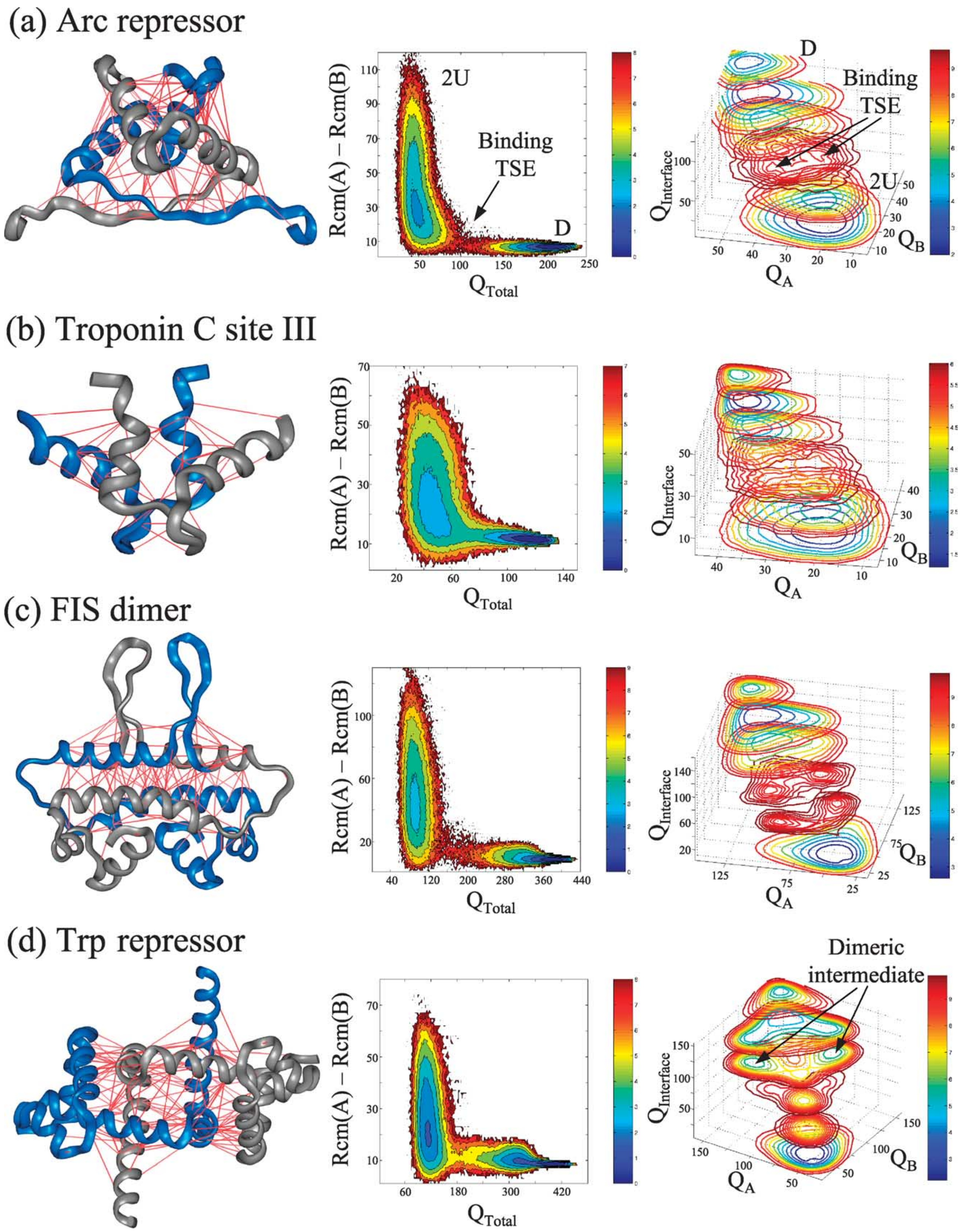

Figure 2. Binding free energy landscape for obligatory protein complexes. The dimer subunits of Arc repressor, troponin C site III, FIS dimer, and Trp repressor are colored grey and blue, and the native contacts that define the dimer interface are denoted by the red lines. The two-state binding thermodynamics is shown by the projection of the free energy on the total number of native contacts (Q_{Total} is the sum of the monomeric native contacts, Q_{A} and Q_{B}, and the interfacial native contact, $Q_{\mathrm{Interface}}$). The coupling between monomer folding and interface formation is seen by the projection of the free energy on Q_{A}, Q_{B}, and $Q_{\mathrm{Interface}}$. In this four-dimensional plot, a contour of the free energy as a function of Q_{A} and Q_{B} is plotted at six different values of $Q_{\mathrm{Interface}}$. U and D stand for an unfolded monomer and a folded dimer, respectively.

structured chain that has achieved a shape that supports recognition. Due to the symmetry of the dimer, the asymmetric binding pathway cannot distinguish between monomers A and B, as both can serve as a pseudo-template for binding to the other chain. For Trp repressor, a dimeric intermediate is populated at equilibrium, and its formation is the rate-determining step for dimerization. The four-dimensional energy landscape for Trp repressor indicates that a folded free monomer is not stable. The dimeric intermediate is comprised of a relatively well-folded monomer with the second monomer being only partially folded. In a previous analysis, we found that this intermediate gives rise to a significant fly-casting effect in the dimer formation of Trp-repressor.[8]

To microscopically analyze the nature of the transition state, we calculate the Φ values in the binding transition state ensemble. The Φ values are calculated for each contact formed in the native state (equation (3)), as well as averages for each residue that take into account all contacts made by that residue (equation (4)). The contact Φ values for the dimers that form concomitantly with monomer folding are shown in Figure 3 in the format of a contact probability map (the contact probabilities in the native state, P_{ij}, is also shown). The Φ values for individual contacts are not easily measured experimentally. However, they are useful in providing a more complete description of the structure of the transition state ensemble. Contact Φ values are particularly valuable in the case of the binding transition state because they treat separately the intra- and inter-molecular contacts of a given residue. The whole residue Φ value measured in the laboratory correspond to both intra-molecular and inter-molecular interactions.

Examining the contact Φ value maps for the binding transition state reveals two major points that are valid for all four studied homodimers. First, all the contact Φ values have relatively low values (centered around 0.5 or less) and, in general, no region is as completely structured as the native state. Second, the intra and inter-monomeric contacts generally have similar Φ values, which demonstrate that the monomeric and interfacial contacts are formed to nearly the same degree at the transition state. This is an outcome of coupling between folding and binding. The significant non-native character of the binding transition state is illustrated by the histogram of the calculated contact Φ values, which shows no value above 0.8. Our computed Φ values can be tested by comparing to measured ones. Of the dimers we simulated, experimental Φ values are currently available only for Arc-repressor.[46,49] The histograms of the computed and experimental Φ values for Arc-repressor satisfactorily overlap. Both histograms indicate a partially structured transition state; however, the computed Φ values are generally higher than those from experiment. Figure 4(a) and (b) show the correlation between the Φ values from experiments and simulations, giving a correlation coefficient of 0.31. The main deviation is found for the N-terminal region, which was more structured in the simulations than in the laboratory. Although there are detailed deviations between the simulated and experimental Φ values of particular residues, both methods support the overall view that residues 9–12 (located in the middle of the β-strand) and residues 25–30 (the C-terminal of the first helix) are the most structured at the binding TSE.

Binding of pre-folded subunits

The binding transition state ensembles for forming homodimers *via* a three-state mechanism were studied for three systems; λ repressor,[72] λ Cro repressor,[73] and LFB1 transcription factor.[74] These systems are all found experimentally to form by the association of already folded monomers. We have already shown that Gō model simulations for these homodimers successfully reproduce the gross features of their three-state binding mechanism.[8] However, while in all cases a stable folded monomer is formed during the association of these three homodimers, each one differs in the details of how the subsequent monomer is recognized.

λ repressor forms *via* the so-called induced-fit mechanism.[8,75] Figure 5(a) illustrates that two unbound monomeric λ repressors are prerequisite for the dimer formation. For LFB1 transcription factor and λ Cro repressor, a folded monomer is stable on its own; yet, in these systems association does not follow an induced fit mechanism. Rather, in these cases, the existence of a single folded monomer acts as a template for the folding of the other monomer. This asymmetric binding pathway resembles the asymmetric structure of the transition state found for the case of binding of monomers that are intrinsically unstructured. The folded subunit can be viewed as catalyzing the folding of the unfolded partner by forming interfacial contacts between the two chains. This catalysis of a monomer folding is shown in Figure 6. In each case the barrier for folding of a monomeric λ repressor, LFB1 transcription factor, and λ Cro repressor in the dimeric environment is lower than the folding barrier for an isolated monomer. The catalytic effect is more significant for LFB1 transcription factor and for λ Cro repressor than for λ repressor. A mild effect, like that found for λ repressor, has also been observed for HIV-1 PR (data not shown). This latter set of homodimers follows the induced-fit binding mechanism,[75,76] and the effect of the dimeric environment on the folding of their subunits can be interpreted as a crowding effect. The enhanced folding is expected to be concentration dependent, being reduced upon dilution.

The binding of both LFB1 transcription factor and λ Cro repressor occurs readily between unfolded and folded subunits. For LFB1 transcription factor, the complex between unfolded and folded monomers is transient, and its formation is coupled with

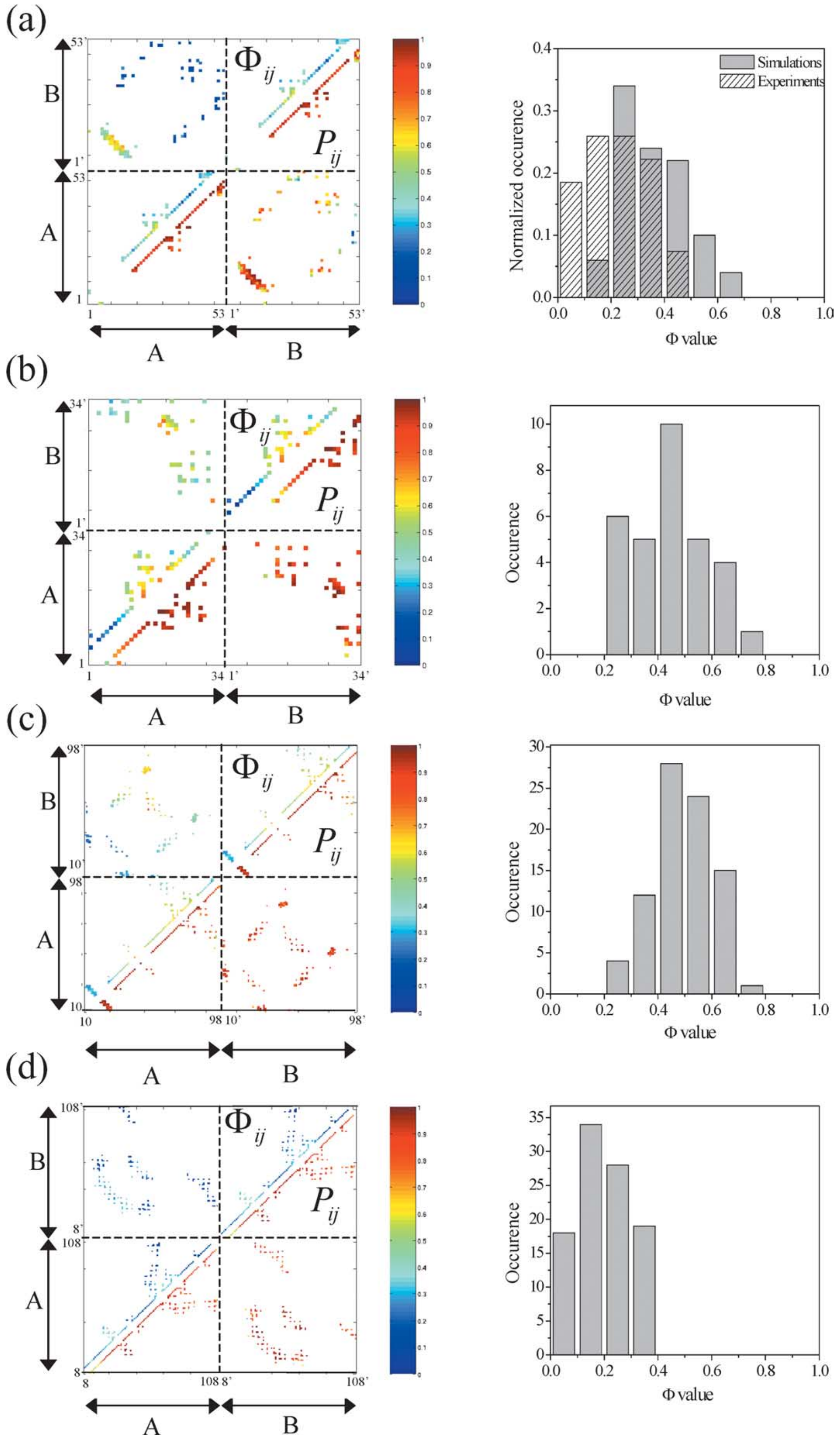

Figure 3 (*legend opposite*)

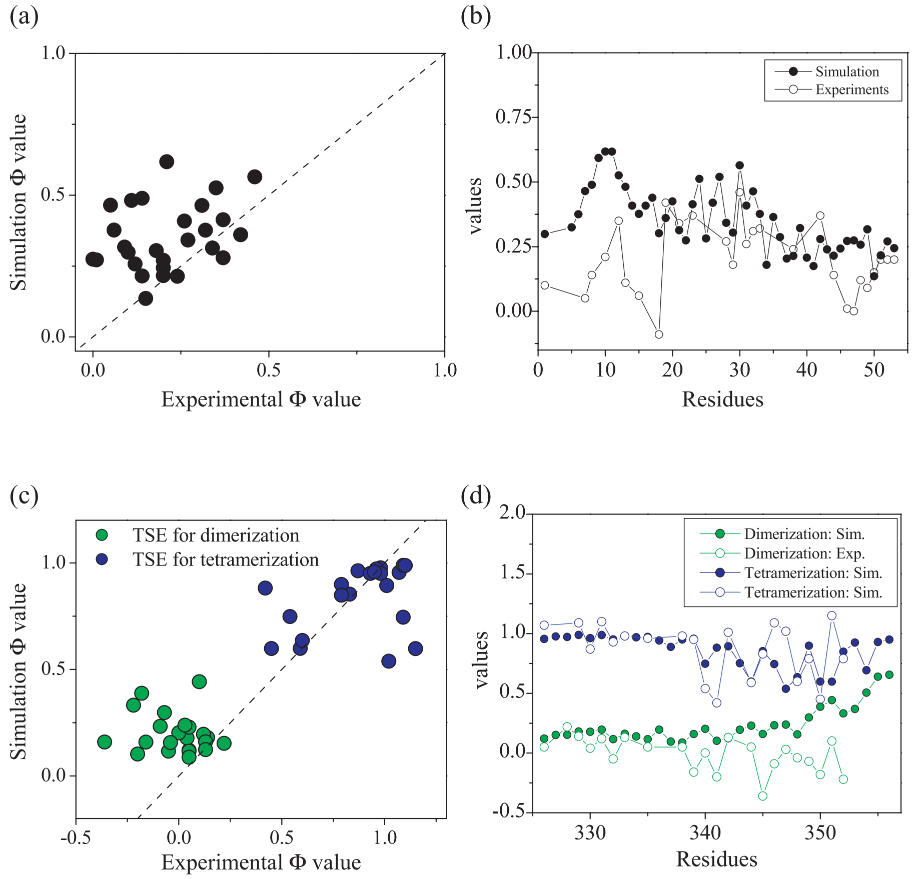

Figure 4. Comparison between the simulated and experimental Φ values for binding of Arc repressor (a and b) and the tetramerization domain of p53, p53tet (c and d).

the folding of the unfolded monomer (Figure 5(b)). This complex for λ Cro repressor is, however, more stable and was significantly populated in our simulations (Figure 5(c)). Once the complex is formed, the unfolded monomer can fold on its partner. The existence of this intermediate state was verified by simulating λ Cro repressor with a linker of 12, 20, and 30 glycine residues that connect between the two monomers, as well as in our standard simulations where a linker was avoided.[7] It is plausible that this asymmetric binding mechanism for λ Cro repressor is similar to the effect related to pro-sequences in enzymes such as subtilisin and α-lytic protease.[77] Pro-sequences, which are crucial for the timing of the enzyme activity, have been found to catalyze enzyme folding. The exact mechanism of the folding catalysis by pro-sequences is still unknown, however, it seems that a folded pro-sequence may act as a template for enzyme folding, as was found for LFB1 transcription factor and λ Cro repressor.

The different association mechanisms of the three-state homodimers we studied are reflected by different pattern of the Φ values at the binding transition state (Figure 7). The Φ values of the intra-monomeric contacts are much higher than those corresponding to the inter-monomeric contacts, indicating that at the binding transition state the monomers are nearly folded. A histogram of the Φ values indicates that most of the contact Φ values are close to unity. The Φ values at the binding transition state of LFB1 and λ Cro repressors are lower than those for λ repressor.

The extended and simple interface of λ Cro makes binding possible even before folding, even though the monomer is stable on its own. That there may be an adaptational advantage for this binding mechanism receives support from a recent retro-evolutionary study of Cro proteins.[78,79] The λ Cro repressor is the only one of the Cro proteins that has a β-strand element. This element arises from two specific mutations in a usually α-helical region of the structure for the other family members that are monomeric. Thus to enable dimerization λ Cro

Figure 3. The contact Φ values, Φ_{ij}, at the binding transition state of (a) Arc repressor, (b) troponin C site III, (c) FIS dimer, and (d) Trp repressor, and the contact probability, P_{ij}, in the native dimer. The histogram of the contact Φ values for each dimer is also shown and, for Arc repressor, is compared with the experimental values.

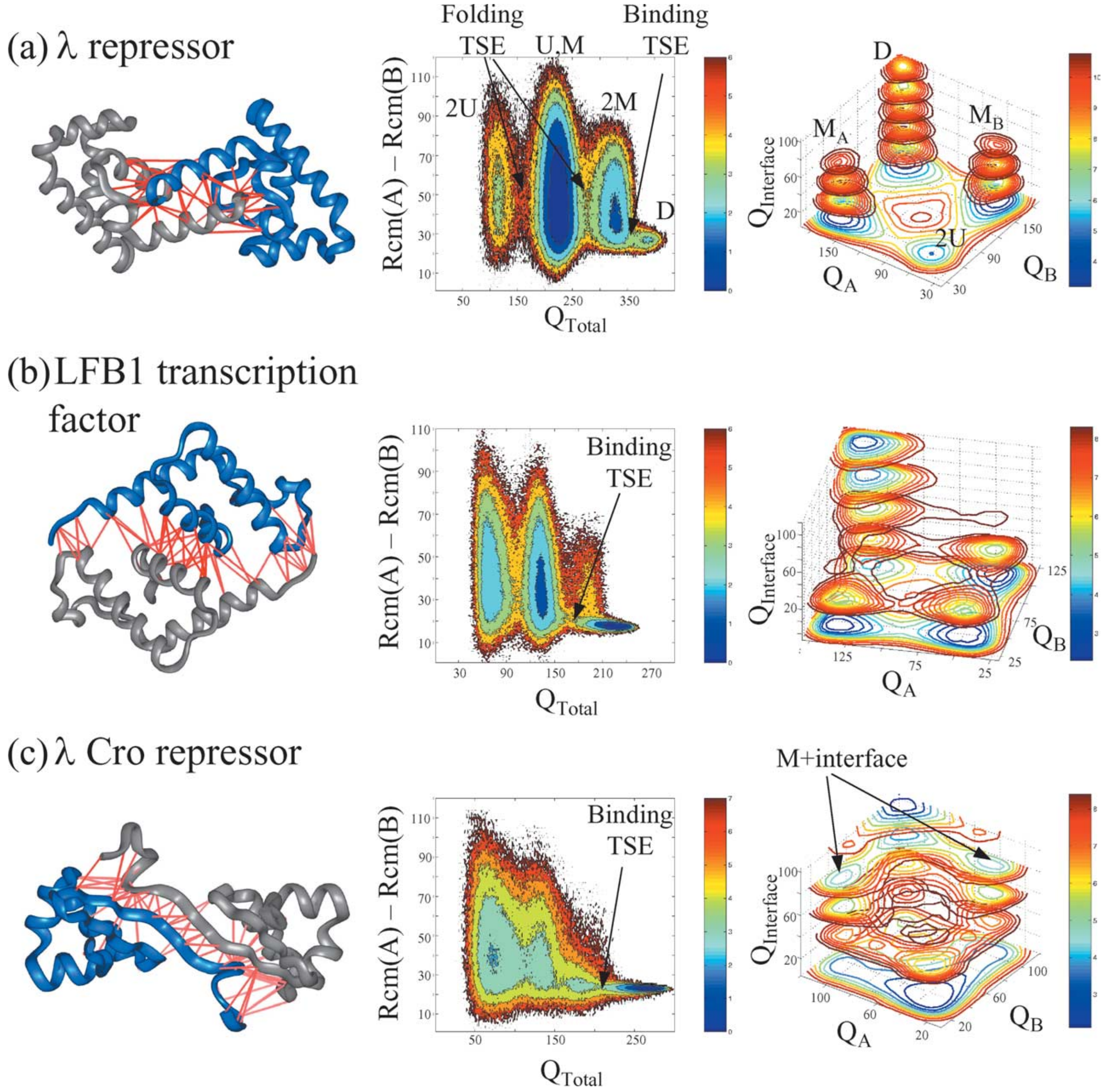

Figure 5. Binding free energy landscape for non-obligatory protein complexes. The complexes include (a) λ repressor, (b) LFB1 transcription factor, and (c) λ Cro repressor. U, M, and D refer to an unfolded monomer, a folded monomer, and a folded dimer.

repressor has evolved a new interface, now leading to an α/β protein rather than an all-α one.

Association by domain-swapping

Domain-swapping[21,22] takes place by interchanging identical structural elements between different subunits. Dimerization occurs by recruiting interactions that originally were evolutionary designed to internally stabilize separate subunits. The experimentally observed swapped dimer was reproduced using a landscape model that allows each intra-molecular interaction to be formed in a symmetric fashion intermolecularly. The postdiction of the native swapped oligomer by this (partially) frustrated model, which actually allows any region in the monomeric protein to swap, indicates that domain-swapping is a consequence of the principle of minimal frustration.[27] Since any monomeric protein has the potential to oligomerize *via* domain-swapping to form an inactive form of the protein, such as amyloids, it is crucial to understand the transition states for this binding reaction.

We study the transition state for domain-swapping by BS-RNase. The BS-RNase is an interesting system for studying the mechanism of converting a monomeric protein into a domain-swapped oligomer, because both of its quaternary structures are relatively stable and have been

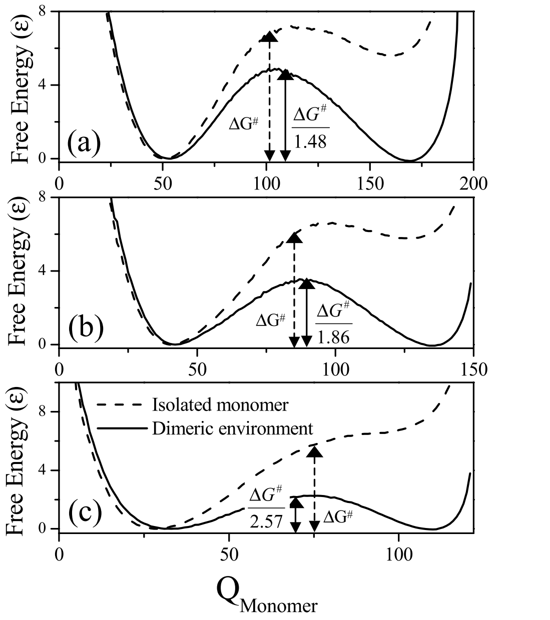

Figure 6. Free energy for folding of (a) monomeric λ repressor, (b) LFB transcription factor, and (c) λ Cro repressor. For each homodimer, the free energy plots are for folding of an isolated monomer (broken line) and a monomer in a dimeric environment (continuous line) and are both plotted at the folding temperature for the monomer in the dimeric environment.

structurally characterized in the laboratory: a domain-swapped dimer, M×M, and a dimer with no interchange, M=M. The interface formed between the two monomers is much larger in the case of domain-swapping (Figure 8). Accordingly, a greater stability is observed for the M×M dimer in association simulations of BS-RNase.[74] The contact Φ values show that for both forms of BS-RNase the Φ values for the interfacial contacts are smaller than those for intra-monomeric contacts. The monomers that form M=M BS-RNase are fully folded upon binding. The monomers that form the swapped structure are also nearly folded at the binding transition state, however, the α-helix at the N terminus (residues 1–19) is not yet fully structured. At the binding transition state of M×M, the contacts of this helix with the other subunit are formed to a larger degree ($\Phi_{ij}\sim 0.6$) than the interfacial contacts between the hinge loops that form at a later stage ($\Phi_{ij}\sim 0.3$).

Although for both dimerization reactions the residue Φ values span a similar range of values, the monomers that bind to form M=M are significantly more folded than the monomers that form the M×M BS-RNase (Figure 9). The Φ values suggest that forming M=M involves much less flexibility, while the formation of the domain-swapped dimer requires more coupling between folding and binding. It is possible that other systems associating *via* domain-swapping will follow a more complete unfolding and thus will show an even larger degree of coupling between folding and binding.

Binding transition state ensemble for trimer and tetramer formation

p53 tetramerization domain: dimerization of dimers

The tetramerization domain of p53 (p53tet, residues 326–356) is an intriguing system to study because it manifests different assembly scenarios in different stoichiometry.[48,80] The completed tetramer is formed by two steps: first dimerization of two unfolded chains, which then in turn further associate to form the tetramer. Accordingly, this tetramer is often called a dimer of dimers. Formation of p53tet dimer is coupled to monomer folding, while the tetramer is formed by the binding of already folded entities. The properties of the network of contacts that define the interface of the dimer and tetramer, as well as the topologies of the monomeric and dimeric p53tet (see Figure 10(a)), are consistent with the different assembly modes for formation of dimeric and tetrameric p53tet. There are 0.77 intra-subunit contacts per residue for the dimeric p53tet. The tetrameric p53tet has 1.58 such contacts per residue. There are 1.65 interfacial contacts per residues for the dimer but only 0.71 for the tetramer. Accordingly, the ratio between interfacial and intra-subunits contacts for dimeric p53tet is 2.13 and for the tetramer is 0.44. The interface hydrophobicity of the dimer is 0.35 while that for the tetramer is 0.40. The phase diagrams in Figure 1(a) and (b) would indicate, therefore, that the p53tet dimer follows a two-state equilibrium mechanism, while assembly into a tetramer should follow a three-state equilibrium binding. Moreover, the average clustering coefficient of the monomer is 0.04 ± 0.01 and that for the dimeric interface is 0.1 ± 0.02. The average shortest path-length for the monomer is 4.15 ± 0.15, suggesting an intrinsically unfolded monomer (see Figure 1(c) and (d)).

The folding and assembly of the WT tetrameric domain of p53 was simulated starting from four unfolded monomers. The specific heat curve for forming of WT p53tet exhibits two peaks (Figure 10(b)). To assign each of the specific heat peaks found for the tetramerization reaction, the change of the specific heat in the dimerization reaction alone was also calculated. Dimerization gives a peak at the higher temperature and thus the other lower-temperature peak corresponds to the dimers associating into a tetramer. This already suggests that the fully formed dimeric p53tet is prerequisite for the tetramer formation. To further analyze the tetramer assembly, the free energy surface was projected along the total number of contacts (Q_{Total}) and the radius of gyration of the tetramer (Figure 11(a)). Four different oligomeric states of p53tet were formed and correspond to all four subunits being unfolded, states where either a

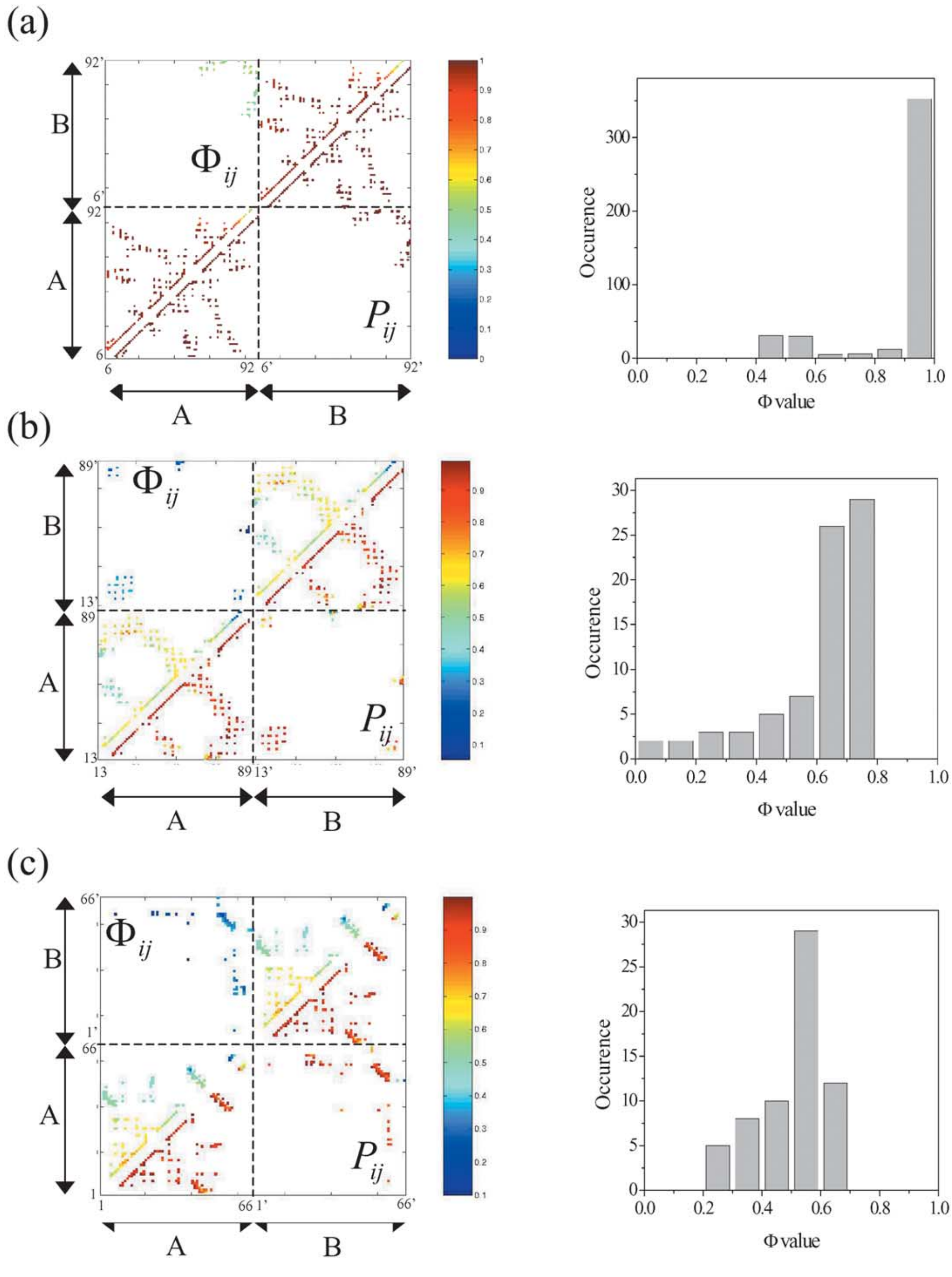

Figure 7. The contact Φ values, Φ_{ij}, at the binding transition state of (a) λ repressor, (b) LFB1 transcription factor, and (c) λ Cro repressor and the contact probability, P_{ij}, in the native dimer. The histogram of the contact Φ values for each dimer is shown.

single dimer or two dimers are formed, and a tetrameric state. The obligatory formation of two dimers to form a tetramer and the fact that no stable trimeric state is found in the simulation is consistent with the experimental designation of the tetrameric domain of p53 as a dimer of dimers (Figure 11(a)).

The two-step assembly is illustrated by the projection of the free energy surface along the reaction coordinates for the dimer formation (Q_{ac} and Q_{bd}) and for the formation of the tetramer interface (Q_{ac-bd}). Each dimer forms independently and is stable on it own. Yet, two folded dimers are prerequisite for tetramerization (Figure 11(b)). Accordingly, a dimer is formed by association of unfolded chains that become folded upon binding, while the tetramer is formed by association of folded molecules. These two different binding modes are reflected by the change of the free energy as a function of the distance between the two molecules participating in each binding reaction

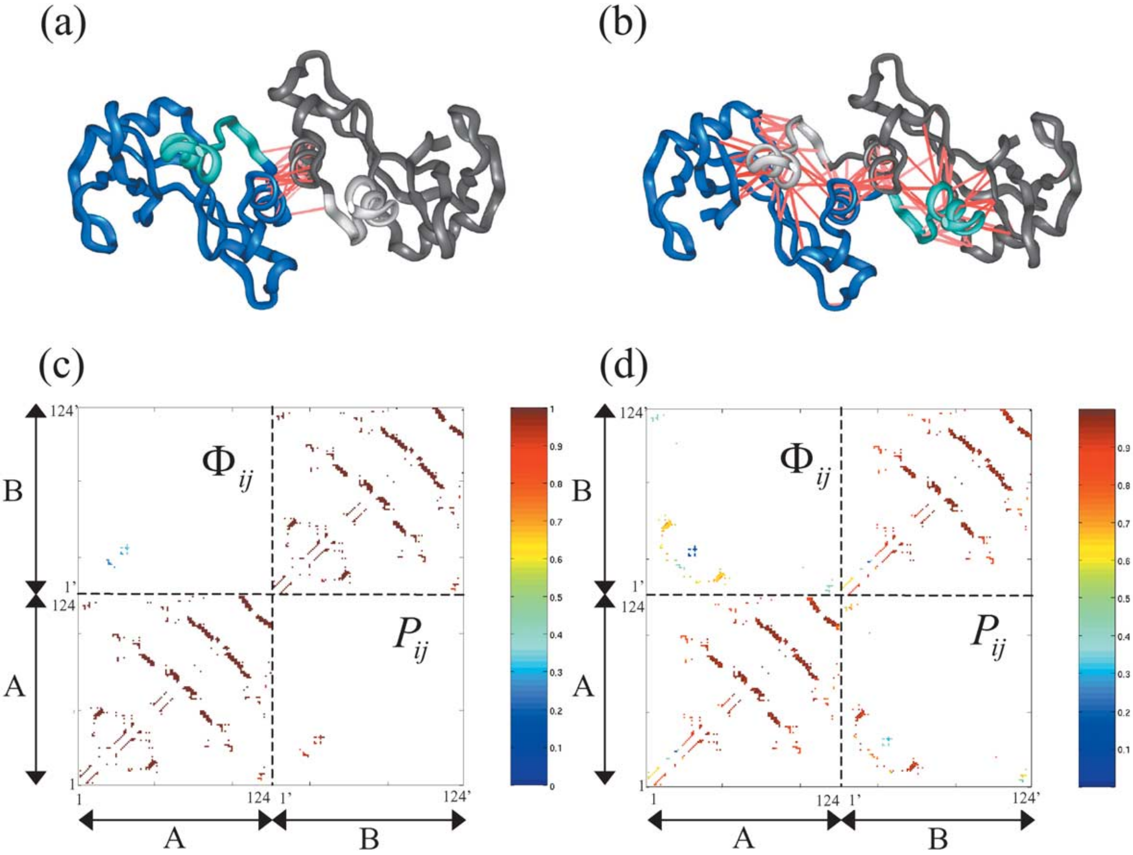

Figure 8. The two quaternary structures of bovine seminal ribonuclease (BS-RNase): (a) M=M and (b) M×M. The contact Φ values at their binding transition state and the contact probability in the native state are shown in (c) and (d), respectively.

(Figure 10(c)). A gradual decrease of the free energy is seen for the association of unfolded monomers even when the separation distance between them is relatively large, reflecting a fly-casting effect due to a transiently binding of unfolded regions. For the tetramerization, a barrier has to be surmounted to form a tetramer from two folded dimers. The barrier's origin is the entropy loss upon the complex formation, owing to the lack of a fly-casting effect. To assess the contribution of the high flexibility of the monomers to their association into dimeric p53tet *via* the fly-casting mechanism, a simulation was also carried out where all the monomeric contacts were kept permanently formed (Figure 10(c)). This corresponds to association reactions of partially flexible monomers (the monomers are not completely rigid as there are almost no interactions between the α-helix and the β-strand in each monomer). A milder fly-casting effect was actually still observed for this case, emphasizing the

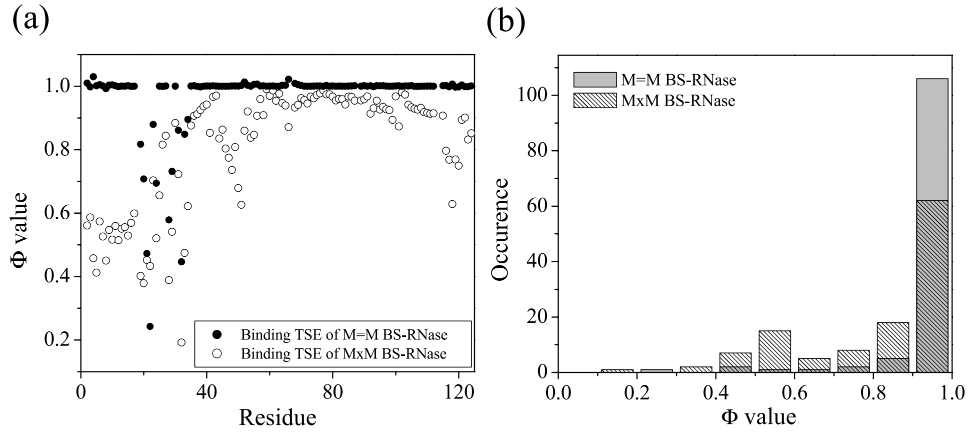

Figure 9. The residue Φ value at the binding TSE of M=M and M×M isoforms of BS-RNase.

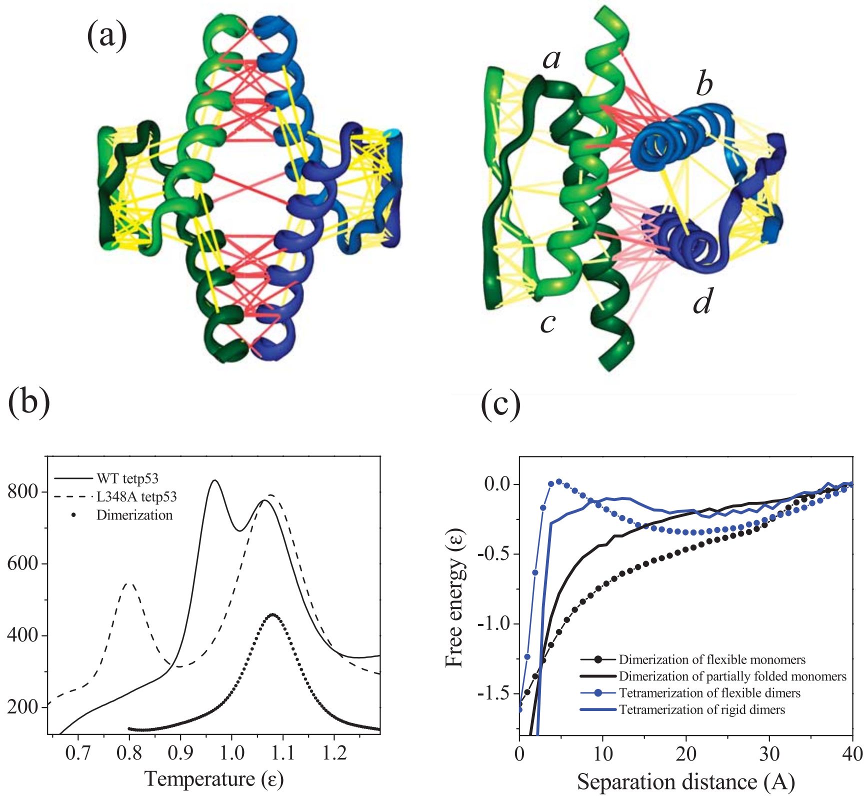

Figure 10. Assembly mechanism of the tetramerization domain of p53 (p53tet). The tetramer, which is shown in two different orientations, is composed of four identical monomers, *a*–*d*. The interfacial contacts between the dimers composed of *a* and *c* as well as between *b* and *d* are shown by yellow lines. (a) The interfacial contacts between these two dimers are shown by the red lines. (b) The specific heat as a function of temperature during the formation of p53tet, as well as of its mutant L348A. For comparison the specific heat during the formation of the dimers *ac* or *bd* is shown by the dotted graph. (c) The free energy as a function of the separation distance during the dimerization reaction (formation of *ac* or *bd*) and the tetramerization reaction (formation of *abcd*) for different degree of flexibility.

crucial role of flexibility in the association of the monomers. The signal for fly-casting is not eliminated, because the monomers are not completely rigid, and the β-strands that primarily participate in the interface formation are still flexible. We would like to point out that the fly-casting effect in the case of dimerization of p53tet might be affected by introducing either non-additive or non-native interactions even in the case of highly flexible recognition. The association of rigid folded dimers to p53tet does not show any fly-casting behavior. However, the barrier for binding is much smaller in comparison to the case where flexibility was not prohibited from the dimers. This is due to less entropy loss during the association, as the subunits have intrinsically less entropy, as dictated by the enforced rigidity.

A quantitative comparison between the association mechanisms of monomers and dimers can be obtained from Φ value analysis. The Φ values were calculated for the dimerization and tetramerization reactions based on the location of their corresponding transition state along *Q*. The contact Φ values for the two association modes exhibit significant differences (Figure 12). The contact Φ values for the dimerization have similar values for intra and inter-monomeric contacts, reflecting a coupled folding-binding process. These contacts have much higher Φ values in the tetramerization reaction where only the inter-dimeric contacts are partially formed. The low Φ values for dimerization in comparison to those for tetramerization parallel the pattern found in two-state *versus* three-state homodimerization. Comparing the calculated Φ values to those measured experimentally by Fersht and colleagues[48] yields a reasonable agreement (Figure 4(c)). The different nature of the two association reactions is correctly captured by the Gō simulations (the correlation coefficient obtained when comparing all the theoretical Φ values of both reactions with the experimental ones is 0.89). The correlation coefficient within each set of Φ values is smaller. For the tetramerization, the correlation coefficient is 0.36, however, upon exclusion of residues 347 or 351, it is 0.48 and 0.51, respectively. When both residues 347 and 351 are discarded, the correlation coefficient is 0.70. Although the Gō simulations indicate, consistently with experiments, low Φ values at the dimerization transition state, the correlation coefficient between the

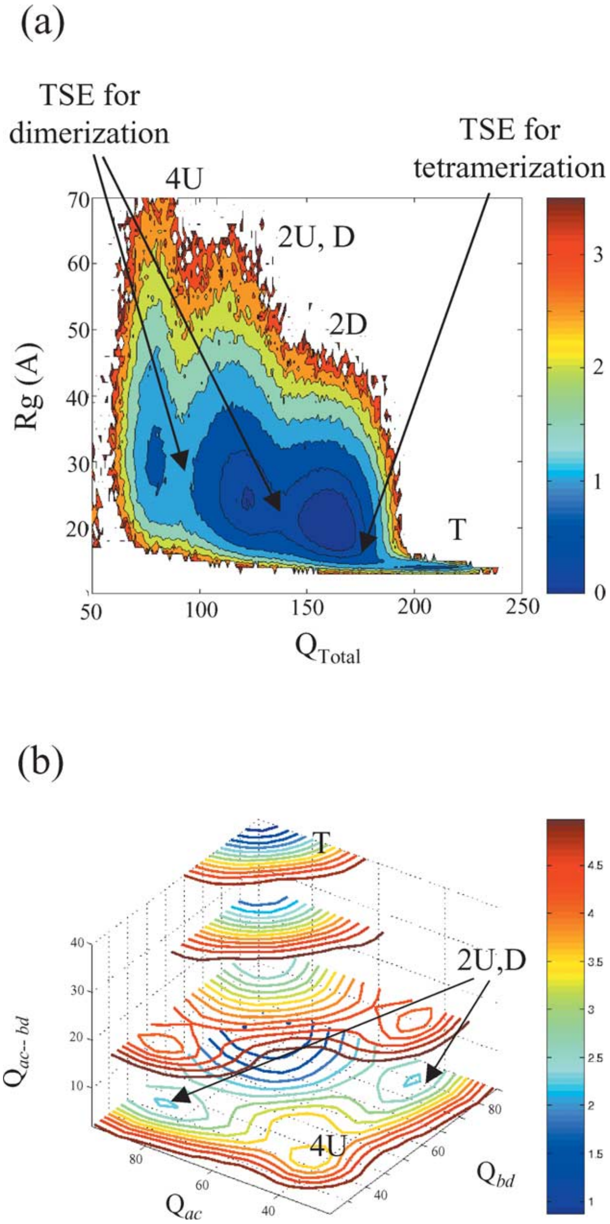

Figure 11. The free energy surface for the assembly of p53tet. (a) The free energy is plotted as a function of Q_{Total} and the radius of gyration of the tetramer. (b) A four-dimensional energy landscape is plotted along the reaction coordinates of formation of the dimers *ac* and *bd* and the tetramer interface (ac–bd). U, D, and T refer to an unfolded monomer, a folded dimer, and a folded tetramer, respectively.

theoretical and experimental Φ values of the association of two unfolded subunits is poor, around −0.08. This weak correlation can be understood by the nature of these small Φ values, which makes them more sensitive to fluctuations. Correlating low and similar Φ values (i.e. those that span a very small range) are intrinsically more vulnerable in comparison to correlating two sets of values that span a large range. Moreover, the fact that many negative Φ values are found experimentally may indicate that the p53tet dimer on its own is an energetically frustrated system, which cannot be captured by a simple minimally frustrated Gō model. We would like to point out that very recently an all-atom molecular dynamics study was done by Pande and his co-workers on the dimerization reaction of the tetramerization domain of p53.[81] The Φ values for the binding TSE from that study, which includes non-native interactions, displays qualitatively similar results to those obtained from our Gō simulations.

Fersht *et al.* have shown that the L348A mutant of p53tet destabilizes the tetramer by destroying the hydrophobic packing in the core of the oligomerization domain.[82,83] The leucine residue at position 348 participates in two contacts that stabilize the dimer and three contacts at the interface between the two dimers, resulting in a potential decrease of four dimeric and 12 tetrameric interfacial contacts upon mutation. The L344A mutant was also predicted to show destabilization of the tetramer in comparison to the dimer. This destabilization effect of the tetramer interface is consistent with the decreased number of contacts, while the dimer is unaffected. In our study, the L348A mutant was simulated by removing all the inter-molecular contacts that it participates in. Like the wild-type, the specific heat curve for forming L348A p53tet exhibits two peaks (Figure 10(b)). The peak at the higher temperature corresponding to the dimerization is similar for both the wild-type and mutant p53tet. The dimerization of dimers, however, is different for the two molecules. For the L348A p53tet, the temperature at which the two dimers are at equilibrium with a tetramer is now much lower than that for the wild-type. For our simulations, the mutant tetramer is less stable than the wild-type. The mutation also stabilizes the dimer and is thus consistent with the Fersht group measurements.[48,83]

The free energy surfaces for the tetramerization of WT and L348A p53tet (at the same temperature, ε = 0.96) plotted along the dimerization reaction coordinates (Q_{ac} and Q_{bd}) show similar thermodynamic properties (Figure 13(a)). In both cases, the formations of the dimers *ac* and *bd* are decoupled and their stability is similar. Figure 13(b) shows the free energy surface along the coordinates to form the dimers *ac* and *ab*. This illustrates that *ab* is formed only upon the formation of the dimer *ac*. Similar coupling is also observed between the folding of the dimer *ac* and the formation of the dimer *ad*, the trimer *abc*, and *acd* (data not shown). The formation of the trimer *abd* is coupled to the association of *a* and *c*. Accordingly, all the trimeric states, as well as the dimers *ab*, *ad*, and *bc*, are not stable and are defined only as part of the tetramer. For L348A, the stability of the tetramer (reflected by the free energy of either the dimer *ab* or the trimer *abd*) is much lower than that of the WT. This indicates that due to the poor packing at the tetramer interface, which is introduced by the

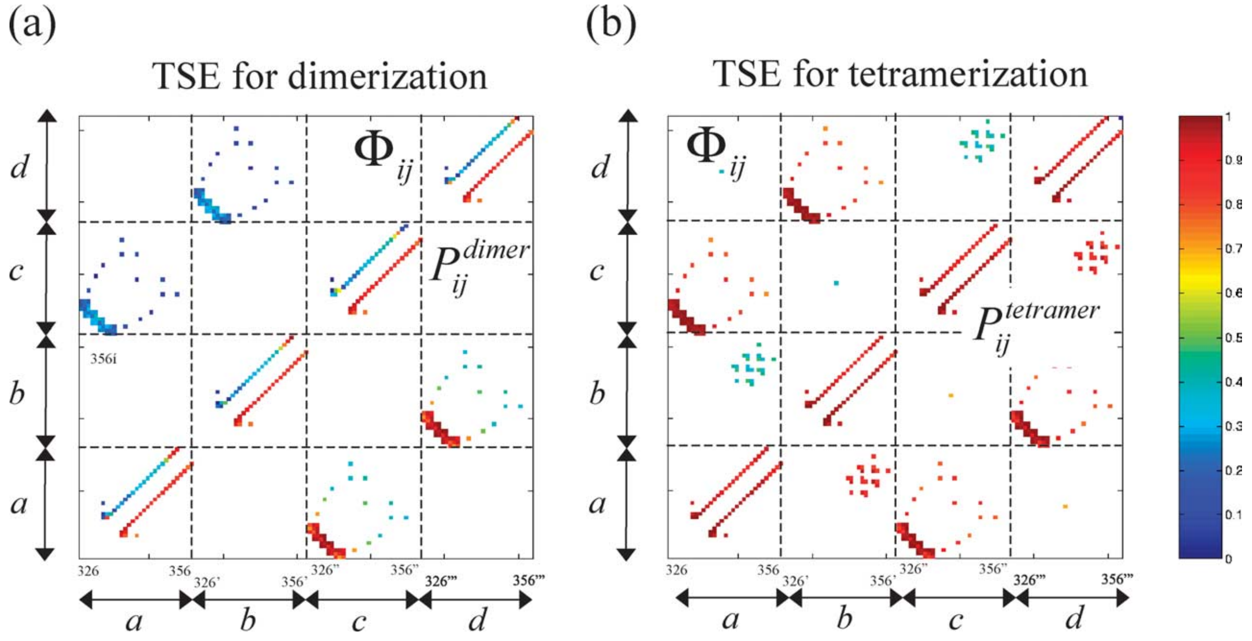

Figure 12. The contact Φ values at the (a) dimerization and (b) tetramerization TSE of p53tet. For comparison, the contact probability at the dimeric intermediate state and tetrameric state are shown.

L348A mutation, the tetramer is much less stable and will be populated only at lower temperatures.

HIV-1 gp41 envelope protein: a two-state trimer formation

The HIV-1 gp41 envelope protein (Figure 14(a)) is a homotrimer of helical hairpins that mediates the entry of immunodeficiency viruses into target cells by promoting the fusion of viral and cellular membranes. The folding thermodynamics of gp41 envelope protein follows two-state characteristics, which has been suggested to have implications for membrane fusion.[84,85] The ratio between interfacial and monomeric contacts is 1.51, the interface hydrophobicity is 0.51, the average clustering coefficient for monomeric and interfacial residues are 0.08 and 0.19, respectively, and the mean shortest path-length is 4.03. According to the phase diagrams in Figure 1, these values would place monomeric gp41 envelope protein (i.e. a single helical hairpin) with the two-state homodimers, that is, the free monomer is intrinsically unfolded under conditions where the trimer can form.

The free energy surface as a function of Q_{Total} ($= Q_a + Q_b + Q_c + Q_{ab} + Q_{ac} + Q_{bc}$, where Q_x counts the number of contacts in monomer x and Q_{yz} counts the number of interfacial contacts between monomers y and z) and the radius of gyration of the trimer indicates a two-state thermodynamics (Figure 14(b)). The coupling between the binding and folding of the helical hairpin is tested by plotting the free energy along the reaction coordinates for the formation of dimeric helical hairpins (*ab*, *ac*, and *bc*) (Figure 14(c)). This free energy plot illustrates that the dimeric species of gp41-envelope protein are not stable on their own, and they are defined only as part of the formed homotrimer.

Φ value analysis supports the existence of coupling between monomer folding and trimer formation. The contact Φ values for intra and inter-monomeric contacts have similar values (Figure 15(a)) and are relatively low (0.2–0.65, see Figure 15(b)). Accordingly, the HIV-1 gp41 envelope protein exhibit Φ values similar to those found for two-state homodimers. Our simulation study indicates that the residues at the C terminus of helix 1 (residues 20–32) and at the N terminus of helix 2 (residues 39–42) have the highest Φ values. We predict that the nucleation region for the folding assembly is centered around the turn that connects the two helices.

Antigen–antibody complexation: a possible role of water at the binding transition state

Since Fisher's time, components of the immune system have been extensively used as paradigms of protein–protein interactions.[86] X-ray structures of the complexes have guided kinetic and site-directed mutagenesis investigations of several antigen–antibody complexes, including that of the chicken lysozyme and its specific antibody,[87,88] to decipher the high affinity and specificity of the antibody to the antigen. The antibodies recognize their target antigens using their variable domains, along with fragments of the variable regions. Fab molecules, which are the fragments of antibodies, are composed of two polypeptide chains (light and heavy), each composed of variable and constant domains. Fv molecules consist of light and heavy chain

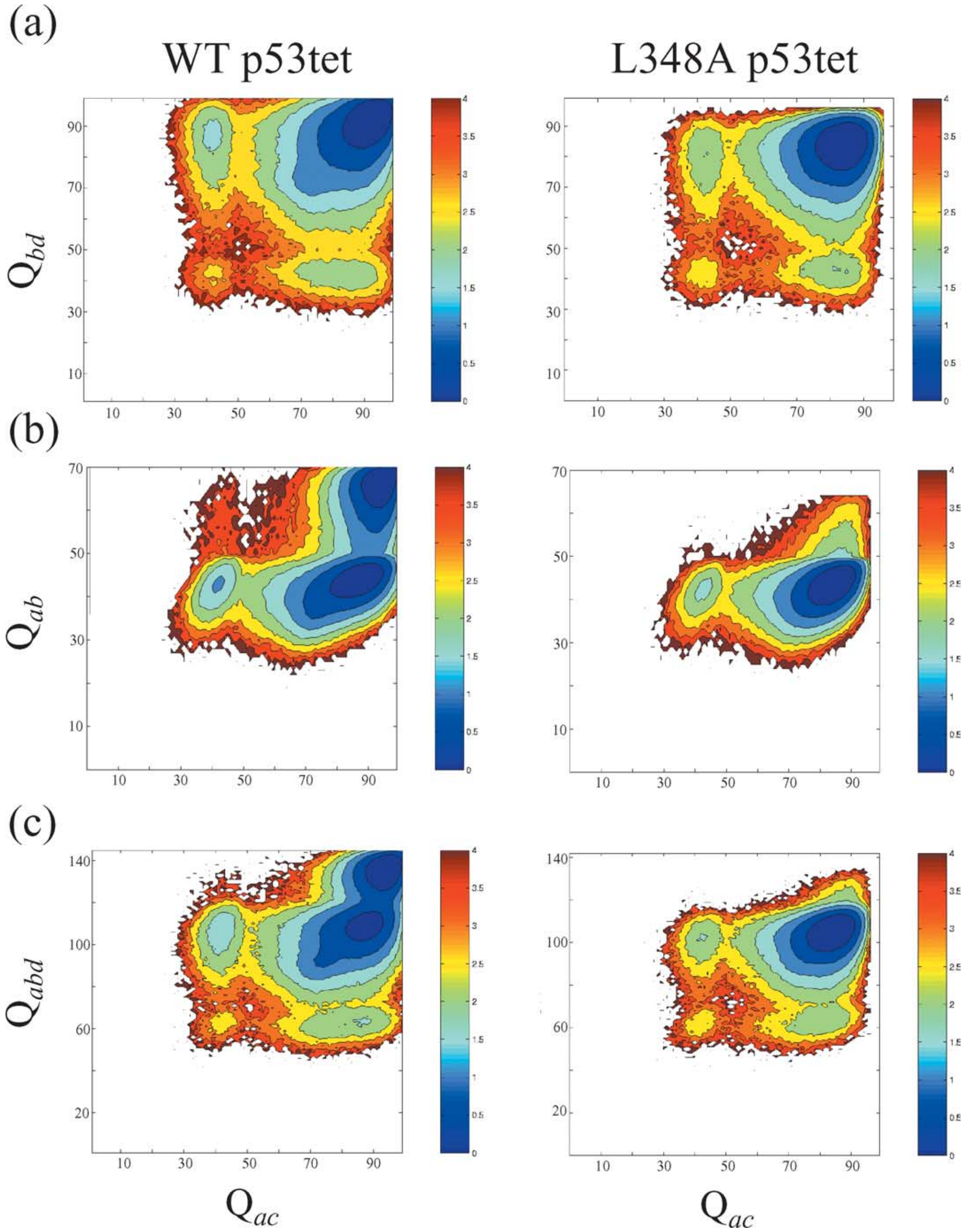

Figure 13. The free energy surface for assembly of WT and L348A p53tet. The free energy surfaces are plotted against (a) the reaction coordinates for formation of the dimers *ac* and *bd*, (b) the formation of the dimer *ab*, and (c) the formation of the trimer *abd*.

variable domains (VL and VH) of antibodies only and are thus approximately half the molecular mass of Fab (see Figure 16). The Fv molecule exhibits antigen binding specificity and affinity similar to those for Fab fragments. Extensive amino acid replacements have been introduced into the lysozyme (HEWL)-HyHEL-10 Fab complex to evaluate the free energy contributions of the chicken lysozyme epitope residues.[87,89]

To study the transition state ensemble for forming the antigen–antibody complex, we simulated the lysozyme (HEWL)-HyHEL-10 Fab complex (PDB entry 3HFM) at a temperature range where both the HEWL and Fab molecules are folded, but undergo binding/unbinding transitions. The recognition between HEWL and Fab results from the formation of 29 and 23 contacts between the lysozyme and VL and VH, respectively. The calculated Φ values for the 19 HEWL residues at the interface with Fab were calculated. Grouping the Φ values into two sets of values between $0.25 < \Phi < 0.60$ and $0.60 < \Phi < 0.89$ reveals a

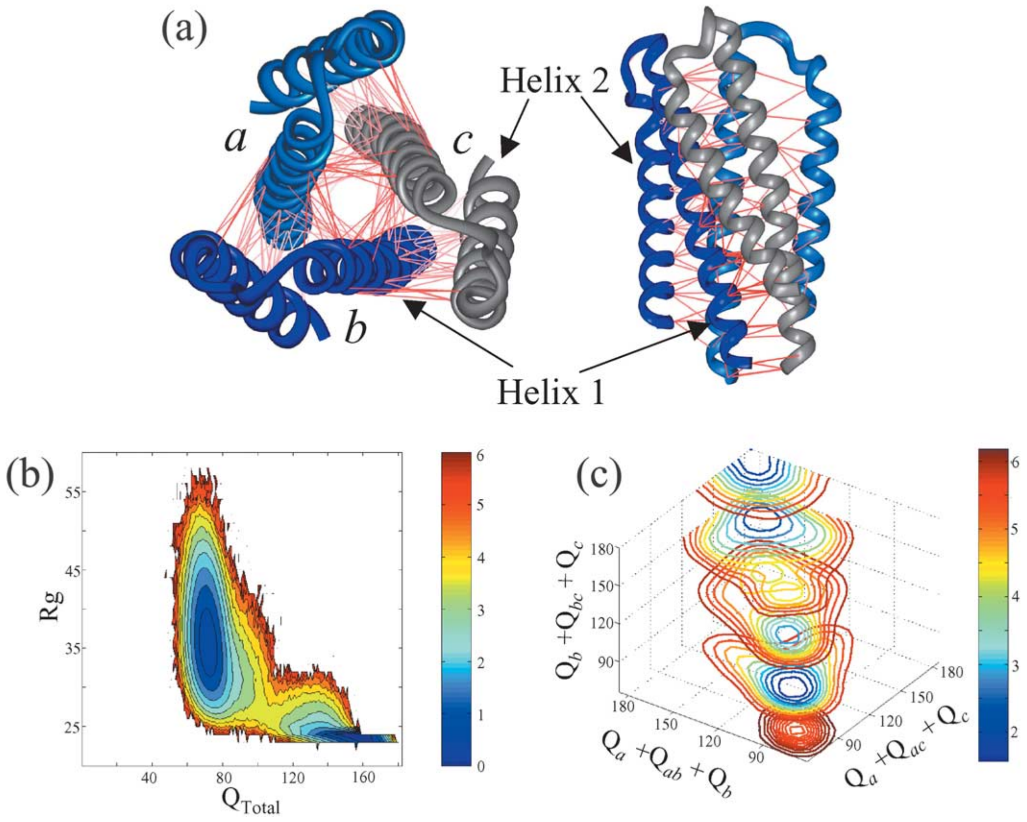

Figure 14. The trimerization of HIV-1 gp41. Each of the identical monomers is composed of two helices. (a) The interfacial interactions are shown by the red lines. The binding free energy surface is projected along Q_{Total}, and (b) the radius of gyration of the trimer and (c) the reaction coordinates for dimerization reactions (formation of the dimers *ab*, *ac*, and *bc*).

distinct correlation with the lysozyme interface. The low Φ values correspond to residues that interact with VH, while the higher Φ values correspond to residues that interact with VL. These Φ values present a polarized scenario where, at the transition state, the interactions with VL are much more formed than with VH. Comparison of the Φ values from our simulations with experimental ones can be done on only a limited basis because Φ values were measured in the lab for only three HEWL residues: R21, K97, and D101.[47] R21 mainly interacts with VL (participates in four contacts with VL and two contacts with VH) and the R97 and D101 residues interact solely with VH (two and three contacts with VH, respectively). The experimental Φ values for K97 and D101 are much more similar to the simulated Φ values than those correspondingly found for R21 (see Table 2). However, due to the low resolution of the Gō model, one may compare the simulated and experimental Φ values for

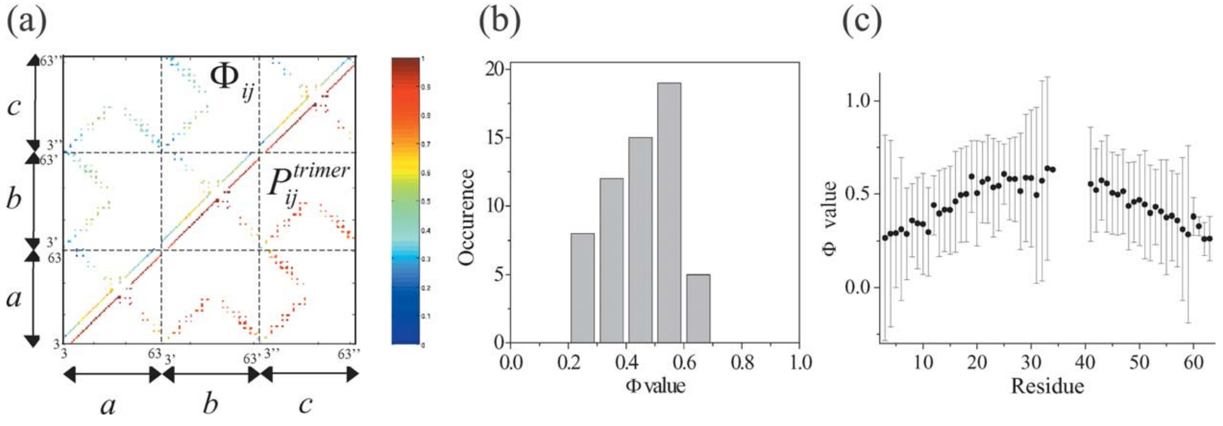

Figure 15. Φ value analysis for the trimerization of HIV-1 gp41. The Φ values are shown (a) per contact, (b) as a histogram, and (c) per residue.

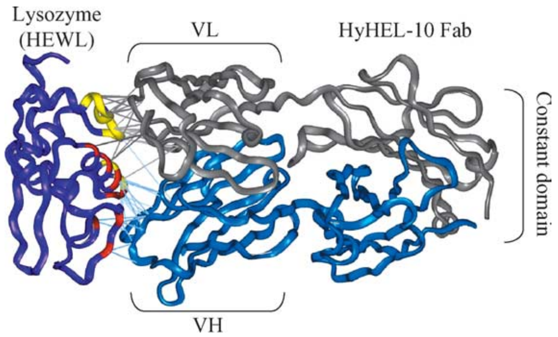

Figure 16. The complex between hen egg-white lysozyme (HEWL) and the Fab antibody. The Fab is composed of light (VL, colored grey) and heavy (VH, colored blue) chains. Interfacial contacts between HEWL and VL and VH are colored grey and blue, respectively. Accordingly, the interfacial contacts between HEWL and VL and VH are shown by grey and blue lines, respectively. The HEWL residues with Φ value larger than 0.6 are colored yellow and those below 0.6 are colored red.

regions of residues rather than for a single isolated position.

The different Φ values obtained for position 21 from the topology based model and the experimental studies on the HEWL-Fab complex (0.84 and 0.14, respectively) can be explained in the light of the crystal structure of HEWL with Fv. The overall structures of the complexes of HEWL with Fv and Fab, which includes in addition to the variable domains, VL and VH that constitute Fv, also a constant domain, are similar. However, at the interface of HEWL-Fv, 11 water molecules were found, while the complex with Fab is almost dry and contains only a single water molecule.[90–93] The water molecule at the interface of HEWL with Fab mediates a contact between the lysozyme with VH. The numbers of water-mediated interactions of the lysozyme made with the VH and VL domains of Fv are four and 11, respectively. Clearly, the water molecules at the interface of the complex HEWL-Fv are important in mediating hydrogen bonds, especially between the lysozyme and the antibody VL domain and indicate a pure shape complementarity between the HEWL and VL. The involvement of water in stabilizing the complex with Fv suggests that the deletion of the constant region results in an unfavorable effect that is compensated by solvent-mediated hydrogen bond formation that may lead to a minimum decreased affinity. The role of the constant domain of Fab in the recognition of Fab to HEWL is supported by normal mode dynamics.[94]

Table 2. Φ values for complexation of FAB with HEWL

HEWL residue	Simulated Φ value	Experimental Φ value
VL		
13	0.67	NA
14	0.68	NA
15	0.70	NA
16	0.68	NA
17	0.89	NA
18	0.79	NA
20	0.83	NA
21	0.84	0.14
22	0.81	NA
96	0.68	NA
100	0.69	NA
VH		
63	0.50	NA
73	0.35	NA
75	0.25	NA
89	0.36	NA
93	0.52	NA
97	0.57	0.36
101	0.52	0.36
102	0.44	NA

The abundance of water in mediating contacts of the lysozyme with VL domain in the absence of constant domain (Fv fragment) can explain the discrepancy between the Φ value at position 21 from experiments and the Gō simulations at the binding transition state of the HEWL-Fab complex. Although the interface between lysozyme with Fab is very dry, the presence of several solvent-mediated contacts between R21 and its neighbor residues and the VL domain of Fv fragment may indicate an important role of the water in effecting antigen–antibody association. Solvent molecules, thus, can assist the initial association to form the encounter complex or, alternatively, the main binding transition state, which will be squeezed out at a later stage and result in a dry interface, which is stabilized by shape complementarity. The latter is probably affected in the case of HEWL-Fv due to the lack of the constant domain, which is compensated by water molecules at the interface that improve lack of structural fitting. Accordingly, the complex HEWL-Fv serves as a snapshot in the recognition dynamics between the antibody and antigen. We propose that the interface of HEWL with the VL domain is much more solvated at the binding transition state than the interface of VH with HEWL. This may explain the lower Φ value found for R21 that cannot be captured by a simple Gō model.

Conclusions

Protein topology, currently well accepted as a pivotal factor in determining unimolecular folding, also determines many aspects of protein assembly. The topology of a protein complex is characterized by the network of non-covalent residue–residue interactions that exist within each chain and between the chains. The density of contacts, as well as the average clustering coefficient and the mean shortest path-length, were found to differentiate between proteins that first fold and then bind and proteins that, in contrast, fold and bind simultaneously. Beyond this global structural analysis, we have shown that the network of native interactions of a few complexes (including nine dimers, a trimer, and a tetramer) determines the funnel energy landscape that can be used to study the residue-specific dynamics of assembly. Simulations based on these landscapes, which uniformly treat all the intra and inter-chain contacts, reproduce the gross features of binding regarding the coupling between folding and binding. Further support for the role of topology in protein assembly can be found in a recent study that has shown that the folding rate of a two-state homo-heptamer can be predicted based on the topology of the native monomer.[95] The important role of protein flexibility for binding kinetics is manifested by the fly-casting mechanism exhibited by intrinsically unfolded monomers that fold upon association. Inhibiting monomer flexibility results in a shallower fly-casting effect and gives slower binding. Flexibility is an inevitable ingredient of protein recognition. Its importance is reflected also by the instances of monomer folding on an already folded monomer. In these situations, one partner serves as a template. An asymmetric binding pathway for forming even symmetric dimers is also formed where at the binding TSE one monomer is more structured than the other monomer.

The structures of the various transition state ensembles were evaluated from these native topology based simulations and used to calculate the contact and residue Φ values. A direct comparison between the simulated and experimental Φ values was possible for three molecular systems. The simulated Φ values exhibit a general correlation with the experimental Φ values. This suggests that the structure of the binding transition state can indeed be obtained with the knowledge of the final complex's structure alone. For the other protein complexes, the topology based model simulations agreed with experimental findings of whether a folded monomer constitutes a populated intermediate state. The simulated Φ values at the binding TSE for these less-studied assemblies may serve as predictions to further test the thesis that minimal frustration and funneled landscapes dominate protein–protein recognition across the genome. While the topological model captures the gross binding mechanism and the finer features of the binding transition state for most systems, the simulation of the antibody–antigen complex points out the possible additional role of water molecules in recognizing rigid proteins. The magnitude of desolvation, electrostatic, and trapping effects in protein binding have to be quantified and compared with the role of topology dominance and structural complementarity in protein assembly.

Models and Method

The studied protein complexes

With the aim of surveying a range of mechanisms of protein association, the complexes selected included dimers, trimers, and a tetramer. In all cases, at least some experimental evidence for the mechanism of assembly was available. The selected complexes follow different binding routes and have varying interface topology. The studied set of dimeric proteins include three homodimers that bind concurrently with their monomer folding (Arc-repressor (1arr, residues $[1–53]_2$), troponin C site III (1cta, residues $[1–34]_2$), and FIS dimer (1f36, residues $[1–34]_2$)), one homodimer system that associates through an on-pathway dimeric intermediate (Trp repressor (2wrp, residues $[8–108]_2$)), and three homodimers for which monomer folding is prerequisite to their association (λ Cro repressor (1cop, residues $[1–66]_2$), LFB1 transcription factor (1lfb, residues $[13–89]_2$), and λ repressor (1lmb, residues $[6–92]_2$)). Bovine seminal ribonuclease (BS-RNase, 11ba, residues $[1–124]_2$), which adopts two quaternary structures (association of fully folded monomers designated M=M and a domain-swapped structure designated as M×M),[96,97] represents association *via* domain-swapping. Since a structure of M=M is not available, its structure was modeled based on the coordinates of the M×M conformation.[75] In addition, the complex between hen egg-white lysozyme (HEWL) and its Fab antibody (PDB entry 3hfm) that associates as folded subunits was studied. The trimeric protein investigated was the HIV gp41 envelope protein, which shows two-state thermodynamics (1i5x, residues $[3–63]_3$). The tetrameric domain of p53 (1sak residues $[326–356]_4$), which has been classified as a dimer of dimers, was studied as well.

In addition to the oligomeric proteins mentioned above, a data set of 122 non-redundant homodimers, which was generated by Janin and his colleagues,[64] was studied. The thermodynamics and kinetics of the association of the majority of these homodimers have not yet been studied. Analyzing the structures of these 122 homodimers together with those where the binding mechanism is known may indicate the relative frequency of occurrence of the various binding mechanisms as reflected by the Protein Data Bank (PDB).

Structural and topological analysis of protein complexes

Various average structural properties of the monomers, as well as of the interfaces, were analyzed. The dimer structure may be crudely described by the number of interactions in the folded state (the total number of native contacts). Two definitions were applied to calculate the number of native contacts. In the first definition, an interaction between a pair of residues (*i,j*) exists if the distance between the C^α atoms of residues i and j is less than 8 Å or if the distance between any side-chain heavy atoms in the two residues is smaller than 4 Å. In the second definition, the residues i and j in the native structure are considered to be in contact according to the Contacts of Structural Units (CSU) software,[98] which is available from the PDB. Intra-monomeric native contacts between pairs of residues (*i,j*) with $|i-j|<4$ were discarded from the native contact list, because any three or four contiguous residues already interact through the angle and dihedral terms. A set of contacts stabilizing the native oligomeric protein defines the protein topology and its subsets address the topology of the complex subunits and interfaces. In general, a larger set of native contacts is obtained by using the CSU definition of native contact, yet, a similar trend is obtained when comparing sets of proteins by either definition. A water-mediated interaction is defined by there being a crystallographic water molecule where an oxygen atom is within a distance of 4 Å from at least a single heavy atom of two residues and that the distance between their C^{β} atoms (or C^{α} for glycine) is between 4 Å and 9.5 Å.

In all cases, the resulting map of native interactions for each protein complex was used to evaluate the average number of contacts in and between the monomers. The occurrence of amino acid residues that participate in interfacial contacts was used to evaluate the interface hydrophobicity after scaling by the residue hydrophobicity factor.[99] To further characterize the contact map of each protein (i.e. to quantify the protein topology), we analyzed the selected complexes by calculating network parameters that reflect the connectivity of the network of native contacts. In this analysis, each residue is treated as a node in the network and an edge between two nodes is said to exist if the two residues associated with that node interact based on the definition used for a native contact. The clustering coefficient for a node i, C_i, reflects the fraction of nodes that are connected to a given node and are interacting between themselves, thus measures the local density of contacts. For residue i the clustering coefficient is given by:

$$C_i = \sum_{j=1}^{N}\sum_{k=1}^{N}\frac{A_{ij}A_{ik}A_{kj}}{2N_i(N_i-1)} \qquad (1)$$

Here, N is the protein length, and A_{ij} is equal to 1 if a contact is defined between residues i and j, otherwise it equals 0. The average clustering coefficient for a certain set of residues (e.g. residues at the complex interface) is calculated by averaging the corresponding residue clustering coefficients. The second average calculated is the mean shortest path-length, L, which is the average over the minimum number of connections that must traverse to connect residue pair i and j. The mean shortest path-length of a given contact map is:

$$L = \frac{1}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}L_{ij} \qquad (2)$$

where L_{ij} reflects the shortest path that connects node (residue) i to node j, L is a global parameter that has often been used to measure network efficiency. Due to the sequence discontinuity for oligomeric proteins, L is calculated here only for the complex subunits. These two parameters have been reported to be modestly powerful in analyzing protein dynamics and to predict the folding transition state.[100–103] In this study, the network analysis is used to characterize the residue connectivity in the monomer and the interface, and to differentiate between protein complexes that follow different association mechanisms.

Association simulation model

The dynamic nature of the association of flexible proteins was studied for the selected protein complexes using a native topology based model (Gō model). This model takes into account only interactions that exist in the native structure and, therefore, does not include energetic frustration (or, alternatively, includes only topological frustration). The Gō-model has already been used to study the folding of many monomeric proteins that fold in a two-state fashion. A significant correlation is found between the experimental folding rates and the rates (or the free energy barrier heights) obtained from the topology based simulations.[104,105] In addition, the Gō model was successfully used to predict intermediates observed experimentally during the folding of larger proteins.[106,107] Recently, the effect of gatekeepers on the folding rate of the ribosomal protein S6 has also been reproduced by topology based model simulations.[108] The impressive agreement between the results for perfectly funneled energy landscapes, and experimental studies strongly indicates the validity of the idea that proteins are energetically minimally frustrated for both folding and binding. However, to resolve quantitative discrepancies between the minimalist Gō models based on pair contacts, and the folding and binding behavior found in the laboratory, creating supplemental minimalist models is unavoidable. Such supplements may include adding cooperativity,[109–111] introducing roughness to the perfectly funneled landscape (adding non-native interactions),[112] and introducing the solvent environment.[113,114]

Here, we used an off-lattice Gō model, where each residue is represented by a single bead centered on its α-carbon (C^{α}) position.[106] Adjacent beads are strung together into a polymer chain by means of a potential encoding bond length and angle constraints. The secondary structure is encoded in the dihedral angle potential and the non-bonded (native contact) potential. In the framework of the model, all native contacts are represented by the 10–12 Lennard-Jones form without any discrimination between the various chemical types of interaction. The details of the model, including its parameters, can be found in previous studies.[8,75,76] To enhance the sampling of binding events, a constraint is applied on the complex subunits. The first constraint, which was applied in studying dimers only, is a polyglycine linker, and the second type of constraint is a harmonic potential on the distance between the center of mass of the subunits. The force constant used to constrain the distance between the center of mass is 0.04, i.e. 2500 times smaller than the force constant for bond between two adjacent residues. The linker holds the two unbound subunits (folded or unfolded) in close proximity during their motions; essentially the local concentration is enhanced. The linker's length was determined by the distance between the C terminus of subunit A and the N terminus of subunit B. This length is sufficient to ensure that the linker will not interfere with any intra or intersubunit contacts that stabilize the folded dimer. To optimize its conformation with respect to the dimer, a minimization was performed on the linker including the two residues to which the linker is directly connected. Covalently linked Arc repressor[67] has been experimentally found to be fully functional with an enhanced folding rate and stability, suggesting indeed that the linker plays a passive, largely entropic role of keeping the unbound monomers at high local concentrations during folding. To further ensure the linker's role is only entropic, the linker residues have no non-bonded interaction (native contacts) with either subunit. All the parameters for the bonded terms of the linker residues were chosen to be smaller by one order of magnitude to enhance its flexibility and to reduce its energetic contributions. To check the effect of the linker length on the association mechanisms simulation were performed for different linker sizes,[8] and also when a harmonic constraint was applied to prevent the center of mass distance of the two subunits from becoming greater than twice its value in the native complex.

For the antibody–antigen complex, studying the folding of each subunit is very computationally expensive, due to the size of the system (the antigen and antibody are composed of 129 and 429 residues, respectively). For this reason and because this complex is predicted by the phase diagram (Figure 1) to bind only after folding, its simulations were done at a temperature where only binding takes place. The tetrameric domain of p53, which is simulated starting from four unfolded monomers, was additionally simulated with models where constraints are applied to limit the monomer flexibility. In one case, the monomer was set to be relatively rigid by fixing all the monomeric native contacts to be permanently formed. In the second case, the dimers (*ac* and *bd*) were rigid and cannot dissociate into two monomers. These models were designed to address the role of flexibility in binding and to quantify the magnitude of the fly-casting effect, which is monitored by plotting the free energy as a function of the distance between the association subunits.

For each system, several constant temperature molecular dynamics simulations were performed (using the simulation package AMBER6 as an integrator†) starting from either the folded dimeric conformation or the unfolded and unbound monomers. At least a single binding/unbinding transition was required to ensure an equilibrated sampling. The multiple trajectories were combined using the weighted histogram analysis method (WHAM)[115] to provide the transition temperatures from the peaks of the specific heat *versus* temperature and to calculate thermodynamic properties of the systems. The temperatures and the free energy are in units of ε, the stability gain from formation of a single native contact. The free energy surface of a binding process is projected onto several candidate reaction coordinates for folding and binding: the fractions of monomeric native contacts, interfacial native contacts, the total number of native contacts, and the distance between the center of mass of the two subunits. In the free energy calculations, the energy terms associated with the linker residues were neglected to enable a comparison between a dimer and an isolated monomer folding.

To probe the nature of the transition state ensemble, we computed the ϕ_{ij} value for each native contact pair between i and j from the probability of formation, P_{ij}:

$$\phi_{ij} = \frac{\Delta\Delta F^{\mathrm{TS-U}}}{\Delta\Delta F^{\mathrm{F-U}}} \approx \frac{P_{ij}^{\mathrm{TS}} - P_{ij}^{\mathrm{U}}}{P_{ij}^{\mathrm{F}} - P_{ij}^{\mathrm{U}}} \qquad (3)$$

where $\Delta\Delta F$ is the free energy difference between the wild-type and mutant protein, P_{ij} is the probability of formation of contact between i and j. For folding of a monomeric protein the subscripts F, U, and TS correspond to folded, unfolded, and folding transition state ensembles, respectively. Similarly, for a binding reaction, the subscripts F, U, and TS correspond to the folded dimer, unfolded monomers, and binding transition state ensembles, respectively. Since all non-bonded contacts in the Gō model have the same energies, the ϕ_i value of residue i can be calculated from the contact values, ϕ_{ij}, by averaging all the ϕ_{ij} values that are involved with residue i:

$$\phi_i = \frac{1}{n}\sum_{j}^{n} \phi_{ij} \qquad (4)$$

The computational ϕ_i value prediction can be compared with the experimental data once they are available. The Φ values were calculated based on the use of Q, the fraction of native contacts, as a reaction coordinate. Q successfully distinguishes between F, U, and TS ensembles and has been shown to reasonably reproduce experimental Φ values.[106,116,117]

† amber.scripps.edu/doc6/

Acknowledgements

This work has been funded by the NSF sponsored Center for Theoretical Biological Physics (grants PHY-0216576 and 0225630) with additional support from MCB-0084797. Computations were carried out at the UCSD KeckII computing facility (partially supported by NSF-MCB). S.S.C is supported by a University of California at San Diego Molecular Biophysics Training Grant.

References

1. Gavin, A.-C., Bosche, M., Krause, R., Grandi, P. & Marzioch, M. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
2. Ho, Y., Gruhler, A., Heilut, A., Bader, G. D., Moore, L. & Adams, S.-L. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
3. Salwinski, L. & Eisenberg, D. (2003). Computational methods of analyses of protein–protein interactions. *Curr. Opin. Struct. Biol.* **13**, 377–382.
4. Janin, J. & Seraphin, B. (2003). Genome-wide studies of protein–protein interaction. *Curr. Opin. Struct. Biol.* **13**, 383–388.
5. Fischer, E. (1894). Einfluss der configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.* **27**, 2985–2991.
6. Koshland, D. E. J. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl Acad. Sci. USA*, **44**, 98–123.
7. Frauenfelder, H., Sliger, S. G. & Wolynes, P. G. (1991). The energy landscapes and motions of proteins. *Science*, **254**, 1598–1603.
8. Levy, Y., Wolynes, P. G. & Onuchic, J. N. (2004). Protein topology determines binding mechanisms. *Proc. Natl Acad. Sci. USA*, **101**, 511–516.
9. Verkhivker, G. M., Bouzida, D., Gehlhaar, D. K., Rejto, P. A., Freer, S. T. & Rose, P. W. (2003). Simulating disorder–order transitions in molecular recognition of unstructured proteins: where folding meets binding. *Proc. Natl Acad. Sci. USA*, **100**, 5148–5153.
10. Gupta, N. & Irback, A. (2004). Coupled folding-binding *versus* docking: a lattice model study. *J. Chem. Phys.* **120**, 3983–3989.
11. Ma, B., Kumar, S., Tsai, C.-J. & Nussinov, R. (1999). Folding funnels and binding mechanisms. *Protein Eng.* **12**, 713–720.
12. Bosshard, H. R. (2001). Molecular recognition by induced fit: how fit is the concept? *News Physiol. Sci.* **16**, 171–173.
13. Ma, B. Y., Shatsky, M., Wolfson, H. & Nussinov, R. (2002). Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.* **11**, 184–197.
14. Goh, C.-S., Milburn, D. & Gerstein, M. (2004). Conformational changes associated with protein–protein interactions. *Curr. Opin. Struct. Biol.* **14**, 1–6.
15. Foote, J. & Milstein, C. (1994). Conformational isomerism and the diversity of antibodies. *Proc. Natl Acad. Sci. USA*, **91**, 10370–10374.
16. Berger, C., Weber-Bornhauser, S., Eggenberger, J.,

Hanes, J., Plucjthun, A. & Bosshard, H. R. (1999). Antigen recognition by conformational selection. *FEBS Letters*, **450**, 149–153.

17. Volkman, B. F., Lifson, D., Wemmer, D. E. & Kern, D. (2001). Two-state allosteric behavior in a single-domain signaling protein. *Science*, **291**, 2429–2433.
18. Jeffery, C. J. (1999). Moonlighting proteins. *Trends Biochem. Sci.* **24**, 8–11.
19. Copley, S. D. (2003). Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol.* **7**, 265–272.
20. James, L. C. & Tawfik, S. S. (2003). Conformational diversity and protein evolution—a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **28**, 361–368.
21. Bennett, M. J., Schlunegger, M. P. & Eisenberg, D. (1995). 3D domain swapping: a mechanism for oligomer assembly. *Protein Sci.* **4**, 2455–2468.
22. Liu, Y. & Eisenberg, D. (2002). 3D domain swapping: as domains continue to swap. *Protein Sci.* **11**, 1285–1299.
23. Bryngelson, J. D. & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA*, **84**, 7524–7528.
24. Leopold, P. E., Montal, M. & Onuchic, J. N. (1992). Protein folding funnels: a kinetic approach to the sequence–structure relationship. *Proc. Natl Acad. Sci. USA*, **89**, 8721–8725.
25. Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z. & Socci, N. D. (1995). Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl Acad. Sci. USA*, **92**, 3626–3630.
26. Taverna, D. M. & Goldstein, R. A. (2002). Why are proteins so robust to site mutations? *J. Mol. Biol.* **315**, 479–484.
27. Yang, S., Cho, S., Levy, Y., Cheung, M.S., Levine, H., Wolynes, P. G. & Onuchic, J. N. (2004). Domain Swapping is a consequence of the principle of minimal frustration. *Proc. Natl Acad. Sci. USA*, **101**, 12786–12791.
28. Rousseau, F., Schymkowitz, J. W. H. & Itzhaki, L. S. (2003). The unfolding story of three-dimensional domain swapping. *Structure*, **11**, 243–251.
29. Neet, K. E. & Timm, D. E. (1994). Conformational stability of dimeric proteins: quantitative studies by equilibrium denaturation. *Protein Sci.* **3**, 2167–2174.
30. Xu, D., Tsai, C.-J. & Nussinov, R. (1998). Mechanism and evolution of protein dimerization. *Protein Sci.* **7**, 533–544.
31. Dyson, H. J. & Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **12**, 54–60.
32. Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27**, 527–533.
33. Uversky, V. N. (2002). Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756.
34. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S. *et al.* (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59.
35. Papoian, G. A. & Wolynes, P. G. (2003). The physics and bioinformatics of binding and folding—an energy landscape perspective. *Biopolymers*, **68**, 333–349.
36. Kirwacki, R., Hengst, L., Tennant, L., Reed, S. & Wright, P. (1996). Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl Acad. Sci. USA*, **93**, 11504–11509.
37. Spolar, R. & Record, M. (1994). Coupling of local folding to site-specific binding of protein to DNA. *Science*, **263**, 777–784.
38. Gunasekaran, K., Tsai, C.-J., Kumar, S., Zanuy, D. & Nussinov, R. (2003). Extended disordered proteins: targeting function with less scaffold. *Trends Biochem. Sci.* **28**, 81–85.
39. Shoemaker, B. A., Portman, J. J. & Wolynes, P. G. (2000). Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl Acad. Sci. USA*, **97**, 8868–8873.
40. Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J. E., Vajda, S. *et al.* (2003). CAPRI: a critical assessment of predicted interactions. *Proteins: Struct. Funct. Genet.* **52**, 2–9.
41. Ben-Zeev, E., Berchanski, A., Heifetz, A., Shapira, B. & Eisenstein, M. (2003). Prediction of the unknown: inspiring experience with the CAPRI experiments. *Proteins: Struct. Funct. Genet.* **52**, 41–46.
42. Mendez, R., Laplae, R., Maria, L. D. & Wodak, S. J. (2003). Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins: Struct. Funct. Genet.* **52**, 51–67.
43. Fersht, A. R. (1994). Characterizing transition states in protein folding: an essential step in the puzzle. *Curr. Opin. Struct. Biol.* **5**, 79–84.
44. Daggett, V. & Fersht, A. R. (2003). The present view of the mechanism of protein folding. *Nature Rev. Mol. Cell Biol.* **4**, 497–502.
45. Fersht, A. R. & Sato, S. (2004). *F*-values analysis and the nature of protein-folding transition states. *Proc. Natl Acad. Sci. USA*, **101**, 7976–7981.
46. Milla, M. E., Brown, B. M., Waldburger, C. D. & Sauer, R. T. (1995). P22 Arc repressor: transition state properties inferred from mutational effects on the rates of protein unfolding and refolding. *Biochemistry*, **34**, 13914–13919.
47. Taylor, M. G., Rajpal, A. & Kirsch, J. F. (1998). Kinetic epitope mapping of the chicken lysozyme HyHEL-10 Fab complex: delineation of docking trajectories. *Protein Sci.* **7**, 1857–1867.
48. Mateu, M. G., Pino, M. M. S. D. & Fersht, A. R. (1999). Mechanism of folding and assembly of a small tetrameric protein domain from tumor suppressor p53. *Nature Struct. Biol.* **6**, 191–198.
49. Nolting, B. & Andert, K. (2000). Mechanism of protein folding. *Proteins: Struct. Funct. Genet.* **41**, 288–298.
50. Wu, L. C., Tuot, D. S., Lyons, D. S., Garcia, K. C. & Davis, M. M. (2002). Two-step binding mechanism of T-cell receptor recognition of peptide-MHC. *Nature*, **418**, 552–555.
51. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 539–594.
52. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. & Wolynes, P. G. (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold. Des.* **1**, 441–450.
53. Tsai, C.-J., Kumar, S., Ma, B. & Nussinov, R. (1999). Folding funnels, binding funnels, and protein function. *Protein Sci.* **8**, 1181–1190.
54. Zhang, C., Chen, J. & DeLisi, C. (1999). Protein–protein recognition: exploring the energy funnels near the binding sites. *Proteins: Struct. Funct. Genet.* **34**, 255–267.
55. Tovchigrechko, A. & Vakser, I. A. (2001). How

common is the funnel-like energy landscape in protein–protein interactions? *Protein Sci.* **10**, 1572–1583.

56. Kumar, S., Ma, B., Tsai, C.-J., Sinha, N. & Nussinov, R. (2000). Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci.* **9**, 10–19.
57. Verkhivker, G. M., Bouzida, D., Gehlhaar, D. K., Rejto, P. A., Freer, S. T. & Rose, P. W. (2002). Complexity and simplicity of ligand–macromolecule interactions: the energy landscape perspective. *Curr. Opin. Struct. Biol.* **12**, 197–203.
58. Wang, J. & Verkhivker, G. M. (2003). Energy landscape theory, funnels, specificity, and optimal criterion of biomolecular binding. *Phys. Rev. Letters*, **90**, 188101.
59. Papoian, G. A., Ulander, J. & Wolynes, P. G. (2003). Role of water mediated interactions in protein–protein recognition landscapes. *J. Am. Chem. Soc.* **125**, 9170–9178.
60. Papoian, G. A., Ulander, J., Eastwood, M. E. & Wolynes, P. G. (2004). Water in protein structure prediction. *Proc. Natl Acad. Sci. USA*, **101**, 3352–3357.
61. Jones, S. & Thornton, J. M. (1996). Principles of proetin–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
62. Conte, L. L., Chotia, C. & Janin, J. (1999). The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
63. Ofran, Y. & Rost, B. (2003). Analyzing six types of protein–protein interfaces. *J. Mol. Biol.* **325**, 377–387.
64. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2003). Dissecting subunit interfaces in homodimeric proteins. *Proteins: Struct. Funct. Genet.* **53**, 708–719.
65. Sheinerman, F. B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153–159.
66. Schreiber, G. (2002). Kinetic studies of protein–protein interactions. *Curr. Opin. Struct. Biol.* **12**, 41–47.
67. Robinson, C. R. & Sauer, R. T. (1996). Equilibrium stability and sub-millisecond refolding of a designed single-chain Arc repressor. *Biochemistry*, **35**, 13878–13884.
68. Bowie, J. U. & Sauer, R. T. (1989). Equilibrium dissociation of the Arc repressor dimer. *Biochemistry*, **28**, 7139–7143.
69. Monera, O. D., Shaw, G. S., Zho, B.-Y., Sykes, B. D., Kay, C. M. & Hodges, R. S. (1992). Role of interchain a-helical hydrophobic interactions in Ca2+ affinity, formation, and stability of a two-site domain in troponin C. *Protein Sci.* **1**, 945–955.
70. Hobart, S. A., Ilin, S., Moriarty, D. F., Osuna, R. & Colon, W. (2002). Equilibrium denaturation studies of the *Escherichia coli* factor for inversion stimulation: implication for *in vivo* function. *Protein Sci.* **11**, 1671–1680.
71. Gloss, L. M. & Matthews, C. R. (1998). The barriers in the bimolacular and unimolecular folding reactions of the dimeric core domain of *Escherichia coli* Trp repressor are dominated by enthalpic contributions. *Biochemistry*, **37**, 16000–16010.
72. Huang, M. & Oas, T. (1995). Submillisecond folding of monomeric l repressor. *Proc. Natl Acad. Sci. USA*, **92**, 6878–6882.
73. Jana, R., Hazbun, T. R., Mollah, A. K. M. M. & Mossing, M. C. (1997). A folded monomeric intermediate in the formation of lambda Cro dimer–DNA complexes. *J. Mol. Biol.* **273**, 402–416.
74. Leiting, B., Francesco, R. D., Tomei, L., Cortese, R., Otting, G. & Wuthrich, K. (1993). The three-dimensional NMR-solution structure of the polypeptide fragment 195–286 of the LFB1/HNF1 transcription factor from rat liver comprises a non-classical homeodomain. *EMBO J.* **12**, 1797–1803.
75. Levy, Y., Papoian, G. A., Onuchic, J. & Wolynes, P. G. (2004). The energy landscape of protein dimers. *Isr. J. Chem.* in the press.
76. Levy, Y., Caflisch, A., Onuchic, J. N. & Wolynes, P. G. (2004). The folding and dimerization of HIV-1 protease: evidence for a stable monomer from simulations. *J. Mol. Biol.* **340**, 67–79.
77. Bryan, P. N. (2002). Prodomain and protein folding catalysis. *Chem. Rev.* **102**, 4805–4816.
78. LeFevre, K. R. & Cordes, M. H. J. (2003). Retroevolution of lambda Cro toward a stable monomer. *Proc. Natl Acad. Sci. USA*, **100**, 2345–2350.
79. Newlove, T., Koniezka, J. H. & Cordes, M. H. (2004). Secondary structure switching in Cro protein evolution. *Structure*, **12**, 569–581.
80. Shakhnovich, E. I. (1999). Folding by association. *Nature Struct. Biol.* **6**, 99–102.
81. Chong, L. T., Snow, C. D., Rhee, Y. M. & Pande, V. S. (2005). Dimerization of the p53 oligomerization domain: identification of a folding nucleus by molecular dynamics simulations. *J. Mol. Biol.* in the press.
82. McCoy, M., Stavridi, E. S., Waterman, J. L. F., Wieczorek, A. M., Opella, S. J. & Halazonetis, T. D. (1997). Hydrophobic side-chain size is a determinant of the three-dimensional structure of the p53 oligomerization domain. *EMBO J.* **16**, 6230–6236.
83. Mateu, M. G. & Fersht, A. R. (1998). Nine hydrophobic side chains are key determinants of the thermodynamic stability and oligomerization status of tumor. *EMBO J.* **17**, 2748–2758.
84. Jelesarov, I. & Lu, M. (2001). Thermodynamics of trimer-of-hairpins formation by the SIV gp41 envelope protein. *J. Mol. Biol.* **307**, 637–656.
85. Marti, D. N., Bjelic, S., Lu, M., Bosshard, H. R. & Jelesarov, I. (2004). Fast folding of the HIV-1 and SIV gp41 six-helix bundles. *J. Mol. Biol.* **336**, 1–8.
86. Braden, B. C. & Poljak, R. J. (2000). Structure and energetics of anti-lysozyme antibodies. In *Protein–protein Recognition* (Leanthous, C., ed.) 1st edit., Oxford University Press, Oxford.
87. Rajpal, A., Taylor, M. G. & Kirsch, J. F. (1998). Quantitative evaluation of the chicken lysozyme epitope in the HyHEL-10 Fab complex: free energies and kinetics. *Protein Sci.* **7**, 1868–1874.
88. Li, Y., Urrutia, M., Smith-Gill, S. J. & Mariuzza, R. A. (2003). Dissection of binding interactions in the complex between the anti-lysozyme antibody HyHEL-63 and its antigen. *Biochemistry*, **42**, 11–22.
89. Pons, J., Rajpal, A. & Kirsch, J. F. (1999). Energetic analysis of an antigen/antibody interface: alanine scanning mutagenesis and double mutant cycles on the HyHEL-10/lysozyme interaction. *Protein Sci.* **8**, 958–968.
90. Bhat, T. N., Bentley, G. A., Boulot, G., Greene, M. I., Tello, D., Dall'Acqua, W. *et al.* (1994). Bound water molecules and conformational stabilization help mediate an antigen–antibody association. *Proc. Natl Acad. Sci. USA*, **91**, 1089–1093.
91. Goldbaum, F. A., Schwarz, F. P., Eisenstein, E., Cauerhff, A., Mariuzza, R. A. & Pojak, R. J. (1996). The effect of water activity on the association

constant and the enthalpy of reaction between lysozyme and the specific antibodies D1.3 and D44.1. *J. Mol. Recogn.* **9**, 6–12.

92. Kondo, H., Shiroishi, M., Matsushima, M., Tsumoto, K. & Kumagai, I. (1999). Crystal structure of anti-hen egg white lysozyme antibody (HyHEL-10) Fv-antigen complex. *J. Biol. Chem.* **274**, 27623–27631.
93. Yokota, A., Tsumoto, K., Shiroishi, M., Kondo, H. & Kumagai, I. (2003). The role of hydrogen bonding *via* interfacial water molecules in antigen–antibody complexation. *J. Biol. Chem.* **278**, 5410–5418.
94. Adachi, M., Kurihara, Y., Nojima, H., Takeda-Shitaka, M., Kamiya, K. & Umeyama, H. (2003). Interaction between the antigen and antibody is controlled by the constant domains: normal mode dynamics of the HEL-HyHEL-10 complex. *Protein Sci.* **12**, 2125–2131.
95. Bascos, N., Guidry, J. & Wittung-Stafshede, P. (2004). Monomer topology defines folding speed of heptamer. *Protein Sci.* **13**, 1317–1321.
96. Piccoli, R., Tamburrini, M., Piccialli, G., Donato, A. D., Parente, A. & D'Alessio, G. (1992). The dual-mode quaternary structure of seminal RNase. *Proc. Natl Acad. Sci. USA*, **89**, 1870–1874.
97. D'Alessio, G. (1999). Evolution of oligomeric proteins. The unusual case of a dimeric ribonuclease. *Eur. J. Biochem.* **266**, 699–708.
98. Sobolev, V., Wade, R. C., Vriend, G. & Edelman, M. (1996). Molecular docking using surface complementarity. *Proteins: Struct. Funct. Genet.* **25**, 120–129.
99. Pacios, L. (2001). Distinct molecular surfaces and hydrophobicity of amino acid residues in proteins. *J. Chem. Inf. Comput. Sci.* **41**, 1427–1435.
100. Vendruscolo, M., Dokholyan, N. V., Paci, E. & Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E*, **65**, 061910–061911.
101. Dokholyan, N. V., Li, L., Ding, F. & Shakhnovich, E. I. (2002). Topological determinants of protein folding. *Proc. Natl Acad. Sci. USA*, **99**, 8637–8641.
102. Atilgan, A. R., Akan, P. & Baysal, C. (2004). Small-world communication of residues and significance for protein dynamics. *Biophys. J.* **86**, 85–91.
103. Greene, L. H. & Higman, V. A. (2003). Uncovering network systems within protein structures. *J. Mol. Biol.* **334**, 781–791.
104. Koga, N. & Takada, S. (2001). Roles of native topology and chain-length scaling in protein folding: a simulation study with a Go-like model. *J. Mol. Biol.* **313**, 171–180.
105. Chavez, L. L., Onuchic, J. N. & Clementi, C. (2004). Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.* **126**, 8426–8432.
106. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000). Topological and energetical factors: what determines the structural details of the transition state ensemble and "En-route" intermediate for protein folding? An investigation of small globular proteins. *J. Mol. Biol.* **298**, 937–953.
107. Clementi, C., Jennings, P. A. & Onuchic, J. N. (2000). How native-state topology affects the folding of dihydrofolate reductase and interleukin-1 β. *Proc. Natl Acad. Sci. USA*, **97**, 5871–5876.
108. Stoycheva, A. D., Brooks, C. L., III & Onuchic, J. N. (2004). Gatekeepers in the ribosomal protein S6: thermodynamics, kinetics, and folding pathways revealed by a minimalist protein model. *J. Mol. Biol.* **340**, 571–585.
109. Eastwood, M. P. & Wolynes, P. G. (2001). Role of explicitly cooperative interactions in protein folding funnels: a simulation study. *J. Chem. Phys.* **114**, 4702–4716.
110. Chan, H. S., Shimizu, S. & Kaya, H. (2004). Cooperativity principles in protein folding. *Methods Enzymol.* **380**, 350–379.
111. Ejtehadi, M. R., Avall, S. P. & Plotkin, S. S. (2004). Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proc. Natl Acad. Sci. USA*, **101**, 15088–15093.
112. Clementi, C. & Plotkin, S. (2004). The effects of nonnative interactions on protein folding rates: theory and simulation. *Protein Sci.* **13**, 1750–1766.
113. Cheung, M. S., Garcia, A. E. & Onuchic, J. N. (2002). Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl Acad. Sci. USA*, **99**, 685–690.
114. Levy, Y. & Onuchic, J. N. (2004). Water and proteins: a love–hate relationship. *Proc. Natl Acad. Sci. USA*, **101**, 3325–3326.
115. Ferrenberg, A. M. & Swendsen, R. H. (1989). Optimized Monte Carlo data analysis. *Phys. Rev. Letters*, **63**, 1195–1198.
116. Clementi, C., Garcia, A. E. & Onuchic, J. N. (2003). Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism all-atom representation study of protein L. *J. Mol. Biol.* **326**, 879–890.
117. Shea, J.-E., Onuchic, J. N. & Brooks, C. L., III (1999). Exploring the origins of topological frustration: design of a minimally frustrated model of fragment B of protein A. *Proc. Natl Acad. Sci. USA*, **96**, 12512–12517.

Edited by M. Levitt

(Received 1 October 2004; received in revised form 3 December 2004; accepted 9 December 2004)