Critical Reviews in Biochemistry and Molecular Biology, 36(5):399–433 (2001)

Protein Folding: Binding of Conformationally Fluctuating Building Blocks *Via* Population Selection

Chung-Jung Tsai,[1] Buyong Ma,[2] Sandeep Kumar,[2] Haim Wolfson,[3] and Ruth Nussinov[1,4]*

[1]Intramural Research Support Program — SAIC, Laboratory of Experimental and Computational Biology, NCI-Frederick,Bldg 469, Rm 151, Frederick, MD 21702; [2]Laboratory of Experimental and Computational Biology, NCI-Frederick, Bldg 469, Rm 151, Frederick, MD 21702; [3]School of Computer Science, Sackler faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel; [4]Sackler Inst. of Molecular Medicine, Department of Human Genetics and Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

Referee: Ilya A. Vakser, Department of Cell and Molecular Pharmacology, Medical Univerisity of South Carolina 173 Ashley Avenue, PO Box 250505, Charleston, SC 294425, phone: (843)792-2471, fax:(843)792-2475, email:vakseri@musc.edu

Abstract: Here we review different aspects of the protein folding literature. We present a broad range of observations, showing them to be consistent with a general hierarchical protein folding model. In such a model, local relatively stable, conformationally fluctuating building blocks bind through population selection, to yield the native state. The model includes several components: (1) the fluctuating building blocks that constitute local minima along the polypeptide chain, which even if unstable still possess higher population times than all alternate conformations; (2) the landscape around the bottom of the funnels; (3) the consideration that protein folding involves intramolecular recognition; (4) similar landscapes are observed for folding and for binding, and that (5) the landscape is dynamic, changing with the conditions. The model considers protein folding to be guided by native interactions. The reviewed literature includes the effects of changing the conditions, intermediates and kinetic traps, mutations, similar topologies, fragment complementation experiments, fragments and pathways, focusing on one specific well-studied example, that of the dihydrofolate reductase, chaperones, and chaperonines, *in vivo* vs. *in vitro* folding, still using the dihydrofolate example, amyloid formation, and molecular "disorder". These are consistent with the view that binding and folding are similar events, with the differences stemming from different stabilities and hence population times.

KEY WORDS: dynamic landscapes, conformational ensembles, binding, folding, funnels, induced conformational change.

1040-9238/01/$.50

I. INTRODUCTION

In recent decades, considerable effort has focused on the understanding of how a polypeptide chain folds into its native state. The observation that protein folding can proceed without the aid of molecular factors has allowed studies of protein folding both in the test tube, as well as computationally.

In vitro the conditions differ from those *in vivo*. Perhaps the largest difference is with respect to protein concentration. However, in addition other factors such as pH, ion concentration, and temperature may vary as well. The spontaneous folding *in vitro* has enabled studies of protein structure, its folding pathways, kinetics, intermediate states, and the effect of single engineered mutations.

Several models have been proposed to describe the protein folding process. The major ones can be classified into (1) the framework model, (2) the nucleation and growth mechanism, (3) the diffusion-collision model, (4) the hydrophobic collapse, and (5) the hierarchical model. In the (1) framework model,[1-3] secondary structure formation is independent of formation of tertiary interactions, and usually precedes these. If tertiary interactions occur first, then they are not necessarily the native ones. In the second, (2) nucleation and growth[4] or nucleation-condensation mechanism,[5,6] folding initiates by formation of a "nucleus", followed by its extension. It has been proposed that the formation of such a nucleus is dependent on contacts between key residues that have been conserved through evolution. This model has led to searches for specific residue conservation in families of related proteins. In the third (3) model, preformed folded largely secondary structure elements assemble into complete folds though random diffusion and collision.[7] If their assembly is favorable, they may lock to yield the native conformation. In contrast to these models, the hydrophobic collapse (4) model highlights the hydrophobic effect, the driving force of protein folding.[8-10] According to the hydrophobic collapse model, folding initiates by the molecule collapsing, burying extensive non-polar surface area. Secondary structure formation, and specific interactions follow. In the hierarchical model (5), protein folding initiates locally, and hierarchically local, folded elements assemble in a largely stepwise fashion to yield the native fold.[11,12]

Below we illustrate that these models are not necessarily exclusive of each other. The hierarchical model may include elements of hydrophobic collapse in the assembly of local folded elements, consistent with the molten globule state. Such a hydrophobic assembly would constitute the first stage of the elements coming together, followed by the optimization of the specific (van der Waals, electrostatic, disulfide bonds, etc.) interactions. The hierarchical model may further include elements of the nucleation and growth, or nucleation condensation. The nucleation does not necessarily have to be restricted to specific residues. We can imagine that the nucleus can be an element of the polypeptide chain whose folded structure forms local minima. Such an element may then act as a template for further folding of the protein. A nice example of such an event is the proregion of subtilisin, or of α-lytic protease. Similarly, with respect to the framework model, we may substitute single secondary structure element formation by such chain-linked local building block minima. Thus, depending on the interpretation of these models, each may be viewed as a specific case of the more general hierarchical model.

Here we review some of the recent results from the literature, showing that they coherently fit into such a simple, hierarchical protein folding scheme. The experimental observations that we touch on here range from protein folding under different conditions (e.g., Refs. 13 to 18); fragment complementation experiments;[19-33] studies of the effect of mutations on protein folding (e.g., Refs. 34 to 38); intermolecular chaperones and chaperonins (e.g., Refs. 39,40); intramolecular chaperones (Refs. 41 to 48); disorder in protein structures, and stabilization through binding;[50-52] similar pathways in topologically similar proteins,[55,56] despite different stabilities, for example, thermophiles vs. mesophiles;[57] intermediates and kinetic traps;[58-64] and *in vivo* vs. *in vitro* folding. We also address theoretical descriptions such as energy landscapes and folding funnels.[65-72]

For clarity, we first briefly describe the model. We proceed to outline the components on which it is based. Into these we weave the various experimental, and theoretical observations. Figure 1 presents a schematic diagram of its major features. According to the model, protein folding is a hierarchic event.[11,12] At the first step, local *transient* building block elements fold. The conformations they obtain are not necessarily stable, but they have higher population times than all other, alternate conformations. In the next step the building blocks associate, mutually stabilizing each other. The association is via *selection* of the most favorable conformers for binding. Hence, the critical point is that the flexible building block fragments do not induce conformational change in each other. Rather, via selection, the intramolecular complex of the chain-linked conformationally fluctuating building blocks gradually grows in a manner similar to that observed in intermolecular oligomers, or any multimolecule complex formation.

The model is based on a number of elements:[73-81] first, intermolecular binding and intra-molecular folding are similar processes, with similar underlying principles. The sole difference between them is chain connectivity. Second, both processes have similar energy landscapes that can be similarly described by a series of successive fusion events of microfunnels. Third, binding mechanisms are immediately implied from the bottom of the funnels. If the bottom has a steep V shape, the molecules are rigid, and can be expected to be highly specific. On the other hand, if the funnel bottoms are wide and rugged, a broad range of binding is implied. Rugged bottoms imply a range of conformations, with low barriers between them. Such a situation straightforwardly illustrates why there is no need to resort to the age-old (largely mistaken) belief of the “induced fit” mechanism. Fourth, the energy landscape is dynamic, the outcome of changing conditions. This implies that under different conditions, whether physical, or binding to other molecules, we observe population shifts. Fifth, this principle holds for binding cascades, enzyme pathways, and allostery. Sixth, the recently devised building block folding model states that folding initiates with the formation of the conformationally fluctuating building blocks. These constitute local minima along the chain. Through combinatorial assembly, and mutual stabilization, the building blocks associate to form the independently folding, compact, hydrophobic folding units, with a strong hydrophobic core. These further associate to form domains,

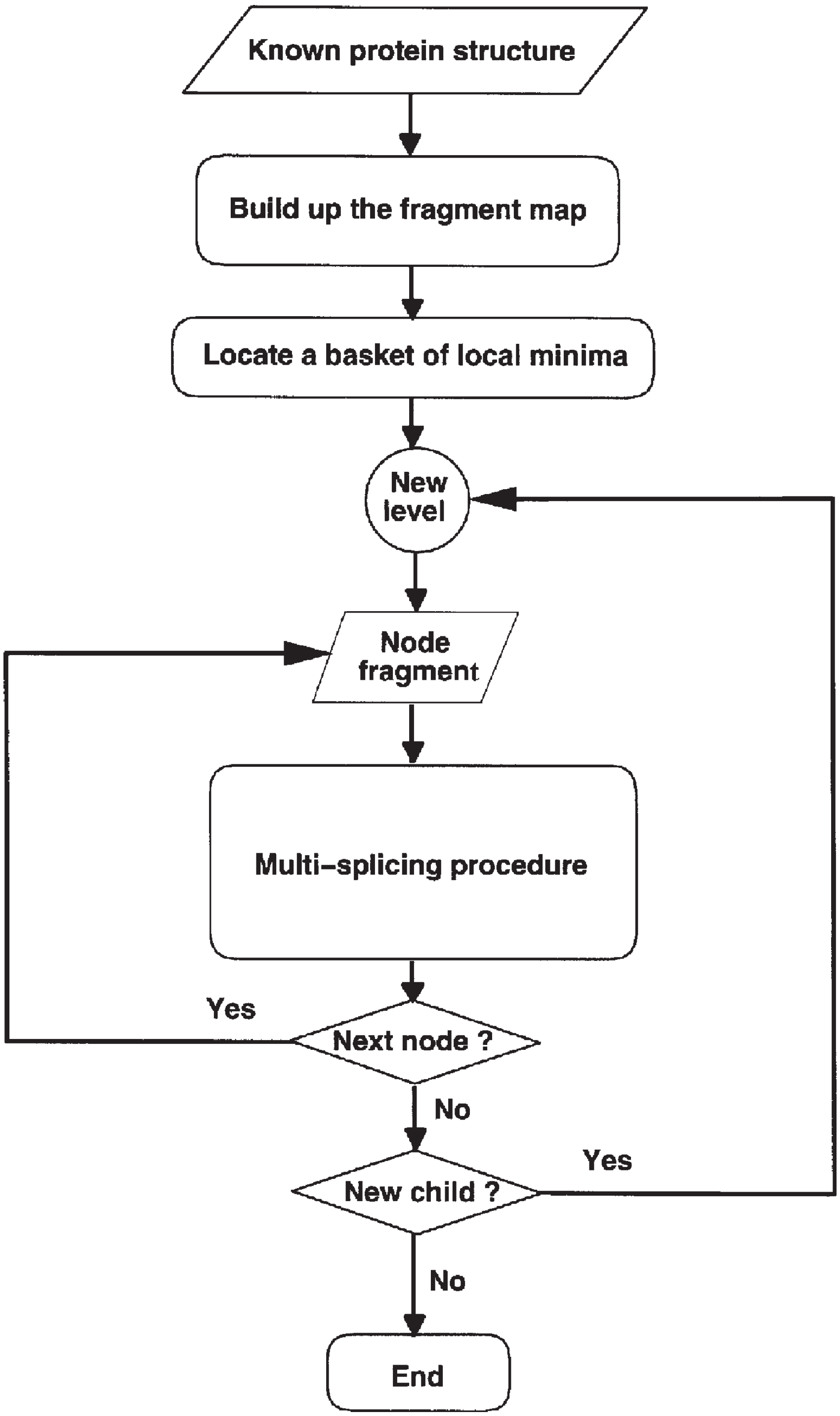

FIGURE 1. A flow chart of the algorthm: the procedure is based on the notion that protein folding is a hierarchical process.[11,12] Through a combinatorial process, *building blocks* assemble to form the compact, independently folding hydrophobic units.[73,74] Unlike the stable folding units that possess a strong hydrophobic core, the building blocks are contiguous fragments, with highly populated conformations. Building blocks may be composed of a single or of several interacting secondary structure elements. If the fragments constituting the building blocks were to be cut out of the protein structure, the most highly populated conformations of the resulting peptides would most likely resemble the conformations of the building blocks seen in native conformations. Nevertheless, not all building block conforma-

FIGURE 1. (continued)

tions are preserved in the final, native structures. The mutually stabilizing interactions between the associating building blocks during the combinatorial assembly may favor selection of alternate conformations that exist in lower concentration in the solution. Under such circumstances, the equilibrium will shift in favor of the depleted building block conformer. In the next stage, the hydrophobic folding units associate to form domains, and subsequently the entire proteins. The algorithm initiates with the native structure. We iteratively cut it, from the top down, allowing multiple dissections at each iterative level. This results in a hierarchy of contiguous fragments. Each node in the descending *anatomy tree* is a building block segment. The native structure is the root node of the tree. The locations of the building blocks correspond to the end nodes of the top-down tree. To be able to carry out the dissection, the most critical component is a fragment-size independent scoring function. This function measures the relative stability of all candidate building blocks. There are three ingredients in this empirical scoring function: measurements of the compactness of the candidate building block conformation, its degree of isolation, and its hydrophobicity. The multicut dissection is applied progressively to the most stable fragments. An attractive feature of the anatomy tree is that at the completion of the dissecting procedure, it yields the most likely folding micropathway. Furthermore, examination of the minima among the fragments straightforwardly yields the alternate routes.

subunits, and finally protein oligomers. Here we show how, when putting these together, a broad range of observations, experimental and theoretical, can be explained.

Next, we describe these elements, and follow up with the comprehensive model. We show how these seemingly disparate components unite to yield a coherent picture. In particular, we discuss a wide range of examples taken from the current and recent literature showing their consistency with building blocks formation and binding *via population selection*. We end with a short general discussion.

II. THE COMPONENTS THAT GO INTO THE MODEL

A. The First Component: the Building Blocks[73-77]

At the first stage building blocks are formed. Building blocks are contiguous sequence fragments with variable sizes. Their stability derives from local interactions. A building block is a highly populated conformation. A given fragment may have several alternate favorable conformations. Some building blocks are highly stable, whereas others may be only marginally so. Hence, while in most of the cases the conformations of the building blocks that we observe in the native state are those that the building blocks would manifest as peptide fragments in solution, this is not necessarily the case. A building block may twist and open, changing its conformation. Nevertheless, for the most part, the building block conformation that we see when it is present within the native protein fold is the one the building block would have early in the folding process. Hence, the local, intrabuilding block interactions are native, already in the initial stages of the folding.

B. The Second Component: the Landscape Around the Bottom of the Funnels[74,78,80]

The energy landscape of protein folding has been depicted frequently as a funnel, where, via multiple routes, the conformations race toward a sharp, relatively pointed bottom. However, in reality proteins are highly flexible molecules, existing in a range of conformations. Hence, the bottoms of the funnels are frequently rugged, manifesting this situation. The barriers between the conformations are low. This immediately argues against the longheld view of “induced fit”. Instead, within this broad range of conformers, the ones that actually bind to the ligand are those that are most favorable and produce the most stable associations. The equilibrium will then shift in favor of the binding conformations. This implies *conformational selection*, not *induced fit*. Clearly, this does not entirely rule out the possibility that a certain extent of induced fit, optimizing the intermolecular interactions through side-chain movements, or a small extent of backbone movement can still take place.

C. The Third Component: Binding and Folding Are Similar Processes[73,74]

Protein folding involves intra-molecular recognition. Such recognition implies that a native structure is the outcome of mutual stabilization between its more favorable structural units. Hence, protein folding, and protein-protein association, are similar processes, governed by similar principles. All considerations that apply to protein binding apply equally well to protein folding. The sole exception is the presence of chain connectivity in folding, and its absence in binding. The polypeptide connectivity limits the degrees of freedom in folding, when compared with binding.

D. The Fourth Component: Similar Landscapes for Binding and for Folding[74]

Folding and binding can be similarly described by a series of sequential fusions and modifications of micro-funnels. In binding we fuse the corresponding funnels of the two molecules. In folding, we also sequentially fuse the micro-funnels, first of the building blocks to obtain hydrophobic folding units, followed by fusion of the micro-funnels of hydrophobic folding units, and so on, down the hierarchical folding process. As the number of binding events increases, the funnels progressively get more complex. The bottom of the micro-funnels are inhabited by building block conformations and, as we go down the funnel, by their associations. The conformationally more stable building blocks have steeper funnel bottoms, whereas those thatare conformationally more flexible exist in a broader range of conformers, and consequently with more rugged micro-funnel floors.

E. The Fifth Component: The Landscape is Dynamic, Changing with the Conditions[75,79]

The energy landscape is not static. Rather, it is dynamic, changing with the external conditions, or with the binding state of the molecule. The external conditions can be, for example, pH, temperature, ionic strength, or pressure. Alternatively, it can be the binding of the molecules forming the growing complex. The most populated conformer at the bottom of the folding funnel may differ from the most populated conformer at the bottom of the complexed form of the molecule. Similarly, as we go down the cascade of an enzyme pathway, or an allosterically regulated protein, or multimolecular complex formation as in signal processing, the most populated conformer at the bottom of the first complex may differ from the most populated conformer in the higher-molecular assembly formation. What we may observe is a sequential enrichment of conformers, largely through the process of *selection*. The landscape is dynamic, implying shifts in populations.

III. THE MODEL: BUILDING BLOCKS FORMATION AND BINDING, *VIA POPULATION SELECTION*

The components described above lead to a logical scheme. Protein folding is a hierarchical process. The building blocks form. The most highly populated conformers are those observed in the native state. Even if the native building block conformation is marginally stable and would exist in a relatively low population still the native conformer is more highly populated than all other conformers. Furthermore, it would not constitute a local minima in our candidate building-blocks fragment map.[77] In the next hierarchical step, the building blocks associate via combinatorial assembly. This association is a binding event. The only difference between the binding of building blocks and the binding of larger stable units, such as domains, or subunits, or different molecules in a complex is the shorter population time of the building block conformer. In this binding event, among the range of conformations present at the bottom of the building block micro-funnel, the ones that bind are the most complementary. Hence, it is largely a process of *selection of building blocks conformations*. With the binding of the native conformers, the population would shift in their favor, further driving the folding reaction. Through their binding, they mutually stabilize each other, leading to the formation of the compact, stable, hydrophobic folding units. In the next step, again through selection, the most complementary hydrophobic folding units within the population at the bottoms of their respective micro-funnels bind to form the domains. In each of the binding events, the changing conditions lead to shifts in the populations. Hence, the population times of given conformers differ at the bottom of consecutive (micro-)funnels, whether in intramolecular folding or in intermolecular binding.

Hence, the critical issue is the population times of the conformationally fluctuating building blocks. If they are very high, and if these building blocks are sequentially connected, fast folding kinetics would be observed. Conversely, if the population times of the native building block conformers whose associations produces the native state are low, slow kinetics is likely to be seen. Traps can occur both in sequential and in nonsequentially folding protein. For example, Che Y is a three-state sequentially folding protein. However, traps may be expected to occur more frequently in cases of nonsequential folding, particularly in those cases where the native building block conformation is unstable, and hence has low population time. The traps largely contain the native conformations of the building blocks; however, their association is nonnative. Thus, traps serve a purpose, namely, to increase the concentration of the native building block conformations. Even though the association is nonnative, the end result is shifting the equilibrium in their direction.

This comprehensive model further implies that conceptually there is no real difference between two-state and threestate folding proteins. The distinction between two-state and three-state is in the population time of the intermediates. In apparent two-state proteins, the population time of the intermediate is very short. On the energy landscape it would be observed as residing in a shallow well. On the other hand, in three-state proteins, the well is deeper, implying higher population times. However, by changing the external conditions, such as by adding a reagent like trifluoroethanol, the intermediates are stabilized, that is, their wells get deeper, and their population times increase allowing us to observe them. The reagents do not change the pathways. Changing of the external conditions results in shifts in population times via changing the relative stabilities.[75,79]

Further, the model states that the difference between the Levinthal view and the "new view" can be reconciled.[71] While hypothetically there are multiple paths going down the funnel toward its bottom, in reality this is not the case. If we accept that protein folding is hierarchical, and that folding initiates from local elements, and that these local (building block) elements have higher population times than alternate conformations, then the theoretically huge number of potential pathways reduces to the combinations between building blocks conformers. Hence, in practice, simulations can be enhanced not by restricting the conformational space search. Rather, enhancement would derive from increasing the population times of given conformations. Not all combinations are equally likely. Moreover, the number of these decreases as we go toward the bottom of the funnel.

This model directly implies that sequential folding is faster and less prone to errors, because building blocks adjoing each other on the chain bind first. Furthermore, the model also implies that even in nonsequential folding, folding is likely to proceed through first trial binding of sequentially connected building blocks, because such a route is kinetically more favorable.

It has been suggested that proteins with similar topologies fold through similar pathways (e.g., Refs. 55, 82). This suggestion is consistent with the model proposed here. Proteins with similar folds are likely to have similar building blocks. Similar building blocks are likely to associate in similar ways, as they would most favorably similarly complement each other, producing favorable, stable associations, that is, similar, high population time building blocks would bind to similar partners via a similar *selection* process.

Hence, by considering that (1) folding is similar to binding, and therefore (2) that mechanisms observed in intermolecular binding, that is, selection of most favorable conformers, with population shifts to keep the equilibrium, are operative in folding, that (3) the building block elements that we observe have the highest population times, and that (4) the landscape is dynamic, we are able to provide a practical protein folding — and binding — model.

Below, we illustrate how this model explains a broad range of observations.

IV. THE MODEL EXPLAINS A BROAD RANGE OF OBSERVATIONS

A. Change in Conditions

Recent refolding experiments of acylphosphatase (AcP) in the presence of trifluoroethanol (TFE) have shown that when the denatured protein is refolded in increasing concentrations of TFE, the refolding rate initially increases. The refolding rate is maximal when the concentration of TFE reaches about 11%. Then it decreases. A similar observation has been made for hexafluoroisopropanol (HFIP), showing that this behavior is not specific to only TFE.[16] AcP is a 98-residue protein. Its structure consists of two parallel α-helices, stacked against a five-stranded, antiparallel β-sheet. Previously, NMR and optical spectroscopy studies have shown that this protein folds in a highly cooperative manner. No significant amount of intermediates have been observed, leading to its classification as a two-state protein. On the other hand, the study by Dobson and his colleagues has illustrated that after increasing the content of these alcohols, the rate of folding initially increases, and that partially structured species accumulate as intermediates.[13] Dobson and his colleagues point out that the increase in TFE or HFIP promote α-helical formation, because alcohols stabilize the hydrogen bonds. The rate of helix formation is higher than that of β-sheet. Thus, the rate of folding initially increases with the increase in the concentration of TFE, as first the native helices would form. Next the nonnative helices are formed, corresponding to the sequences destined to adopt a β-sheet. These regions change their conformation subsequently to adopt their native β-state. Hence, in our terminology the AcP is a three-state protein. Under physiological conditions, no intermediate states are identified, because this state is short lived, with low population time. By adding the alcohol, the intermediate state with both the native α-helices, and with largely nonnative α-helices elsewhere is formed. This conformation is stabilized, and hence detected. Regardless of the presence or absence of the TFE, the pathways are the same. The helices would form before the sheet whether the reagent is present or absent. However, the effect of the alcohol is to increase the population time of all helices, including the nonnative ones, enabling detection of the intermediate. Furthermore, this example illustrates the effect of the change in the conditions on the landscape. TFE changes the environment, and hence what we observe is a shift in the populations.

More recent results from the same group[14] on a larger set of proteins are consistent with this interpretation. The initial acceleration observed after increasing the TFE concentration is higher for the apparent two-state when compared with the three-state folding proteins. The authors have observed that for the two-state proteins, the extent of acceleration of folding correlates with the number of local hydrogen bonds in the native structure. For the multistate proteins, the extent of acceleration is smaller. Thus, by stabilizing the local hydrogen bonds, the folding rates increase. The effect is smaller for proteins whose wells are already populated with intermediates. These wells largely consist of native conformations of the building blocks, which are misassociated. Hence, here a stabilization of the local interactions will affect the folding rates to a significantly lesser extent.

Park et al.[18] have observed a similar phenomenon. They have studied the folding kinetics of IgG binding domain of streptococcal protein G, a small, 57-residue molecule. Although this domain has been cited previously as an example of a protein obeying the two-state folding mechanism, by varying the solvent conditions, and using sensitive stopped-flow fluorescence methods, the authors have observed an early folding intermediate. This intermediate was detected under stabilizing, low denaturant, and the presence of sodium sulfate conditions. Interestingly, a 66-residue fragment with an N-terminal extension containing five apolar side-chains exhibits three-state kinetic behavior, practically identical to that observed for the 57-residue domain. An additional example is provided by the *E. coli* αsubunit of tryptophan synthase.[17] At around 3.2 *M* urea, this α/β barrel protein displays a stable intermediate. However, when monitored at the 5 to 9 *M* urea range, an additional cooperative process is observed. Hence, all the above indicate that whether two-state or three-state (or multistate) the pathways are the same, and intermediates exist. However, depending on the conditions, their populations differ. We are able to observe these only when their populations are large enough, trapped in their wells.

Kiefhaber[58] has also investigated folding, by using interrupted refolding experiments. This method enables monitoring the amount of native molecules during the folding reaction. This is carried out by utilizing the high energy barrier between the native state and folding intermediates. By first completely diluting unfolded molecules to initiate refolding, and subsequently interrupting the folding by altering the conditions of the solution, one may observe the difference in the unfolding rates of native lysozymes (on the order of seconds) vs. partially folded intermediates (miliseconds). By following the unfolding amplitudes after various times of refolding, the fraction of the native molecules can be obtained. Applying it to hen egg white lysozyme, the results show that under strongly native conditions lysozymes can refold in parallel routes. In particular, Kiefhaber has observed that 86% fold in a three-state kinetics. Previously, this molecule has been classified as folding via a two-state kinetics, suggesting that the well on the energy landscape in which the intermediates reside is not too deep. Plate 1A* illustrates the three levels of building blocks cutting of this hen enzyme. At the first level, the enzyme is still one unit, illustrating its compactness. At the second level of cutting, three building blocks are observed, forming two hydrophobic folding units. The green forms an independent unit and is a more stable, single polypeptide segment. This folding unit can be classified as an insertion domain. The second hydrophobic folding unit is made of two segments, the red and yellow building blocks, comprising the amino- and carboxy-termini, respectively. The C-terminus building block is the least stable. In both two-state and three-state scenarios, as the green building block is the most stable it is likely to form first. On the other hand, the red and yellow building blocks are unstable. While each has some population time, their association leads to higher stability. The slow route might conceivably be caused by the nonnative association of the yellow and green building blocks, in the absence of the mediating

* Plate 1A follows page 426.

red building block leading to the observed three-state. The situation is different for the T4 lysozyme (Plate 2B*), as we discuss below. This leads us directly to the discussion of kinetic traps and two-state vs. three-state folding, with respect to the building blocks model. A two-state is actually a three-state folding event, where the intermediates are too transient to be detected. However, if the conditions are changed, or the experimental equipment is sensitive enough, the intermediate state will be detected.

B. Intermediates, Two-State, Three-State, and Kinetic Traps

Qualitatively, the building blocks folding model accounts for three-state vs. two-state protein folding. It is also in agreement with the faster *vs.* slower folding rates of the two-state proteins. The size of the building blocks, the way they associate in the native structure, the number of ways they can assemble, as well as their population times explain a broad range of folding rates. Furthermore, it explains the folding rates of proteins that fold with two-state when compared with data for proteins that fold via three-state kinetics, with observed, stable intermediates. Above, we touched on cases that are apparent two-state, but have been shown to be three-state, consistent with the model. Below, we describe two-state cases, those of villin and T4 lysozyme, which owing to the existence of two hydrophobic cores, might have been expected to be three-state, yet have been shown experimentally to be two-state.

Two-state systems have been considered to constitute the simplest models of protein folding. In two-state protein folding, only the unfolded and the native states are the populated forms of the protein. Numerous studies have been devoted to two-state systems (reviewed in Refs. 55, 83, and references therein). Unlike the situation in two-state, proteins considered to fold via a three-state folding pathway, illustrate populated, stable intermediate state(s).

Among the important determinants of folding rates, two factors have been shown to stand out: the relative contact order (Plaxco et al.[56] and the secondary structure type. The first essentially measures the weight of sequential, local, *vs.* nonlocal interactions in any given protein molecule. Hence, it is a function of the topology. Experimentally derived rates of protein folding for small single domain proteins have been shown to nicely correlate with their relative contact order.[55,56] However, the relative contact order suffers from four limitations: first, it does not account for the effect of mutations on the folding rate; second, it does not account for the different folding kinetics seen under different conditions; third, it applies only to small, two-state folding proteins; fourth, a residue-based model cannot put the folding process in the context of the energy landscape, and the folding funnels. In principle, a residue-based approach cannot capture the downhill folding pathway, as it basically relates to a random search process. No such limitations exist in the building block folding model. The concept of the combinatorial assembly of the conformationally fluctuating set of building blocks, with their most highly populated conformations being those observed at the native state, accounts for these remarkably well. Mutations and different solvent conditions will result in changing the energy landscape, the outcome of dynamic shifts in the populations of the building blocks.[76,79] The experimental observation of an intermediate state in a previously cited example of a protein obeying a two-state kinetics after changing the solvent conditions[18] is consistent with the building block folding model.

Villin (2VIK) is a fast folding two-state protein, despite its possessing two hydrophobic cores.[61] Plates 2A,B* illustrate the second and third levels of building blocks cutting of this protein. In the second level of cutting (Plate 2A), the red and green building blocks constitute a single building block fragment. The two observed hydrophobic cores[61] are within the building blocks. In Plate 2B, one hydrophobic core is within the green, and the second within the yellow. The red building block enhances the green on one side of the hydrophobic core (which is the reason for the red and green forming a single building block in the second level of cutting), and the blue enhances both, through a continuation of the green into the yellow on the upper side of the figure. Villin is a practically sequentially folding protein, with long helices and strands also contributing to its high folding rate.

Llinas and Marqusee[84] have investigated the role of subdomain interactions in the folding and stability of T4 lysozyme. They found that thermodynamically this enzyme behaves as a cooperative unit, with the unfolding transition fitting into a two-state model. Yet, visual (and computational) inspection of the T4 lysozyme illustrates two subdomains: an N-terminal subdomain (residues 13 to 75) and a C-terminal subdomain (residues 76 to 164 and 1 to 12). Permutation of the termini of the enzyme has shown that their location has an important effect on protein stability, although not on its fold. In particular, the experiments have shown that when the two subdomain fragments were isolated, the C-terminal subdomain folds into a marginally stable structure, whereas the *N*-terminal subdomain is predominantly unfolded. Plate 1B illustrates the building blocks cutting at the second level. The more compact C-terminal subdomain folds on itself, dragging the N-terminal subdomain along with it. Hence, while visually and computationally the *N*-terminus forms a stable subdomain, because the C-terminus subdomain is more stable, no intermediate state is detected. There is an intermediate state, because there are two hydrophobic cores, reflected in two hydrophobic folding units. However, as the barrier is low, the intermediate state is not observed.

Both villin and T4 lysozyme examples demonstrate the consistency of the building block folding model with experimental results illustrating that proteins with two hydrophobic cores still fold via a two-state kinetic. An additional example is the *E. coli* dihydrofolate reductase. There, too, two stable hydrophobic folding units are observed, both by eye and computationally. Yet, experimentally it is a two-state protein.[20] The building blocks cuttings, and assembly, demonstrate a similar picture.[85]

C. Mutations

Rothwarf and Scheraga[36] have studied the folding kinetics of the wild type and two mutants of hen egg white lysozyme. The mutants involved Trp-62 and Trp-108 individually replaced by tyrosines. They have observed that both mutants fold significantly faster than the

[*] Plate 2A and 2B follow page 426.

wild-type structure (13- and 7-fold, respectively). The authors have suggested that these results point out to the rate-limiting step in the folding of lysozyme arising not from an inherent slowness in the formation of the native structure. Rather, the difference in rates is a consequence of the formation of a highly stable nonnative intermediate. Inspection of Plate 1A reveals that Trp 62 is on the surface of the yellow building block in the β-domain in the third level of the cutting. On the other hand, Trp-108 is buried in the α-domain. The hydrophobic core of the α-domain is stronger than that of the β-domain. Thus, the β-domain is less stable, and its formation is likely to depend on the formation of the α-domain.

While both mutations increase the folding rates significantly, the enhancement is more dramatic for the Trp-62 mutant. In the third level of cutting, the yellow building block is still stable enough to have a higher population time than all alternate conformations. In the wild type, the yellow building block may then associate with other building blocks to produce non-native conformations. In substituting the surface Trp, the chance for nonnative association of the yellow building block is significantly reduced. A similar argument applies to the Trp-108 mutation. There, the mutated residue is on the surface of the blue building block in the third level of cutting. Thus, both mutations lower the likelihood of the nonnative associations of the building blocks, which conceivably create the trap in the wild-type enzyme.

Shao and Matthews[38] have studied single Trp mutants in a monomeric form of the tryptophan repressor. Previously, they have shown that the dimeric wild-type tryptophan repressor (WT TR) forms a stable highly populated nonnative monomeric intermediate. This monomeric intermediate was proposed to collapse into a compact, nonnative conformation, with the hydrophobic core formed within the monomer, rather than at the interface of the two monomers that are associated into a functional, intertwined dimer conformation. By creating a L39E TR mutation, they have further stabilized the nonnative monomeric form and proceeded to study the roles of two Trp residues, Trp-19 and Trp-99, replacing each with Phe. Their spectroscopic analysis has illustrated that the monomeric repressor can adopt a well-folded conformation with the Trp side-chains distinctly different when compared with the situation in the dimer. The authors take this result to suggest that the hierarchical protein folding model does not apply to the formation of the functional dimer.

Plate 3* illustrates the building blocks cutting of the wild-type trp repressor monomer. As the figure shows, the monomer is cut into two building blocks. In the functional dimer, the α-helical building blocks of the two monomers are swapped. Thus, we propose that the nonnative form of the monomer involves misassociated building blocks, where the two building blocks from the same monomer bind to each other. This situation is reminiscent of domain swapping.[86]

Shortle[37] has described an unusual mutant of staphylococcal nuclease (Gly88Val) that reduces the stability of the native state by stabilizing the denatured state. A high-resolution analysis of this mutant in a shorter fragment has revealed a significant alteration in the packing of the hydrophobic core of the β-barrel, and a concomitant large in-

* Plate 3 appears following page 426.

crease in the mobility of several loops that connect the β-strands. These illustrate that this substitution has altered the conditions, and hence the landscape, with the equilibrium shifting toward the altered conformer.

D. Similar Topologies

Recently, a number of experimental and theoretical results have illustrated that protein folding rates and mechanisms are largely determined by the native protein topology (e.g., Refs. 55–57; 87-89).

Consistently, the approach illustrated here is based on the notion that the native state is a critical determinant of the folding mechanism, and of its rate. Rather than imagining that in the search for an optimal fold, all potential contacts are tried, leading to a vast number of intermediates formed in parallel in the transition state, here we suggest that the native contacts predominate as the chain undergoes its folding process. Furthermore, these native contacts are largely those existing relatively near each other on the linear chain. Hence, such native contacts are kinetically favored. Even if eventually longer range contacts prevail, the chain would still go through the trial formation of nearby contacts first, because they are kinetically more favorable than the ones that are further away. Furthermore, the native contacts that are made early in the folding process are those involving intrabuilding blocks folding. Most building blocks have high population times, as evident from their observation in solution when in a peptide-fragment form. This suggests that similar folds will manifest similar folding mechanisms, regardless of the variability in their sequences, and in their stabilities.[55]

Plate 4* illustrates an example of the application of the building blocks folding model to two topologically similar proteins, with different stabilities. In this plate, we present side by side the glutamate dehydrogenases from a mesophile (1 hrd) and from a thermophile (1 gtm), sharing similar native state topology. As these are complex proteins, for simplicity only the first two levels of cutting are presented in the plate. Despite the 60°C difference in their melting temperatures, the similar folds reveal similar anatomical features. While the two enzymes differ in sequence and stability, their folding pathways are similar. Plate 4A presents the anatomy trees for the mesophilic (from *Clostridium symbiosum*) glutamate dehydrogenases, and Plate 4B depicts the situation for the thermophilic (from *Pyrococcus furiosus*) glutamate dehydrogenases. The *Pyrococcus furiosus* glutamate dehydrogenase is extremely thermostable, with a half-life of 12 h at 100°C. Its melting temperature (T_m) has been reported to be 113°C. The mesophilic *Clostridium symbiosum* glutamate dehydrogenase has a half-life of only 20 min at 52°C, and its melting temperature (T_m) is 55°C. It shares only 34% sequence identity with its thermophilic homologue. Despite the large difference in melting temperatures, inspection of the anatomy trees shows a remarkable similarity in their anatomies, that is, in their folding pathways.

Nevertheless, as noted above, while the folding pathways in topologically similar proteins are similar, the folding rates are affected by additional factors

* Plate 4 appears following page 426.

as well, such as the secondary structure type, with α-helices forming faster than β-strands; external conditions that change the environment alter the folding rates.[13,16] Similarly, rates may be changed by mutations, without a change in the native topologies.[16,83]

E. Fragment Complementation

Over the last few years, fragment complementation experiments have become increasingly popular in investigations of protein folding.[21,22,91,92] The protein is cut to produce fragments of various sizes, and these are examined to see if they are capable of folding, and what are their higher population time native/non-native conformations. While peptides which represent individual elements of secondary structures may not fold to any large extent, larger fragments, or domains, may mutually stabilize each other. Thus, for example, Oas and Kim[93] have linked pairs of peptides that individually were unstructured in the bovine pancreatic trypsin inhibitor. The resulting structure was native-like. Studies carried out on fragments from the tryptophan repressor and the cytochrome c have illustrated that some of these folded spontaneously into autonomous folding domains.[94,95] More recent examples include the numerous studies carried out by Matthews and his colleagues. Among the cases they have investigated are the dihydrofolate reductase and the α subunit of tryptophan synthase.[17,20]

The fact that for most cases cutting a protein and mixing its fragments results in the same conformation is consistent with binding and folding being similar processes. Similarly, the fact that ligating separate subunits usually results in a structure similar to that where the subunits are separate again illustrates the same principle. The building block folding model is also consistent with the experiments cutting and spectroscopic measurements.

F. Fragments and Pathways: The *E. coli* Dihydrofolate Reductase Example

During the last few years, Matthews and his colleagues have been carrying out a series of experiments,[19,20,96] with the goal of identifying folding pathways and intermediate states.[91,92] Their approach has been to produce a set of fragments of various sizes, and to determine whether these independent fragments are able to fold. Building on previous successful applications of this approach (reviewed in Fontana et al., Ref. 92), Matthews and his colleagues have applied it to the *Escherichia coli* dihydrofolate reductase.[20] Specifically, they have cut the dihydrofolate reductase enzyme into eight overlapping fragments. CD and fluorescence spectroscopy results have demonstrated that six out of these, including the adenine binding domain (ABD, residues 37 to 86), were largely disordered. A stoichiometric mixture of fragments 1–36 and 87–159 also did not show evidence of folding beyond that observed for the isolated fragments. Fragment 1–107 showed some apparent secondary and tertiary structure; however, it spontaneously self-associates. On the other hand, fragments 37–159 have shown considerable secondary and tertiary structure, with a well-defined two-state unfolding behavior. This result was surprising, because inspection of the DHFR structure illustrates that first the adenine binding domain appears to constitute an independent domain by visual and functional criteria, and second fragments 1–37 are inserted within the structure of 37–159. Additionally, Matthews et al.[19,20,96] have observed that this fragment may fold cooperatively also in the absence of ammonium sulfate, raising the possibility that partially or fully folded forms of this fragment act as a kinetic trap for the folding of the full-length protein.

Computationally, we draw on our recently developed and implemented algorithm, which enables cutting the protein structure into progressively smaller units, revealing its anatomy in terms of folding pathways.[77] We apply this tool to both the *E. coli* and the human DHFRs to obtain their folding pathways. For the *E. coli* enzyme, our results are consistent with those of Gegg et al.[20] While the pathway they have observed is seen in our hierarchical dissection path, an additional alternate route exists. This route is consistent with the ABD being an autonomous unit. This second pathway is in agreement with the proposition of two populations of folding intermediates, as suggested by NMR.[96] Moreover, taken together the results obtained for the *E. coli* and the human DHFRs (discussed below) are in agreement with the *in vivo* chaperonin-assisted vs. the *in vitro* observations.[97,98] They further enable making some propositions as to the nature of the intermediates.

The *E. coli* dihydrofolate reductase is a 159-residue protein. It catalyzes the reduction of 7,8-dihydrofolate to 5,6,7,8-tetrahydrofolate.[99,100] The enzyme uses NADPH as the reducing factor. The tertiary structure is a doubly wound, parallel α/β-sheet, with 4 α-helices and 8 β-strands. The crystal structure (Plate 5A, PDB code: 7dfr, 101) shows that residues 38 to 88 (green in Plate 5A) form the adenine binding domain. These residues are flanked by the amino terminus fragment (1–37, red in Plate 5A) and the carboxy terminal segment (89–159, yellow in Figure 1). Neither of these three fragments, nor a mixture of the first and last fragments, have been observed to possess secondary and/or tertiary structure to an appreciable extent in the CD and fluorescence studies. Further, fragments containing the ABD attached to the amino terminus (1–86) did not show significant folding either. Gegg et al.[20] have also probed fragments 37–107 and 1–107. Although the signals vary, in all cases the fragments have illustrated largely, although not entirely, disordered structures. The only fragment that demonstrated the presence of substantial secondary and tertiary structure is 37–159. While fragments 1–36, 37–107 and 108–159 displayed little or no dependence of the signals on the urea concentration, fragment 37–159 has shown a cooperative unfolding transition. Gegg et al.[20] conclude that the chemical cleavage of DHFR into fragments that visually reflect its domain structure do not yield the result, which might have been expected. Comparison of these results with the visual picture illustrates why they were unexpected: the functional, (Rossman fold) adenine binding domain appears compact, and similarly a stoichiometric mixture of the amino and carboxy fragments.

Applying our recursive dissection algorithm to the DHFR, all fragment candidates are generated, and a stability score is assigned to each. We proceed to locate local minima on the fragment map of the protein. A local minimum is the highest value in a defined local region.

[*] Plate 5A follows page 426.

All highest scoring candidates within their respective local region are registered. The next step is a recursive, top-down cutting process. Starting with the native structure that constitutes the root-node of the protein and the set of registered building block candidates, we scan the list for a set of fragments whose combination yields the entire node fragment. We allow a small overlap between them. If the fragment is short it is left unassigned. If it is long but has a low score, it is considered a linker fragment, or a fragment whose conformation has changed during the assembly process. There is no limit on the number of branches that can sprout from a node. The recursive node-splitting procedure continues until no two nodes whose scoring sum exceeds a threshold value can be located. The entire tree growth stops when no new children nodes can be generated. Further details are given in the legend of Figure 1. The detailed algorithm will be given elsewhere.[77] A listing of the building blocks that are local minima for the *E. coli* DHFR is given in Table 1. Furthermore, at each sprouting level of the tree, we identify the hydrophobic folding units through a combinatorial assembly of the collection of building blocks.

Plate 5B* represents the most likely folding pathway depicted on the fragment map of DHFR. The blue horizontal lines are the local minima, and the building blocks participated in the most likely pathway are drawn in red. Inspection of the figure illustrates that the building blocks assemble in multiple routes to finally yield the native fold at the top. The anatomy tree, with its detailed step-by-step micro-paths for the folding of the enzyme is shown in Plate 5C. The branches sprouting from a node form the respective node. Each branch is a building block. It is labeled with its position and score. Additionally, for level one and two the assignments into HFUs are noted. These are also given at the top right-hand side of the figure.

Plate 5D depicts the step by step dissection (top row) and assembly into HFUs (bottom row) graphically. By going through the last two plates, we can reconstruct the most likely folding pathways of the DHFR, based on our algorithm and scoring function. By inspecting the fragment map of Plate 6B and the listing of the building blocks and their scores in Table 1, we can identify the alternate routes.

Several findings are clearly observed:

1. Starting with the native conformation, no dissection was observed at the first level (labeled L1 in Plate 5D). This indicates the association between building blocks is critical for the overall folding stability. At the next level the molecule is dissected into three fragments (labeled L2). At the further level, the green building block (35–103) is now split into three building blocks (labeled L3), green (31–63), yellow (59–84) and blue (85–103). The red (5–35) and yellow (104–159) building blocks of the second level (labeled L2) are not dissected further at the next stage. The yellow building block of the second level is now colored cyan in level three (labeled L3). Plates 5B,C show that the native fold is obtained through multiple routes.
2. Plate 5B illustrates how the building blocks assemble via a stepwise hierarchical procedure. However, it does not reveal whether DHFR is a

* Plate 5B follows page 426.

TABLE 1
The Local Minima Found in the Fragment Map (Plate 1) of Dihydrofolate Reductase

#	Range		Size	Z	I	H	score
1	1	159	159	1.604	0.000	0.792	4.091
2	35	103	69	1.466	0.118	0.739	4.020
3	31	103	73	1.507	0.105	0.740	3.875
4	35	94	60	1.456	0.138	0.719	3.353
5	31	94	64	1.497	0.122	0.721	3.294
6	37	85	49	1.462	0.153	0.701	2.853
7	1	128	128	1.576	0.086	0.766	2.851
8	31	85	55	1.495	0.150	0.698	2.406
9	18	103	86	1.637	0.164	0.707	0.583
10	2	103	102	1.660	0.161	0.724	0.485
11	59	84	26	1.501	0.190	0.628	0.466
12	14	103	90	1.654	0.160	0.706	0.347
13	30	159	130	1.805	0.099	0.742	0.052
14	18	94	77	1.628	0.182	0.687	0.045
15	132	159	28	1.607	0.204	0.628	-0.686
16	31	126	96	1.772	0.175	0.706	-0.807
17	18	85	68	1.625	0.208	0.664	-0.815
18	95	159	65	1.716	0.199	0.673	-0.901
19	128	159	32	1.618	0.193	0.620	-0.925
20	109	159	51	1.681	0.206	0.649	-1.097
21	31	63	33	1.492	0.301	0.626	-1.131
22	104	159	56	1.695	0.205	0.654	-1.156
23	58	103	46	1.646	0.228	0.642	-1.251
24	5	85	81	1.722	0.223	0.666	-2.134
25	77	159	83	1.828	0.203	0.670	-2.356
26	58	159	102	1.926	0.174	0.684	-2.810
27	53	159	107	1.946	0.182	0.679	-3.521
28	75	103	29	1.628	0.318	0.583	-3.628
29	18	63	46	1.602	0.333	0.601	-3.629
30	5	63	59	1.690	0.328	0.617	-4.175
31	59	133	75	1.916	0.248	0.635	-4.486
32	58	128	71	1.893	0.259	0.630	-4.581
33	5	35	31	1.648	0.334	0.564	-4.598
34	53	133	81	1.931	0.249	0.632	-4.994

sequential or a nonsequentially folding protein. In general, while a consistent pattern of binary branches at each node in the tree will always yield a sequentially folding protein, a tree where some nodes sprout into three or more branches might or might not be a sequential folder. On the other hand, this information is straightforwardly given by Plate 5C. A comparison of the building block cuttings with the hydrophobic folding units assignment illustrates that DHFR has a complex fold. The figure illustrates that the most likely pathway involves parallel assembly of the three central building blocks, and assembly of the first and last building blocks. This is also observed by a comparison of the top and bottom rows in Plate 5D.

3. As the bottom row of Plate 5D shows, the two hydrophobic folding units roughly correspond to the visual inspection. At the second level (labeled H2 in Plate 5D) the first HFU includes the fragment 35–103, whereas the second HFU corresponds to fragments 5–35 and 104–159. The adenine binding domain spans residues 37–87. On the other hand, Plate 5C illustrates that the second HFU consists of three building blocks: 31–63, 59–84, and 85–103. A combination of the first two building blocks corresponds

nicely to the ABD. However, the addition of the third 19-residue-long building block improves the HFU score.

4. An examination of Table 1 and of Plate 5B illustrates additional local minima. Among these, there is a building block spanning residues 30–159. This building block may associate with the 5–35 building block to produce the native protein. Fragment 87–159 could arise via an alternate micro-path, assembling the red building blocks 85–103 and 104–159 in Plate 5B. Together with building blocks 5-35, 31–63 and 59–84, the native fold could be formed. Fragment 2–103 in Plate 5B also constitutes a local minima close to 1–107. Through an alternate route, via assembly with 104–159, it too could lead to the native fold. These routes involve the native conformation of each of the blocks. We shall come back to this point below.

A comparison with the results of Gegg et al. yields a good correspondence. First, all fragments observed by CD and fluorescence to show some secondary or tertiary structure are in the basket of our building blocks. By way of alternate pathways, they could lead to a complete protein fold.

Second, fragment 37–159 was observed by Gegg et al. to possess significant secondary and tertiary structure. In our case, this fragment would arise in the assembly of fragments 35–103 and 104–159. However, our most favorable pathway suggests that 35–103 would form an independent HFU, whereas 104–159 would most favorably associate with 5–35 to yield a second HFU (labeled B in Plate 5C). Hence, at first sight there appears to be a discrepancy between the CD and fluorescence results when compared with those obtained computationally. This, however, is not the case. There are several points to consider: (1) the results of Gegg et al. illustrate that the transition state of this fragment is broad, suggesting that 37–159 exists in a nonnative conformation, which is not very stable. This apparently facilitates the insertion of the red (1–37) building block. Furthermore, (2) visual inspection of DHFR immediately indicates that in the absence of the 1–35 fragment (red in Plate 5A), the association between the ABD (green) and the carboxy-fragment (yellow), the 37–159 fragment cannot exist in the native structure. The red fragment mediates the interactions between the green and yellow fragments. The β-strand in the amino terminus (unassigned fragment in gray) interacts with the green β-strand to continue the sheet. The continuation of the amino terminus strand into the assigned red building block interacts with the yellow β-strand. (3) Our algorithm addresses only native conformations, whether of the entire fold, or of any of its fragments. We do not see the 37–157 building block, since if kept at its native conformation, it would be unstable. On the other hand, 35–103 is a very stable HFU, and 104–159 is also a relatively stable building block. Taken together, it suggests that in the 37–159 fragment the conformations of its composite building blocks are the native ones. However, the tertiary associations between them are nonnative. Hence, this fragment is an intermediate conformer on the folding pathway to the native state. It will be trapped at a minima well, which is not too deep, as its stability is not high. This trap in on-pathway, rather than off-pathway, because the trap actually contributes to achieving the native fold through its acting as a trap for the native conformers of the building block fragments. By transiently holding these, it shifts the population of the building blocks conformers in the direction of the native ones. This is further consistent (4) with the NMR results, suggesting the existence of two populations of intermediates.

This rationale is consistent with the observation that the other fragments observed in the spectroscopic experiments of Gegg et al. are largely, although not completely, disordered, possessing some secondary structure. This is in agreement with the results of Goldberg et al.[97] for the human DHFR. All of these are building blocks present in our fragment map (Plate 5B, and Table 1), and constituting alternate, less heavily traveled pathways. Hence, their associations can also be trapped in some minima wells, again contributing to achieving the native fold via shifting the conformer-equilibrium in their favor. Like previously, subsequently these nonnative associations between the native building block conformers will climb up from their wells to reassemble through a combinatorial assembly process to yield the stable hydrophobic folding units. Hence, here we argue that the traps down the slopes of the landscape contain nonnative associations of native building block conformations. As such, they actually aid in the folding process. Traps are largely on- rather than off-pathway.

It is also noteworthy that while the conformation of the 37–159 fragment is relatively unstable, its barriers might be high, as indicated by its existence both in the presence, but also in the absence of the ammonium sulfate.

The agreement between the experimental and the computational results is gratifying, validating the building blocks model, and illustrating its usefulness. Furthermore, the dissection into an *anatomy tree*, and the combinatorial assembly yielding hydrophobic folding units, provides information about the complexity of the folding pathway(s). We note, however, that the algorithm can consider only native conformations. Furthermore, while the scoring function it utilizes appears adequate throughout our extensive testing, nevertheless, there is no assurance that this is the case. Detailed analysis of the *E. coli* dihydrofolate reductase fragments has been presented elswhere.[85]

G. Chaperonins and Chaperones: Inter- and Intramolecular-Assisted Folding

Chaperones assist proteins to fold. There are two types of chaperones: The first, inter-molecular case I, such as in the case of GroEL and the heat shock proteins (Hsp), the chaperones lead to *macroscopic* changes in the energy landscape, by lowering the barrier heights for the unfolding. On the other hand, in the second case as in the proregion[48] and in the uncleaved intramolecular chaperones (reviewed in Ma et al. Ref. 81), the change is both *macroscopic and microscopic*, because they not only lower the barriers to open nonnative interactions, but additionally directly assist in the folding by providing a template. In intramolecular chaperones, the chaperone is a fragment of the molecule. It may be cleaved and digested after fulfilling its chaperoning (and inhibitory catalysis) role, as in subtilisin and in α-lytic protease. Alternatively, it may also remain part of the structure if its presence is essential for protein function (as in

E. coli DHFR (Ma et al. Ref. 81 and in adenylate kinase [102]). Regardless of whether it is cleaved or remains intact, intramolecular chaperones are building block fragments, essential for attaining correctly folded molecules. In both inter- and intramolecular chaperone cases, the folding energy landscapes are similar, as macroscopically both lower the barriers. However, the *microscopics* differs. Given their different mechanisms, in the second case the more stable the intramolecular fragment that serves as a chaperone (whether a proregion or an uncleaved building block sequence-fragment), and the higher the population time, the faster the folding rate.

In type I, the association of the chaperones with the target proteins is weak, characterized by unstable, transient binding. Consistent with their general role, their binding is nonselective, carried out via a diffusion-collision mechanism. This nonselective binding is in contrast to *intramolecular* chaperones, which are highly specific. In the type II case, the association of intramolecular chaperones (particularly the uncleaved) with their parent protein molecule may be very tight, buried in its core, and serving as a *critical building block* fragment.[81,102] Hence, here too the same building block folding and binding principle applies: the mechanism of lowering the barriers is via building block *conformational selection*. The conformationally fluctuating building block intramolecular chaperone template selects the most favorable conformation of another building block, binds to it, and thereby stabilizes it. This "complex" further selects the most favorable conformation of additional building block(s). By stabilizing the building block(s) next to it, the intramolecular chaperone (or critical building block) lowers the barrier. Through such binding, the building block minimizes the chance of nonnative association.

Chaperonins like the GroEL have two roles: first to unfold misfolded molecules, and second to prevent aggregation. *If* the protein can fold spontaneously, like the *E. coli* DHFR (and at low concentration the eukaryotic DHFR), the chaperonin is likely to help strictly via the second mechanism (preventing aggregation). *If,* however, it does not fold spontaneously in *in vitro*, then chances are the chaperonin fulfills both roles. The misfolded conformation can be of mis-associated building blocks, in which case the trap is not very deep, with the chaperonin assisting through the second mechanism. However, if the misfolded conformation consists of misfolded building blocks, then the chaperonin may function through both mechanisms.

Hence, there is a difference between a chaperonin and a proregion. For α-lytic protease whose proregion has been removed from the nascent chain prior to folding, a chaperonin will be of little use. Thus, had the α-lytic protease been synthesized without the proregion, and a chaperonin were to be present instead, the native state would not be reached.

H. Back to the Dihydrofolate Reductase Example: *In Vivo* vs. *In Vitro* Folding

Plate 6A* presents the fragment map of the eukaryotic human DHFR enzyme; Plate 6B presents the anatomy tree; and Plate 6C presents the cutting levels (in the top row) and the combinatorially assembled hydrophobic folding units (in

* Plate 6A follows page 426.

the bottom row). The results illustrate that the folding pathways of the *E. coli* and of the eukaryotic enzyme are practically identical. The sole difference is that in the cutting, a helix that belongs to the green building block in the *E. coli* enzyme (Plates 6A,B), is now assigned into the yellow building block in the human enzyme (Plate 6C). This illustrates a slight shift in the position of the helix between the two structures. The similarity in the folding pathways observed in the anatomy trees is consistent with the experimental data: that data show that the folding of both the *E. coli* and the murine DHFRs are better modeled with three intermediate states,[103] when compared with modeling the *E. coli* with three intermediate states, and the murine with two.[104] The anatomy tree further provides some clues to these intermediate states. The first intermediate state may involve the more stable green building block of the ABD; the second intermediate state may relate to the red and yellow building blocks, and the third to the association of both hydrophobic folding units (the green building block with the yellow + red).

Neither of the enzymes is very stable: Clark and Frieden[103] report that native *E. coli* and murine DHFRs contain late-folding intermediates in addition to the native states. The binding to the ligands stabilizes the native conformations. As the native conformers bind, the population is shifted in their favor, further driving the folding-binding reaction.[75,79] Nevertheless, the *E. coli* DHFR is more stable (with a $T_{m\ of}$ 56.3°C) than the eukaryotic one (a $T_{m\ of}$ 49°C for the bovine enzyme). This difference is likely to reflect the effect of the loops on the structure. in both cases, the stabilization by the binding of the ligands is substantial: In the *E. coli* DHFR, binding of the NADP(H) raises the T_m to 59.4°C. Binding of the inhibitor (methotrexate) raises it to 70.5°C, and with both bound to the enzyme the T_m climbs to 78.2°C. For the bovine case, we see the same trend: binding of the NADP(H) raises the T_m to 57.3°C, the methotrexate to 62.2, and together to 78.1°C.[105] The larger stabilization exerted by the binding of the methotrexate inhibitor when compared with the NADP(H) is consistent with the building block cuttings of both enzymes: the yellow building block (Plates 5A,D for the *E. coli* and Plate 6C for the eukaryotic enzyme) is not involved to a considerable extent with the binding of neither of the ligands (only 4 out of 29 residues are in contact with either dihydrofolate or NADP(H)). The NADP(H) binds to an already stable ABD building block. However, the dihydrofolate substrate (or the methotrexate inhibitor) binds largely to the unstable red building block, and thereby confers a larger stabilization to the structure.

Superimposing the two structures (Plate 7) shows their high similarity. The difference between them is apparent at three locations: a loop at the junction of the red and green building blocks (at position 36 in the *E. coli* enzyme); a loop in the green ABD building block, at position 86; and a loop in the yellow building block, at position 136. As the native bacterial enzyme does not bind to the GroEL, and the human and murine enzymes do, it appeared that the origin of the different behavior may reside in the loops. By grafting the loops into the bacterial DHFR, Clark et al.[106] have shown that the first and third loops are responsible for the binding to the GroEL. The finding that the eukaryotic DHFR binds to the chaperonin through its loops is consistent with recent structural data. Crystal[107,108] structures clearly show that the GroES and the peptides bind to GroEL through a mobile loop, with an extended β conformation, or via a β-strand. The first and third loops of the eukaryotic DHFR resemble the mobile loop of GroES. Hence, it is conceivable that the enzyme would bind to the GroEL (and to the hsp60) through these loops. Binding would be strongest if both loops bind simultaneously to two binding sites in the chaperonin subunits. While such simultaneous binding is likely to be sterically infeasible in the native state of the enzyme, it may take place in an intermediate state, suggesting why a misfolded conformation binds to the chaperonin, whereas the native state does not. Further, the tighter binding of the late intermediates is consistent with the building blocks, and hence the loops, already in their native conformations, however, with nonnative contacts between the building block elements. The formation of such an intermediate state may represent the slow step in the folding of the enzyme.

Interestingly, Clark and Frieden[98] have shown that while the native *E. coli* enzyme does not bind to GroEL, a late intermediate does. The fact that an intermediate state of the bacterial DHFR also binds to the GroEL can be understood through the low stability of the red, amino-terminus building block (Plate 5A). As Plate 5C and Table 1 show, the score of this building block is -4.6, pointing to a very low population time. Therefore, it is conceivable that during folding this building may adopt an extended, mobile loop-like conformation, enabling its binding to the GroEL. In such a scenario, when the red building block is flipped out, the green and yellow building blocks collapse onto each other, forming the stable native-like intermediate, with a significant number of native interactions, consistent with the proposition of Goldberg et al.[97] The general lower stability of the eukaryotic enzyme, coupled with the location of the loops in the building blocks cutting, is consistent with the experimental observation of Clark and Frieden, namely, that when complexed with the DHFR enzymes the GroEL stabilizes the *E. coli* enzyme only above 30°C, whereas the murine enzyme is stabilized already in the 8 to 42°C range.

Nevertheless, despite the fact that *in vivo* the folding of the eukaryotic enzyme is assisted by the chaperonin, there is evidence that both enzymes are able to fold spontaneously *in vitro*.[111] Hence, if the concentration is low, the slow-folding DHFR will eventually get out of the well on its energy landscape where the misfolded conformations reside to continue on the folding route. If, however, the concentration is high, the misfolded conformations will aggregate. Once an aggregate forms, it remains in this most stable form. Consistently, *in vivo*, in the absence of the hsp60 chaperonin, the eukaryotic DHFR aggregates.[112] This suggests that there is a higher affinity between the chaperonin and the misfolded DHFR when compared with the affinity between misfolded conformations. The competitive GroEL vs. aggregate binding suggests that the loops play a role in the formation of the aggregate as well. Shtilerman et al.[113] have presented ample data to support the view that the chaperonin acts via unfolding misfolded molecules. Here, however, as the eukaryotic enzyme is able to fold without the assistance of the chaperonin under proper dilute buffer conditions, it appears that the chaperonin acts primarily by preventing aggregation through competition. The tight binding of the DHFR loops to the GroEL is indicated by the need for ATP to release the enzyme. The hydrolysis of ATP coupled with the allosteric movement of the GroEL renders the binding of the DHFR loop weaker, losing the competition to the mobile GroES loop. This scenario, where the misfolded DHFR conformations are able to climb out of the well without assistance, indicates that the well is most likely not too deep.

I. Amyloid Formation

At least for several proteins, it has been shown that a conformational change may take place leading to formation of a regular amyloid fibril structure. Amyloids are typically composed of propagating twisted β-sheets, which are extremely stable to heat, as well as to protease digestion. Analysis of X-ray fiber diffraction has yielded an average of twist between consecutive β-strands. Detailed studies for some of these proteins, such as from A β and prion, have shown that certain peptides derived from these, can already form the characteristic amyloid fibrils observed for their full-length parent proteins. Further, particles originating from these have been observed to be toxic to human and rat cells.

The fact that these proteins, and their corresponding derived short peptides, are amyloidogenic and toxic raises several interesting questions: (1) How is the initial amyloid seed stabilized? (2) What is the smallest size that a seed can be? (3) What is the mechanism of seed growth? Our recent simulations of an 8-residue peptide (AGAAAAGA), derived from Syrian hamster prion protein (Ma and Nussinov, unpublished), have suggested that an octamer of such of a peptide (and of A) is stable enough to serve as a seed. Further, they have shown that the driving force is the hydrophobic effect, and that fibril growth is via conformational *selection*, rather than being *induced* by the preformed seed/oligomer through binding. The slow step is seed formation, rather than its growth. These results are in remarkable agreement with the experimental observations of Serio et al.,[53] and provide a detailed molecular model of seed formation and growth. They have further led to the proposition of the "lamellar model", where peptide oligomers assemble as relatively dynamic and twisted sheets to yield the average experimentally derived 15 twist angle.

The peptide may be considered a conformationally fluctuating, unstable building block. Because it lacks a hydrophobic core, its population time is very low. Hence, it cannot be detected experimentally, but may be observed in simulations. By way of conformational selection, it binds to the amyloid seed, with a redistribution of the populaion, further driving the binding reaction. Plate 8 provides an illustration from our simulation.

J. Disorder

Molecular disorder has long been viewed as local, or as global instability. Molecules (or regions) that display disorder were considered unstructured. Molecules (or their portions) have been considered disordered if no atomic coordinates could be assigned. It has been suggested that for such molecules, the structured state is induced through binding to their cognate ligands. A comprehensive review of such molecules, with numerous examples has been published recently.[50] Nevertheless, even in apparently "disordered" molecules, prevailing conformations exist, with higher population times than for all other potential conformations. Hence, these molecules, or regions, are structured, with preformed conformers binding via conformational selection. Even if the population times of these conformations are low, after their binding there is a redistribution of the populations, further driving the binding reaction.

The conformational times are a function of the presence of hydrophobic cores. If the molecule possesses a hydrophobic core, the conformational time may be high enough to be detected. In its absence, the time a molecule spends in this conformation is likely to be too low to enable such a detection. In addition, the presence of a net charge and thus electrostatic repulsion has also been shown to lead to natively disordered state[52] such as observed in the case of α-lactalbumin[114] or in the adenine binding domain of *E. coli* dihydrofolate reductase.[85] In α-lactalbumin the disordered molten globule state is observed either after lowering of the pH or when removing the C_{α}^{++} ion. Whereas in these two examples after binding (correspondingly to calcium, or to NADPH) the net charge is nullified and "order" is achieved, resulting in well-defined crystal coordinates, this is not necessarily always the case. In the case of the H-Ras protein, the opposite is observed. H-Ras is a G-protein of 189 residues that serves as a molecular switch coupling cell-surface receptors to intracellular signalling pathways.[115] Here, when H-Ras is bound to both GTP and the GAP (GTPase activating transmembrane protein), Ras presents a well-defined, ordered structure. However, after the hydrolysis of GTP to GDP, a (charged) phosphate group dissociates from the complex. This leads to the Ras helix α2 and loop L4 at the binding site of GAP to become disordered, with the outcome of GAP dissociating next, and Ras again displaying an ordered conformation.

"Disordered" molecules are the outcome of rugged energy landscape away from the native state, around the bottom of the funnels. Such ruggedness creates a range of conformations, and as such has biological function, either in regulation and/or in multimolecular complex formation in complex cellular pathways.

K. Hydrogen Exchange Data and Limitations of the Cutting Algorithm

A recent review by Englander and his colleagues[116] summarized the data on folding intermediates as studied by hydrogen exchange (HX) methods. The model presented by Rumbley et al. has many features of similarity to the one outlined here. HX data indicate that the amino acid sequences in proteins stabilize not only the native state, but also a small set of native-like intermediates. The intermediates form from cooperative secondary structure elements of the native protein. Rumbley et al. suggest a process of 'sequential stabilization', where early intermediates guide in a stepwise manner sequential addition of these structural elements, constructing a folding pathway. Most importantly, HX data indicate that the native structure guides the folding, with the sequence determining the intermediates, and their kinetic accessibility. The data are consistent with a model where the intermediates are on-pathway (e.g., Laurents et al.[117] have shown it for ribonuclease A;

Bai[118,119] for Cyt c and hen lysozyme and additional workers [reviewed in Rumbley et al., Ref. 116], for additional cases as well). In cases where the intermediates are off-pathway, it is unclear if these less frequent off-pathway cases reach the native state.

HX-based methods are able to detect conformations of infinitesimally populated intermediates at equilibrium, as well as kinetic intermediates with subsecond lifetimes. HX exchange has enabled exploration of the free energy landscape between the native and the unfolded state.[120–123] Chamberlain et al.[124] have listed cases where where parts of proteins have been prepared and shown to have native-like structure.

Plates 9A to C* depict the fragment map, the anatomy tree and the building blocks cutting of Cyt c (1ycc), and the corresponding hydrophobic folding units (HFU) association, so we can compare it with the HX data as presented by Rumbley et al.[116] The HX data relate to the unfolding pathway. Four distinct cooperative unfolding units can be seen in the HX results (Figure 3D in Rumbley et al.). The least stable is the loop connecting the C-terminal helix to the previous helix, termed the 60s helix by Rumbley et al. The second loop connecting the 60s helix to the N-terminus helix is slightly more stable. According to the HX data, folding initiates by the N and C termini helices coming together, forming a concerted folding unit. Cutting the Cyt c structure (Plate 9C, top row) has also generated four building blocks at the last cutting level. While we observe the termini as separate building blocks, as the anatomy tree (Plate 9B) shows, each is unstable, possibly explaining why they are not seen by the HX. By way of association they get stabilized and observed. This can be seen in the HFU (bottom row, Plate 9C), where the termini form one HFU (green in the figure). We note, however, that while the threshold to obtain this HFU (1.39) is lower than the one set for an HFU (1.65), this latter value has not been determined systematically to reflect experimental results. The building blocks cutting follows the folding pathway. During unfolding, the green and yellow building blocks in the top row of Plate 9C may unfold first, followed by the associated termini. In folding, two terms need to be considered: the stability of the associated helices and the entropy of their coming together. This is not the situation in the unfolding, with the sole term being that of the stability.

This example serves to illustrate a limitation of the cutting algorithm. Since the N and the C termini are not sequentially linked, they cannot be a single building block. This suggests a modification to the algorithm, where the edges be artificially linked, and a recutting of this linked region will be executed. Nevertheless, a favorable association is reflected in the HFU. An additional, critical limitation relates to the scoring function. The criteria in the scoring function are the buried nonpolar surface area, compactness, and the ‘isolatedness’, that is, the area originally buried and exposed after cutting. However, electrostatics are not accounted for. Because electrostatics has been increasing shown to play a critical role in protein stability, this is a major drawback of the function.

CONCLUSIONS

Here we have described a hierarchical protein folding model. We have

* Plate 9Afollows page 426.

outlined the elements that go into the model, namely, (1) the fluctuating building block fragments that constitute local minima along the protein sequence, (2) consideration of the landscape around the bottom of the funnel, (3) the postulate that folding and binding are similar processes conforming to the same principles with similar landscapes describing both, and (4) that the lanscape is dynamic, changing with the conditions. (5) Furthermore, the conformationally fluctuating building blocks bind via conformational *selection*. Thus, even if the population time of the native structure of the building block is low, through conformational selection, the equilibruim will shift in its favor. These are then the two critical elements: fluctuating building blocks binding through selection.

We illustrate that such a model is consistent with a broad range of experimental (and computational) results. Hence, in principle such a model may be applied to eventually aid in prediction of folded structures. Here, one may envision cutting the sequence into building block fragments, simulating these and (iteratively) combinatorially assembling them. The most difficult step in such a venture would no doubt be the combinatorial assembly.

Currently, the algorithm cuts the structure into such local building block elements, estimating their stability. The underlying assumption on which the procedure is based is that the structures of the building blocks that we see in the native fold are the ones with the highest population time. Thus, had the structure been experimentally cut into these peptide fragments, their conformations would resemble those that we see in the entire protein. This implies that native contacts guide the protein folding.

Consistently, current knowledge indicates that while there are multiple pathways, not all are equally probable. Pathways are a function of fragment population times, reflected in the stability of building block protein fragments. Current knowledge further indicates that intermediate states consist largely of native contacts. These mostly derive from *intrabuilding block* interactions. Further, while sequentially folding proteins are likely to further illustrate appreciable native *interbuilding block* contacts, intermediate states in nonsequential protein folders (such as the DHFR), may involve nonnative contacts between misassociating building blocks. Because native contacts guide the folding pathways, the breadth of the transition states is limited, consistent with the view that in reality the difference between the “old view” and the “new view” is not large.[71] The building block folding model further provides a link between two-state “hydrophobic collapse” and the hierarchical model, which is three-state. In two-state protein folding intermediates are not observed, as the process is too fast, the well too shallow, and thus the population times too low. Nevertheless, two- and three-state follow similar events.

Here we have shown detailed analysis for one case, the dihydrofolate reductase. The consistency of the theoretical results with the spectroscopic observations for the *E. coli* DHFR further validates the building block folding model and the hierarchical folding of proteins. The application of the dissecting algorithm to the human enzyme yielded similar pathways. This is not surprising, as the enzymes are highly similar, with the exception of three loops, present only in the eukaryotic DHFR. Two of these have been shown to be the means through which the eukaryotic DHFR binds to the chaperonin. Hence, in particular, the building blocks folding model has enabled us to address the difference between the *in vivo* eukaryotic chaperonin-assisted, and *E. coli* unassisted folding, and the likely intermediate states. It has further enabled addressing the binding modes of the late intermediate states to the GroEL chaperonin.

An advantage of the building block folding model over residue-based models such as the contact order is that unlike contact order it provides for population times. Topology by itself is static. For this reason, contact order cannot provide for changes in conditions, or for mutations. Because it does not divide the sequence/structure into components, the relative contact order is unable to address the hierarchy in folding pathways. As such, the relative contact order is unable to account for the *redistribution* of conformational substates.[79,125]

To conclude, here we argue that the combination of the fluctuating building blocks in their native conformations binding via selection in folding, resembles stable molecules binding via selection in intermolecular binding. The differences stem from the difference in stabilities, and hence in population times. Additionally, because in intermolecular binding there is no chain connectivity between the molecules, the question of sequential vs. nonsequential binding events does not arise. Thus, the model provides a realistic view of binding and folding events.

The results of the anatomy for each chain in PDB can be viewed at the web site **http://protein3d.ncifcrf.gov/tsai/anatomy.html.**

ACKNOWLEDGMENTS

We thank Dr. Jacob Maizel for discussions and encouragement, and the personnel at FCRDC for their assistance. The research of H. J. Wolfson and R. Nussinov in Israel has been supported in part by the Magnet grant, by the Ministry of Science grant, and by the "Center of Excellence in Geometric Computing and its Applications" funded by the Israel Science Foundation (administered by the Israel Academy of Sciences). The research of H.J.W. is partially supported by the Hermann Minkowski-Minerva Center for Geometry at Tel Aviv University. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number NO1-CO-56000. The content of this publication does not necessarily reflect the view or policies of the Department of Health and Human Services, or does the mention of tradenames, commercial products, or organization imply endorsement by the U.S. Government.

REFERENCES

1. Kim, P. S. and Baldwin, R. L. 1982. Specific intermediates in the folding reactions of small proteins, and the mechanism of protein folding. *Annu. Rev. Biochem.* **51**:459–489.
2. Kim, P. S. and Baldwin, R. L. 1990. Intermediates in the folding reactions of small proteins. *Annu. Rev. Biochem.* **59**:631–660.
3. Udgaonkar, J. B. and Baldwin, R. L. 1988. NMR evidence for an early frameork interdediates on the folding

pathway of ribonuclease A. *Nature* **335**:694–699.

4. Wetlaufer D. B. 1973. Nucleation, rapid folding and globular intrachain regions in proteins. *Proc Natl Acad Sci USA* **70**:697–701.

5. Shakhnovitch, E., Abkevitch, V. and Ptitsyn, O. 1996. Conserved residues and the mechanism of protein folding. *Nature* **379**:96–98.

6. Fersht, A. R. 1997. Nucleation mechanism in protein folding. *Curr Opin Struct Biol* **7**:3–9.

7. Karplus, M. and Weaver, D. L. 1994. Protein folding dynamics: the diffusion-collision model and experimental data. *Prot Sci* **3**:650–668.

8. Rackovsky, S. and Scheraga, H. A. 1977. Hydrophobicity, hydrophilicity and the radial and orientational distributions of residues in native proteins. *Proc Natl Acad Sci USA* **74**:5248–5251.

9. Dill, K. A. 1985. Theory for the folding and stability of globular proteins. *Biochemistry* **24**:1501–1509.

10. Dill, K. A. 1990. Dominant forces in protein folding. *Biochemistry* **29**:7135–7155.

11. Baldwin, R.L. and Rose, G.D. 1999. Is protein folding hierarchic? I. Local structure and peptide folding. *TIBS* **24**:26–33.

12. Baldwin, R.L. and Rose, G. D. 1999. Is protein folding hierarchic? II. Folding intermediates and transition states. *TIBS* **24**:77–84.

13. Chiti, F., Taddei, N., Webster, P., Hamada, D., Fiaschi, T., Ramponi, G., and Dobson, C.M. 1999. Acceleration of the folding of acylphosphatase by stabilization of local secondary structures. *Nat Struct Biol* **6**:380–386.

14. Hamada, D., Chiti, F., Guijarro, I., Kataoka, M., Taddei, N., and Dobson, C. M. 2000. Evidence concerning rate-limiting steps in protein folding from the effects of trifluooethanol. *Nat Struct Biol* **7**:58–61.

15. Sabelko, J, Ervin, J and Gruebele, M. 1999. Observations of strange kinetics in protein folding. *Proc Natl Acad Sci USA 96*:6031–6036.

16. Ionescu, R.M. and Matthews, C.M. 1999. Folding under the influence. *Nat Struct Biol* **6**:304–307.

17. Gualfetti, P. J., Bilsel, O., and Mattews, C. R. 1999. The progressive development of structure and stability during the equilibrium folding of the α subunit of tryptophan synthase from *E. coli*. *Protein Sci* **8**:1623–1635.

18. Park, S.H., O'Neil, K.T., and Roder, H. 1997. An early intermediate in the folding reaction of the B1 domain of protein G contains a native-like core. *Biochemistry* **36**:14277–14283.

19. Zitzewitz, J.A., Gualfetti, P.J., Perkons, I.A., Wasta, S.A., and Matthews, C.R. 1999. Identifying the structural boundaries of independent folding domains in the α subunit of tryptophan synthase, a β/α barrel protein. *Protein Sci* **8**:1200–1209.

20. Gegg, C.V., Bowers, K.E., and Matthews, C.R. 1997. Probing minimal independent folding units in dihydrofolate reductase by molecular dissection. *Protein Sci* **6**:1885–1892.

21. Dyson, H. J., Merutka, G., Waltho, J. P., Lerner, R. A., and Wright, P. E. 1992. Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding I. Myohemerythrin. *J Mol Biol* **226**:795–817.

22. Dyson, H. J., Sayre, J. R., Merutka, G., Shin, H.-C., Lerner, R. A., and Wright, P. E. 1992. Folding of peptide fragments comprising the complete sequence of proteins. Models for initia-

tion of protein folding II. Plastocyanin. *J Mol Biol* **226**:819–835.

23. Prat-Gay, G. 1996. Association of complementary fragments and the elucidation of protein folding pathways. *Prot Eng* **9**:843–847.

24. Yokota, A., Takenaka, H., Oh, T., Noda, Y., and Segawa, S.-I. 1998. Thermodynamics of the reconstitution of tuna cytochrome c from two peptides. *Protein Sci* **7**:1717–1727.

26. Sancho, J. and Fersht, A. R. 1992. Dissection of an enzyme by protein engineering. The N and C-terminal fragments of barnase from a native-like complex with restored enzymic activity. *J Mol Biol* **224**:741–747.

27. Yang, X.-M., Yu, W.-F., Li, J.-H., Fuchs, J., Rizo, J., and Tasayco, M. L. 1998. NMR evidence for the reassembly of an α/β domain after cleavage of an α-helix: implications for protein design. *J Am Chem Soc* **120**:7985–7986.

28. Kobayashi, N., Honda, S., Yoshii, H., Uedaira, H., and Munekata, E. 1995. Complement assembly of two fragments of the streptococcal protein G B1 domain in aqueous solution. *FEBS Lett* **366**:99–103.

29. Neira, J.L. and Fersht, A.R. 1999. Exploring the folding funnel of a polypeptide chain by biophysical studies on protein fragments. *J Mol Biol* **285**:1309–1333.

30. Honda, S., Kobayashi, N., Munekata, E., and Uedaira, H. 1999. Fragment reconstitution of a small protein: folding energetics of the reconstituted immunoglobulin binding domain B1 of streptococcal protein G. *Biochemistry* **38**:1203–1213.

31. Prat-Gay, G. and Fersht, A. 1994. Generation of a family of protein fragments for structure-folding studies. I. Folding complementation of two fragments of chymotrypsin inhibitor-2 formed by cleavage at its unique methionine residue. *Biochemistry* **33**:7957–7963.

32. Tasayco, M.L. and Chao, K. 1995. NMR study of the reconstitution of the beta-sheet of thioredoxin by fragment cmplementation. *Proteins* **22**:41–44.

33. Georgescu, R. E., Braswell, E. H., Zhu, D. and Tasayco, M. L. 1999. Energetics of assembling an artificial heterodimer with an α/β motif: cleaved versus uncleaved *Escherichia coli* thioredoxin. *Biochemistry* **38**:13355–13366.

34. Burton, R.E., Myers, J.K. and Oas, T.G. 1998. Protein folding dynamics: quantitative comparison between theory and experiment. *Biochemistry* **37**:5337–5343.

35. Tsuji, T., Yoshida, K., Satoh, A., Kohno, T., Kobayashi, K., and Yanagawa, H. 1999. Foldability of barnase mutants obtained by permutaion of modules or secondary structure units. *J Mol Biol*, **268**:1581–1596.

36. Rothwarf, D. M. and Scheraga, H. A. 1996. Role of non-native aromatic and hydrophobic interactions in the folding of hen egg white lysozyme. *Biochemistry* **35**:13797–13807.

37. Shortle, D. R. 1996 Structural analysis of non-native states of proteins by NMR methods. *Curr Opin Struct Biol* **6**:24–30.

38. Shao, X. and Matthews, C. R. 1998. Single-tryptophan mutants of monomeric tryptophan repressor: optical spectroscopy reveals nonnative structure in a model for an early folding intermediates. *Biochemistry* **37**:7850–7858.

39. Houry, W. A., Frishman, D., Eckerson, C., Lottspeich, F., and Hartl, F. U. 1999. Identification of *in vivo* sub-

strates of the chaperonin GroEl. *Nature* **402**:147–154.

40. Netzer, W. J. and Hartl, F. U. 1998. Protein folding in the cytosol: chaperone-dependent and -independent mechanisms. *Trends Biochem Sci* **23**:68–73.

41. Baker, D., Shiau, A. K., and Agard, D. A. 1993. The role of pro regions in protein folding. *Curr Opin Cell Biol* **5**:966–970.

42. Shinde, U. P., Liu, J. J. and Inoye, M. 1997. Protein memory through altered folding mediated by intramolecular chaperones. *Nature* **389**:520–522.

43. Sauter, N. K., Mau, T., Rader, S. D., and Agard, D. A. 1998. Structure of α-lytic protease complexed with its proregion. *Nature Struct Biol* **5**:945–950.

44. Sohl, J. S., Jaswal, S. S., and Agard, D. A. 1998. Unfolded conformations of α-lytic protease are more stable than its native state. *Nature* **395**:817–819.

45. Wang, L., Ruan, B., Ruvinov, S., and Bryan, P. N. 1998. Engineering the independent folding of subtilisin BPN' prodomain: correlation of pro-domain stability with the rate of subtilisin folding. *Biochemistry* **37**:3165–3171.

46. Anderson, D. E., Peters, R. J., Wilk, B., and Agard, D. A. 1999. α-lytic protease precursor: characterization of a structured folding intermediate. *Biochemistry* **38**:4728–4735.

47. Cunningham, E. L., Jaswal, S. J., Sohl, J. L., and Agard, D. A. 1999. Kinetic stability as a mechanism for protease longevity. *Proc Natl Acad Sci USA* **96**:1108–11014.

48. Ellis, R. J. 1998. Steric chaperones. *Trends Biochem Sci* **23**:43–45.

49. Baker, D., Sohl, J. L., and Agard, D. A. 1992. A protein folding reaction under kinetic control. *Nature* **356**:263–265.

50. Wright, P.E. and Dyson, H.J. 1999. Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. *J Mol Biol* **293**:321–331.

51. Zhang, J. and Matthews, C.R. 1998. Ligand binding is the principal determinant of stability for the p21$^{H\text{–}ras}$ protein. *Biochemistry* **37**:14881–14890.

52. Uversky, V.N., Gillespie, J.R., and Fink, A.L. 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**:415–427.

53. Serio, T. R., Gashikar, A. G., Kowal, A. S., Sawicki, G. J., Moslehi, J. J., Serpell, L., Arnsdorf, M. F., and Lindquist, S. L. 2000. Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science* **289**:1317–1321.

54. Kelly, J. W. 2000. Mechanism of amyloidogenesis. *Nat Struct Biol* **7**:824–826.

55. Alm, E. and Baker, D. 1999. Matching theory and experiment in protein folding. *Curr Opin Struct Biol* **9**:189–196.

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

57. Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M.A., Jaenicke, R., and Schmid, F.X. 1998. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nat Struct Biol*, **5**:229–235.

58. Kiefhaber, T. 1995. Kinetic traps in lysozyme folding. *Proc Natl Acad Sci USA* **92**:9029–9033.

59. Capaldi, A. P., Ferguson, S. J., and Radford, S. E. 1999. The Greek key protein apo-pseudoazurin folds through

an obligate on-pathway intermediate. *J Mol Biol* **286**:1621–1632.

60. Lopez-Hernandez, E. and Serrano, L. 1996. Structure of the transition state for folding of the 129 aa protein Che Y resembles that of a smaller protein, CI-2? *Fold Design*, **1**:43–55.

61. Choe, S.E., Matsudaira, P.T., Osterhout, J., Wagner, G., and Shakhnovitch, E.I. 1998. Folding kinetics of villin 14T, a protein domain with a central β-sheet and two hydrophobic cores. *Biochemistry* **37**:14508–14518.

62. Matagne, A., Radford, S. E., and Dobson, C. M. 1997. Fast and slow tracks in lysozyme folding: insight into the role of domains in the folding process. *J Mol Biol* **267**:1068–1074.

63. van Nuland, N.A.J., et al. 1998. Slow folding of muscle acylphosphatase in the absence of intermediates. *J Mol Biol*, **283**:883–891.

64. Bilsel, O., Zitzewitz, J. A., Bowers, K. E. and Matthews, R. C. 1999. Folding mechanism of the -subunit of tryptophan synthase, an /β barrel protein: global analysis highlights the interconversion of multiple native, intermediate, and unfolded forms through parallel channels. *Biochemistry* **38**:1018–1029.

65. Panchenko, A.R., Luthey-Schulten, Z., Cole, R., and Wolynes, P.G. 1997. The foldon universe: A survey of structural similarity and self-recognition of independently folding units. *J Mol Biol* **272**:95–105.

66. Hansmann, U. H. E., Okamoto, Y., and Onuchic, J. N. 1999. The folding funnel landscape for the peptide met-enkephalin. *Proteins* **34**:472–483.

67. Frauenfelder, H. and Leeson, D. T. 1998. The energy landscape in non-biological molecules. *Nat Struct Biol* **5**:757–759.

69. Shoemaker, B. A. and Wolynes, P. G. 1999. Exploring structures in protein folding funnels with free energy functions: the denatured ensemble. *J Mol Biol* **267**:657–674.

70. Shoemaker Wang, J. and Wolynes, P. G. 1999. Exploring structures in protein folding funnels with free energy functions: the transition state ensemble. *J Mol Biol* **267**:675–694.

71. Pande, V.S., Grosberg, A.Y., Tanaka, T., and Rokhsar, D.S. 1998. Pathways for protein folding: Is a new view needed? *Curr Opin Struct Biol* **8**:66–79.

72. Pande, V. S. and Rokhsar, D. S. 1999. Molecular dynamics simulations of unfolding and refolding of βhairpin fragements of protein G. *Proc Natl Acad Sci USA* **96**:9062–9067.

73. Tsai, C.J., Xu, D., and Nussinov, R. 1998. Protein folding via binding, and vice versa. *Fold Design* **3**:R71–R80.

74. Tsai, C.J., Kumar, S., Ma, B., and Nussinov R. 1999. Folding funnels, binding funnels and protein function. *Protein Sci* **8**:1181–1190.

75. Tsai, C-J, Maizel, J. V., and Nussinov, R. 1999. Distinguishing between sequential and non-sequentially folded proteins: implications for folding and misfolding. *Protein Sci* **37**:73–87.

76. Tsai, C.J., Ma, B., and Nussinov, R. 1999. Folding and binding cascades: shifts in energy landscapes. *Proc Natl Acad Sci USA* **96**:9970–9972.

77. Tsai, C.J., Maizel, J.V., and Nussinov, R. 2000. Anatomy of protein structures: visualizing how a 1D protein chain folds into a 3D shape. *Proc Natl Acad Sci USA* **97**:12038–12043.

78. Kumar, S., Ma, B., Tsai, C.-J., Wolfson, H. and Nussinov, R. 1999. Folding funnels and conformational

transitions via hinge-bending motions. *Cell Biochem Biophys* **31**:23–46.

79. Kumar, S., Ma, B., Tsai, C.J., Sinha, N. and Nussinov, R. 2000. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci* **9**:10–19.

80. Ma, B., Kumar, S., Tsai, C-J. and Nussinov, R. 1999. Folding funnels and binding mechanisms. *Protein Eng* **12**:713–720.

81. Ma, B., Tsai, C.J., and Nussinov, R. 2000. Binding and folding: in search of intramolecular chaperone-like building block fragments. *Protein Eng* **13**:617–627.

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

83. Jackson, S.E. 1998. How do small single domain proteins fold? *Fold Design* **3**:R81–R91.

84. Llinas, M. and Marqusee, S. 1998. Subdomain interactions as a determinant in the folding and stability of T4 lysozyme. *Protein Sci* **7**:96–104.

85. Sham, Y.Y., Ma. B., Tsai, C.J. and Nussinov, R. 2001. Molecular dynamic simulation of Escherichia coli dihydrofolate reductase and its protein fragments: relative stabilities in experiment and simulations. *Protein Sci* 10:135–148.

86. Bennett, M. J., Schlunegger, M. P., and Eisenberg, D. 1995. 3D domain swapping: A mechanism for oligomer assembly. *Protein Sci* **4**:2455–2468.

87. Munioz, V. and Eaton, W. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc Natl Acad Sci USA* **96**:1131–1136.

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

89. Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., Doshi, N., Grantcharova, V.P., Yi, O., and Baker, D. 1997. Functional rapidly folding proteins from simplified amino acid sequences. *Nat Struct Biol*, **4**:805–809.

90. Martinez, J.C., Pisabarro, M.T., and Serrano, L. 1998. Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat Struct Biol* **5**:721–729.

91. Polverino de Laureto, P., Scaramella, E., Frigo, M., Wondrich, F.G., De Filippis, V., Zambonin, M., and Fontana, A. 1999. Limited proteolysis of bovine α-lactalbumin: isolation and characterization of protein domains. *Protein Sci* **8**:2290–2303.

92. Fontana, A., Zambonin, M., Polverino de Laureto, P., De Filippis, V., Clementi, A., and Scaramella, E. 1997. Probing the conformational state of apomyoglobin by limited proteolysis. *J Mol Biol* **266**:223–230.

93. Oas, T. G. and Kim, P. S. 1988. A peptide model of protein folding intermediate. *Nature* **336**:42–48.

94. Tasayco, M.L. and Carey, J. 1992. Ordered self-assembly of polypeptide fragments to form native-like dimeric trp represoor. *Science* **255**:594–597.

95. Wu, L. C., Grandori, R., and Carey, J. 1994. Autonomous subdomains in protein folding. *Protein Sci* **3**:359–371.

96. Jones, B.E. and Matthews, C. R. 1995. Early intermediates in the folding of dihydrofolate reductase from Escherichia coli detected by hydrogen exchange and NMR. *Protein Sci* **4**:167–177.

97. Goldberg, M.S., Zhang, J., Sondek, S., Matthews, C.R., Fox, R.O., and Horwich, A.L. 1997. Native-like structure of a protein-folding intermediate bound to the chaperonin GroEL. *Proc Natl Acad Sci USA* **94**:1080–1085.

98. Clark, A.C. and Frieden, C. 1999. The chaperonin GroEL binds to late-folding non-native conformations present in native *Escherichia coli* and murine dihydrofolate reductases. *J Mol Biol* **285**:1777–1788.

99. Bystroff, C., Oatley, S.J., and Kraut, J. 1990. Crystal structures of Escherichia coli dihydrofolate reductase: the NADP+ holoenzyme and the folate-NADP+ ternary complex, substrate binding and a model for the transition state. *Biochemistry* **30**:8067–8074.

100. Sawaya, M.R. and Kraut, J. 1997. Loop and subdomain movements in the mechanism of *Escherichia coli* dihydrofolate reductase: crystallographic evidence. *Biochemistry* **36**:586–603.

101. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F. Jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. The protein databank: a computer-based archival file for macromolecular structures. *J Mol Biol* **112**:535–542.

102. Kumar, S., Sham, Y. Y., Tsai, C.-J., and Nussinov, R. 2001. Protein folding and function: the N-terminal fragment in adenylate kinase. Biophys J, in press.

103. Clark, A.C. and Frieden, C. 1999. Native *Escherichia coli* and murine dihydrofolate reductases contain late-folding non-native structures. *J Mol Biol* **285**:1765–1776.

104. Clark, A. C. and Frieden, C. 1997. GroEL-mediated folding of structurally homologous dihydrofolate reductases. *J Mol Biol* **268**:512–525.

105. Pfeil W. 1998. Protein Stability and Folding: A Collection of Thermodynamic Data. Springer-Verlag, Berlin. 657 pp.

106. Clark, A.C, Hugo, E., and Frieden, C. 1996. Determination of regions in the dihydrofolate reductase structure that interact with the molecular chaperonin GroEL. *Biochemistry* 35:5893–5901.

107. Xu, D., Lin, S. L., and Nussinov, R. 1997. Protein binding *vs.* protein folding: the role of hydrophilic bridges in protein associations. *J Mol Biol* **265**:68–84.

108. Chen, L. and Sigler, P.B. 1999. The crystal structure of a GroEL/peptide complex: plasticity as a basis for substrate diversity. *Cell* **99**:757–768.

109. Chatellier, J., Buckle, A.M., and Fersht, A.R. 1999. GroEL recognizes sequential and non-sequential linear structural motifs compatible with extended β-strands and α-helices. *J Mol Biol* **292**:163–172.

110. Tanaka, N. and Fersht, A.R. 1999. Identification of substrate binding site of GroEL minichaperone in solution. *J Mol Biol* **292**:173–180.

111. Eilers, M., Hwang, S., and Schatz, G. 1988. Unfolding and refolding of a purified precursor protein during import into isolated mitochondria. *EMBO J* **7**:1139–1145.

112. Ostermann, J., Horwich, A.L., Neupert, W., and Hartl, F.U. 1989. Protein folding in mitochondria requires complex formation with hsp60 and ATP hydrolysis. *Nature* **341**:125–130.

113. Shtilerman, M., Lorimer, G.H., and Englander, S.W. 1999. Chaperonin function: folding by forced unfolding. *Science* **284**:822–825.

114. Demarest, S.J., Fairman, R., and Raleigh, D.P. 1998. Peptide models of

local and long-range interactions in the molten globule state of human α-lactalbumin. *J Mol Biol* **283**:279–291.

115. Sprang, S. R. 1997 G protein mechanisms: insights from structural analysis. *Annu Rev Biochem* **66**:639–678.

116. Rumbley, J., Hoang, L., Mayne, L., and Englander, S. W. 2001. An amino acid code for protein folding. *Proc Natl Acad Sci USA* **98**:105–112.

117. Laurents, D. V., Bruix, M., Jamin, M., and Baldwin, R. L. 1998. A pulse-chase-competition experiment to determine if a folding intermediate is on or off pathway: application to ribonuclease A. *J Mol Biol* **283**:669–678.

118. Bai, Y. 1999. Kinetic evidence for an on-pathway intermediate in the folding of cytochrome c. *Proc Natl Acad Sci USA* **96**:477–480.

119. Bai, Y. 2000. Kinetic evidence of an on-pathway intermediate in the folding of lysozyme. *Protein Sci* **9**:194–196.

120. Bai, Y., Sosnick, T. R., Mayne, L., and Englander, S. W. 1995. Protein folding intermediates: native state hydrogen exchange. *Science* **269**:192–197.

121. Bai, Y. and Englander, S. W. 2000. Future directions in folding: the multistate nature of protein structure. *Proteins* **24**:145–151.

123. Milne, J. S., Mayne, L., Roder, H., Wand, A. J., and Englander, S. W. 1998. *Protein Sci* **7**:739–745.

124. Chamberlain, A. K., Fischer, K. F., Reardon, D., Handel, T. M., and Marqusee, A. S. 1999. Folding of an isolated ribonuclease H core fragment. *Protein Sci* **8**:2251–2257.

125. Freire, E. 1999. The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody D1.3 to lysozyme. *Proc Natl Acad Sci USA* **96**:10118–10122.