Is protein folding hierarchic? II. Folding intermediates and transition states

Robert L. Baldwin and George D. Rose

The folding reactions of some small proteins show clear evidence of a hierarchic process, whereas others, lacking detectable intermediates, do not. Evidence from folding intermediates and transition states suggests that folding begins locally, and that the formation of native secondary structure precedes the formation of tertiary interactions, not the reverse. Some notable examples in the literature have been interpreted to the contrary. For these examples, we have simulated the local structures that form when folding begins by using the LINUS program with nonlocal interactions turned off. Our results support a hierarchic model of protein folding.

TWO CLASSES OF folding reactions are evident in small, single-domain proteins: class I proteins have observable folding intermediates; class II proteins do not. Are there therefore two different mechanisms of folding or does hierarchic folding explain both? Here we define hierarchic folding as a process in which folding begins with structures that are local in sequence and marginal in stability; these local structures interact to produce intermediates of ever-increasing complexity and grow, ultimately, into the native conformation. Non-hierarchic folding is a process in which tertiary interactions not only stabilize local structures but actually determine them. Hierarchic folding is an attractive model because it is both conceptually simple and computationally tractable.

A basic distinction between hierarchic and non-hierarchic folding is that local sequence information is sufficient (in principle) to predict the secondary structure of a native protein if folding is hierarchic but not if folding is non-hierarchic. In Part I of this article[1], we reviewed the evidence for the proposal that α-helices, β-hairpins and β-turns can be studied in peptides and that interactions that stabilize these structures are evident in their folded structures (determined either by X-ray crystallography or by NMR). Here, we examine the folding reaction in order to understand the structures and properties of observable intermediates in class I proteins and transition states in class II proteins.

There is a strong energetic rationale for believing that burial of hydrophobic side chains determines secondary structure[2], and this rationale is often used as the basis for arguments that folding is non-hierarchic. The free-energy change that accompanies formation of an isolated peptide helix is always small (−2.5 kcal mol^{-1} is a generous upper limit), whereas burial of a single phenylalanine side chain gives a comparable change in free energy[3]. In addition, the particular order of hydrophobic residues seems to be determinative. For example, hydrophobic/polar (H/P) patterning experiments that used peptides at an air–water interface (using air to mimic a nonpolar environment) have shown that the H/P pattern signals whether an α-helix or a β-sheet is formed[2]. In another example, screening of a library of quasi-random sequences modeled on four-helix bundles has shown that maintaining the correct H/P pattern, while choosing amino acids with favorable helix propensities, is sufficient to recover helical proteins that have the properties of molten globules[4].

Nevertheless, three lines of experimental evidence indicate that the folding process is hierarchic. First, helix-stop signals, which fix the boundaries of helices in proteins, are encoded in the local sequences that surround each helix terminus, not in residues that make tertiary interactions (see Part I of this article[1]). Second, many peptide fragments excised from proteins either form, or have a measurable tendency to form, the native fold in the absence of longer-range interactions (see Part I of this article[1]). Third, the structures of observed folding intermediates indicate that the latter form through a hierarchic folding process, and growing evidence suggests that, arguably, transition states in proteins can also be regarded as folding intermediates and possess structures that resemble those of observed intermediates. Here, we examine this third line of evidence.

Structures of folding intermediates

The best-characterized kinetic folding intermediates are those that can also be studied at equilibrium. In such cases, at acid pH, the native protein (N) becomes unstable relative to the unfolded form (U) but the folding intermediate (I) does not. Low-pH equilibrium intermediates occur because N, but not I, has some histidine, aspartate and/or glutamate residues that have unusually low pK_a values. At low pH, the equilibrium is pushed towards unfolding N preferentially, because in U and I, but not in N, these residues can be protonated. Two-dimensional-NMR hydrogen exchange, together with stopped-flow pulse-labeling measurements of exchange, has shown that, in the cases of apomyoglobin (apoMb)[5,6] and RNase H (Ref. 7), the two forms of I – the kinetic form and the acid form – are structurally equivalent. Interpretation of similar studies of ferricytochrome *c* (cyt *c*)[8,9] folding is complicated by the ease of formation of a non-native heme ligand at neutral pH (Ref. 10). Even so, the acid and native forms of cyt *c* are closely related in structure[9].

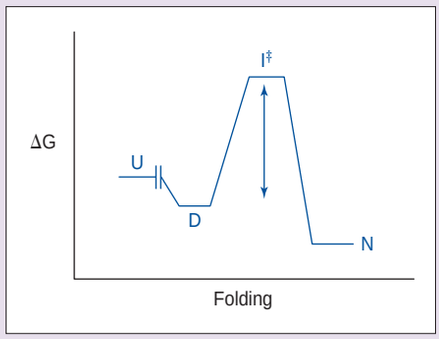
These folding intermediates provide clear evidence for hierarchic folding, because they have native secondary structure in the absence of persisting tertiary interactions: native secondary and supersecondary structures range from partial to complete in these intermediates, but the side chains are not fixed, the hydrophobic core residues remain partly solvated, and tertiary interactions between side chains are weak or absent. There are many well-characterized examples. The acid form of cyt *c* has all three major helices present in N (Ref. 9), whereas the pH 4 intermediate of apomyoglobin (apoMb; the form of myoglobin that lacks the

R. L. Baldwin is at the Dept of Biochemistry, Beckman Center, Stanford University Medical Center, School of Medicine, Stanford, CA 94305-5307, USA; and **G. D. Rose** is at the Dept of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, 725 N. Wolfe St, Baltimore, MD 21205-2185, USA.

0968–0004/99/$ – See front matter PII: S0968-0004(98)01345-0

Box 1. Transition states in protein folding

The figure (on-line, see Fig. I) shows a free-energy diagram of folding that is modeled on an ordinary chemical reaction. Alternative pathways of folding (not represented in the diagram) are usually accessible, but a single, minimal-free-energy path to the native protein (N) might predominate under particular conditions. The unfolded protein (U) is present in a strong denaturant (e.g. 6 M guanidinium chloride) before refolding is initiated. When shifted to refolding conditions, U undergoes very rapid partial folding or compaction. The rapidly formed species is referred to variously as either the denatured protein (D) or an early folding intermediate, depending on the extent of its structure and the degree to which it has been characterized. The free energy of any folding intermediates observable later (not shown) would lie between D and N. The transition-state species ($I^{\ddagger}$), at the top of the highest free-energy barrier, is not detectable.



If the free-energy barrier between D and $I^{\ddagger}$ is large enough (at least 5/2.RT, where R is the gas constant and T is the temperature) then, according to the transition-state approximation, D and $I^{\ddagger}$ equilibrate (approximately) prior to formation of N. Given these assumptions, the rate at which N is formed will be proportional to $[I^{\ddagger}]$, and the folding kinetics will follow an exponential time course. This folding rate can be written as $k_o^{-\Delta G^{\ddagger}/RT}$, where k_o is a prefactor rate constant whose size and meaning is still under discussion, and $\Delta G^{\ddagger} = [\Delta G^{I^{\ddagger}} - \Delta G^{D}]$. In mutagenesis studies, k_o is assumed to be unaffected by mutation, and therefore this term cancels out in the ratio of folding rates between mutant and wild-type species.

ϕ analysis has been used to quantify the extent to which a given side chain stabilizes the transition state, relative to the extent to which it stabilizes the native protein[37]. The ϕ value in the folding direction can be defined as follows[3]:

$$\phi_f = \frac{[\Delta G^{\ddagger}(WT) - \Delta G^{\ddagger}(i)]}{[\Delta G_{N\rightarrow D}(WT) - \Delta G_{N\rightarrow D}(i)]} \quad (1)$$

In the above expression WT and i refer to wild type and mutant, repectively, and $\Delta G_{N\rightarrow D}$ is the free-energy difference between N and D. A similar definition of ϕ_u is made in the unfolding direction, and both ϕ values are then obtained from measured folding and unfolding rates. ϕ_f and ϕ_u are complementary and sum to unity, provided that the transition state is the same under folding and unfolding conditions.

In mutant studies, a residue is often replaced by alanine because the latter has a minimal side chain, but ϕ values can also be measured for other replacements, and they depend on both the residue being replaced and the replacement. If a mutation does not change the stability of N, then the denominator in Equation 1 is zero and the ϕ value is indeterminate, but the ratio of folding rates of wild type and mutant species can still be determined.

When ϕ analysis of transition states was first introduced, it was expected that ϕ values of either 0 or 1 (corresponding to no interaction or full interaction in $I^{\ddagger}$) would be common; in fact, they are found rarely. In retrospect, this might not be surprising if transition states resemble observable intermediates that have molten-globule conformations and loosely packed side chains. For example, analysis of hydrophobic packing mutations[21] show that $\Delta\Delta G$ values of the pH 4 folding intermediate of apoMb are only about half the corresponding values for native apoMb.

The closely related Brønsted plot of $\Delta G^{\ddagger}$ versus $\Delta G_{N\rightarrow D}$ (which contains the arbitrary prefactor in $\Delta G^{\ddagger}$)[37] is used to detect groups of residues that cooperate in forming the transition state. If residues contribute to formation of $I^{\ddagger}$ in proportion to the extent that they stabilize N then their Brønsted plot is linear.

The position of the transition state is estimated from the m values (i.e. gradient) obtained from the folding and unfolding rate constants. For a two-state folding reaction, the plot of ln*k* against the denaturant concentration, [C], resembles a V: $\ln k_f$ decreases linearly with [C] (with a gradient m_f); $\ln k_u$ increases linearly with [C] (with a gradient m_u); and $\ln k_f$ and $\ln k_u$ intersect at the midpoint of the unfolding-transition curve. The extent of folding at $I^{\ddagger}$ is given by $m_f/(m_f - m_u)$. (Note that m_f and m_u have opposite signs.) This measure of the position of $I^{\ddagger}$ is commonly interpreted as the amount of surface area buried upon $D\rightarrow I^{\ddagger}$ folding (normalized by the total area buried in the $D\rightarrow N$ transition).

Certain mutations cause surprisingly large changes in the position of the transition state (see main text). This suggests that the diagram shown here, which is modeled on an ordinary chemical reaction, is not well suited to describing protein folding. A more accurate representation of such data would depict the transition-state barrier instead as being low and broad. Alternatively, mutations that cause large changes in the position of the transition state might do so by causing abrupt jumps between alternative folding pathways that display different transition states (which would invalidate assumptions in the analysis described above). A prime concern of current studies is to distinguish between these possibilities.

heme group) has only three stable helices (A, G and H)[5] – out of eight present in holomyoglobin (holoMb; the form of myoglobin that contains the heme group). Direct NMR spectroscopy of the apoMb intermediate[11] indicates that some helix content is detected in helices B, C and D, and confirms that helices A, G and H form. Comparisons of properties such as sensitivity to unfolding by guanidinium chloride[12] have demonstrated that the kinetic intermediate and an equilibrium intermediate of α-lactalbumin (α-LA) are very similar. NMR analysis of the urea-induced unfolding transition of the acid form of this protein indicates that the structures of the acid and native forms are related[13]. In the case of barnase, the transient folding intermediate is a discrete species that is formed cooperatively[14]; this suggests that it is native-like.

These results, which show that native-like secondary structures are present even though native tertiary interactions are weak or absent, provide the most-compelling evidence for hierarchic folding of class I proteins. To refute this evidence, one has to argue that these are not true folding intermediates, but are instead peptide-like local secondary structures that form rapidly in the unfolded protein but fail to function as building blocks for subsequent folding events. Indeed, some simulations[15–17] suggest that populated intermediates act as traps and that productive folding bypasses such traps on a fast track. Other, more recent, simulations[18] suggest the opposite, however: they imply that the folding process is channeled through particular intermediates, which are entropically favored and stable enough to become observable. A notable innovation in this recent work[18] is the development of a method for locating the transition state within a simulated folding trajectory by equating it with the point at which the folding and unfolding probabilities are equal. Operationally, that point is found by spawning multiple simulations at closely sampled intervals along the parent trajectory and identifying those where half fold and half unfold.

In practice, many informative simulations resort to highly simplified models of the energetics of folding. As a consequence, some known features of early stages in folding are omitted, such as cooperative helix formation. For this reason, other simulations based on more-complete physical models are of particular interest in assessing whether folding is hierarchic or non-hierarchic. For

example, Lazaridis and Karplus[19] recently performed molecular-dynamics simulations of chymotrypsin inhibitor 2 (CI2) unfolding, using high temperature (500K) to accelerate the process into the nanosecond range. From an examination of multiple trajectories (in 24 independent simulations), they could discern a preferred unfolding pathway, although it exhibited considerable variability. According to their simulations, unfolding is hierarchic: tertiary interactions break early, whereas secondary structures remain. If folding is the reverse of unfolding (experimentally, the conditions are different), then the simulations suggest that both the single helix and the β3–β4 strands of the sheet in CI2 form early.

Folding intermediates are coupled systems and are more stable than mere ensembles of independent, fluctuating helices. Protection factors of backbone amide protons in folding intermediates[5,7,9] demonstrate the latter's stability, as do direct measurements of the free energy of unfolding of the acid form of cyt *c* (Ref. 20) and the pH 4 intermediate of apoMb[21,22]. The highly cooperative folding behavior of the apoMb intermediate, which is two state under some conditions[21,22], provides conclusive evidence for the proposal that the intermediate is more than just a set of local helical structures. Cooperative folding is a hallmark of native proteins. A change made in one part affects the stability of the entire structure, which shows that different parts cooperate in the stabilization of the overall structure. This behavior would be surprising in an accidental, off-pathway folding intermediate.

A basic test for native-like structure in a folding intermediate is to truncate buried, nonpolar side chains that play critical roles in stabilizing the native structure, and to ask whether these side chains contribute significantly to the stability of the intermediate. If the intermediate unfolds in a two-state reaction – which is true of apomyoglobin[21,22] and which seems likely for cyt *c* (Ref. 20) – measuring the free energy of unfolding provides a quantitative answer. Kim and co-workers[23,24] have addressed this question in α-LA by measuring the effects of mutations on the stability of a specific disulfide bond. In all three proteins[20,21,23], nonpolar side chains that are important for the stability of the native protein contribute favorably to stabilization of the folding intermediate; this is most evident in the case of cyt *c* (Ref. 20).

A different genetically engineered derivative of α-LA that possesses two disulfide bonds is able to form the native disulfide pairing (one of three possible pairings) when present as a partly folded form (a molten globule) but not when present as the unfolded form in 6 M guanidinium chloride[23]. This result indicates that the molten-globule form of the α-LA derivative has the native tertiary fold. Taken together, these observations indicate that, in these cyt *c*, apoMb and α-LA folding intermediates, the helices interact with each other in a specific and native-like manner.

Note that the experiments reviewed above were conducted primarily in helical proteins or in helical regions of partly helical proteins. The status of β-sheet folding reactions has yet to be studied in similar detail.

Kinetic blocks and alternative pathways

What happens when a kinetic block prevents formation of the native protein? Different ways of imposing a block have been examined. For oligomeric and multidomain proteins, random mutagenesis has produced mutant proteins that fold under one particular condition (condition A) but not under another condition (condition B), although the already-folded form persists when transferred from condition A to condition B. The trimeric tailspike protein of phage P22 is an intensively studied example[25]. Mutants of this kind have yet to be obtained for small, single-domain proteins.

Chemical events involving bond isomerazation, such as the presence of an incorrect proline isomer in the unfolded protein, can also impose kinetic blocks. (Amino acid residues are either of two isomers – *cis* or *trans.* Usually, the *trans* isomer is found in globular proteins, except in the case of proline.) For example, in RNase A, the presence of the incorrect isomer of a particular proline residue prevents extensive folding in marginally native conditions, but folding proceeds in strongly native conditions and results in an intermediate that is both catalytically competent and surprisingly native-like[26,27]. Despite the kinetic block imposed by a wrong heme ligand in cyt *c*, the N- and C-terminal helices form[8] and maintain a key interaction during the folding process[28]. In both of these examples[10,26], when the kinetic block is released through a fluctuation, the folding reaction proceeds to the native structure (i.e. it does not return to the unfolded form to try again). These observations fit neatly with a hierarchic mechanism; unaffected parts of the protein fold and remain poised, and then progress forward towards native structure once the block is removed.

Most folding-process simulations predict that alternative folding pathways are available: studies of several proteins, including barstar[29] and hen lysozyme[30], have confirmed this prediction experimentally. The barstar unfolding results are particularly striking because intermediates that precede the rate-limiting step on each of two unfolding pathways have markedly different properties; therefore, the two transition states must be dissimilar. It is now widely accepted that folding routes are flexible and that competing pathways are available.

Transition states

Protein-folding reactions can be described by three concepts borrowed from chemical reactions: reaction pathway, intermediates and transition state. Folding is not an ordinary chemical reaction, however: no covalent bonds are made or broken; the Eyring rate equation (the fundamental equation of transition-state theory) does not apply; and the concepts of pathway, transition state and intermediates need to be scrutinized carefully. Often, there are ambiguities in terms such as folding pathway and transition state. For example, the term transition state implies a single, well-defined species, but a folding reaction might have a broad ensemble of transition-state species – as revealed in simulations[15–19]. Moreover, such species differ fundamentally from those that occur in ordinary chemical reactions, in which the rate of product formation depends upon the frequency of vibration of a critical bond.

Fortunately, the transition-state approximation is probably valid for protein folding, and it provides the necessary link between mutations, changes in stability, and changes in folding and unfolding rates (see Box 1). Munoz *et al.*[31] have used the transition-state approximation to examine the thermal unfolding of a β-hairpin. They fitted its unfolding kinetics to a statistical mechanical model that specifies the entire distribution of partly folded species as a function of the temperature and time after unfolding starts. The ensemble of transition-state species lies at the top of the free-energy barrier, and the authors estimate that 99% of the molecules follow the minimal-free-energy path. Comparisons of results generated by using the transition-state approximation with exact kinetic calculations for simple model reactions – such as the binding of a specific

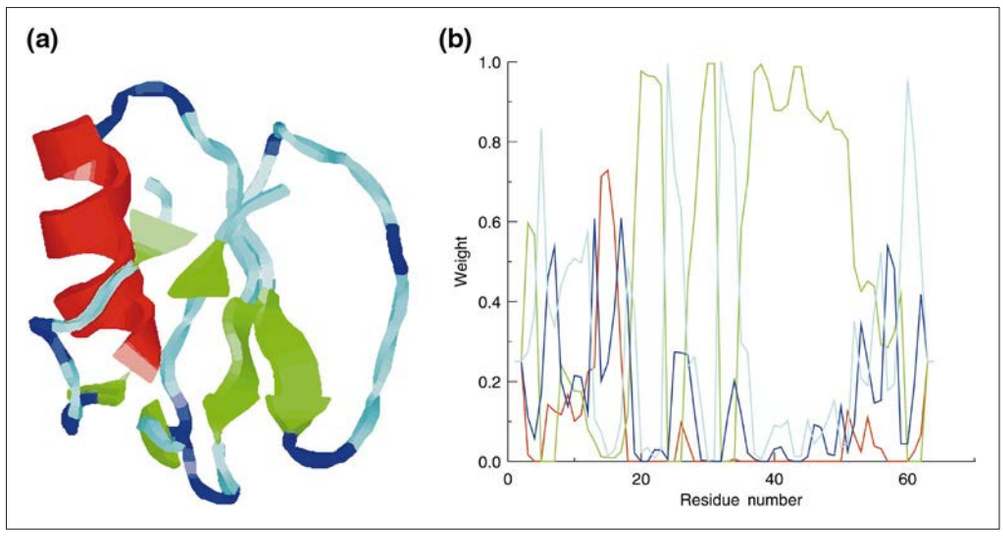


Figure 1

LINUS simulation of chymotrypsin inhibitor 2 (CI2)[40]. **(a)** Rasmol cartoon of the 3CI2 X-ray structure. **(b)** Conformational biases for CI2. Simulations reveal pronounced conformational biases that are distributed throughout the polypeptide chain. The simulation protocol assures that biases are a consequence solely of local interactions. These biases are the ensemble-averaged statistical weights for each residue in the four conformers: helix, strand, turn and coil. By definition, these weights sum to unity. The simulation was performed in three successive 1000-cycle stages[40]. Plots of both secondary structure and statistical biases are colored according to the following code: helix, red; strand, green; turn, blue; coil, cyan. At first, the plots of statistical bias appear overly complex. Each consists of four superimposed plots, which correspond to the four conformers: helix, strand, turn and coil. Segments in which one conformer is clearly dominant are conspicuous because one plot rises above the others in these cases. When no conformer is dominant, two or more plots can form a confusing tangle; in such cases, the reader should resist the urge to disentangle them. Instead, these sites should be interpreted as chain segments where the local LINUS-evolved weights do not resolve into a unique secondary-structure prediction. There is a discrepancy between secondary-structure definitions in the Rasmol cartoons, which are based on both backbone dihedral angles and hydrogen bonds, and the LINUS statistical biases, which are too local to capture hydrogen bonds between non-adjacent strands of sheet. Thus, in the Rasmol cartoons, β-sheet is shown in green but isolated strands are cyan, whereas in plots of statistical biases isolated strands are shown in green. (See the section on peptide folding in Part I for further explanation of this issue[1].) A pronounced bias towards helix is observed only in the region that corresponds to the actual helix; there is a peak at residue 16, one of three key residues in the transition state[37]. Helical bias diminishes abruptly in residues that correspond to the helix C-terminus (supplanted by turn/coil bias), in agreement with experimental data from ϕ analysis[37]. The remaining two key transition-state residues, 49 and 57, also correspond to sites that exhibit a pronounced local conformational bias.

ligand to a protein[32] or a simplified folding reaction in which every step has the same kinetic and equilibrium parameters[33] – increase confidence in the use of the approximation.

How similar are the transition states of different folding reactions, and do their structures resemble those of observable folding intermediates? The first point, on which there is general agreement, is that, unlike the transition-state barriers in ordinary chemical reactions (which are high and sharply peaked) those in folding are low and broad[34–36]. The fact that point mutations sometimes cause surprisingly large changes in the position (see Box 1) of the transition state supports this view. In the case of the Arc repressor[35], mutation causes the position to vary from 0.92 to 0.69 (in the refolding direction). The second point is that, in barnase (as yet the only example), the structure of the transition state is closely related to that of a preceding transient intermediate[14], as judged by their ϕ_f values, which give the interaction strength of each residue (either in the intermediate or in the transition state) relative to its strength in the native state. The ϕ_f values are similar in both species but somewhat higher in the transition state – which one would expect if the intermediate is 'on pathway' but the transition state is more completely folded. The third point is that far less information about secondary structure is available from transition states than is available from observable intermediates. This is because NMR hydrogen exchange, which reveals the stability of individual peptide hydrogen bonds in observable intermediates, cannot be used with transition states. Mutational analyses of transition states provide information chiefly about side-chain interactions, not about secondary structure.

A final point is that the handful of transition states that are well characterized by mutational studies are not yet sufficient to invite generalization. Barnase[14] is discussed above. CI2 (Ref. 37) and the Arc repressor[35] show linear Brønsted plots that include all residues and, therefore, every residue affects the transition state in proportion to the extent it affects the native protein. The strength of the side-chain interactions in the CI2 transition state is less than a third of the strength of those in the native state of the protein; a similar picture is evident for the Arc repressor. However, the SH3 domains of SRC (Ref. 38) and those of α-spectrin[39] show sharply polarized transition states; folded regions represent only a subset of the native structure in these cases, and they include some comparatively stronger side-chain interactions.

Mutational characterization of a transition state yields a snapshot of one stage in the folding process. If both a helix and a tertiary cluster of side-chain interactions are evident, the structure of the transition state does not reveal which forms first or whether both form simultaneously. From current studies of transition states, we can conclude only that it is plausible, but unproven, that class II folding reactions proceed by a hierarchic mechanism.

As noted above, a molecular dynamics simulation of CI2 unfolding[19] indicates that the process is hierarchic. The results of repeated trajectories show great variability and indicate that there is a broad ensemble of transition-state structures. In future work of this kind, it will be interesting to test the validity of the transition-state approximation.

To explore the question of hierarchy, CI2 was simulated by using LINUS (see Fig. 1) with non-local interactions suppressed[40]. Figure 1b shows a plot of the context-induced conformational biases. Only one segment has a strong helix propensity, and it corresponds to the sole helix in the three-dimensional

structure (Fig. 1a), although helical statistical weights tail off towards the C-terminus of the actual helix. The correspondence between strand/turn propensities and actual strands/turns is also high. Strand weights (which reflect the chain's propensity to be extended) dominate in the region of the large loop. As described above, these biases are induced largely by local steric interactions, but they anticipate the actual secondary structure rather closely. Moreover, the biases cannot be promoted by formation of a tertiary nucleus, which is proscribed by the simulation protocol used here. The simulation demonstrates that an extensive structural framework of CI2 is built into the local amino acid sequence and that this framework can be realized in the absence of nonlocal interactions. Of course, stability sufficient for experimental detection might require the framework to be fortified by tertiary interactions; such fortification would be analogous to the stabilization of intermediates that can be observed directly in class I proteins.

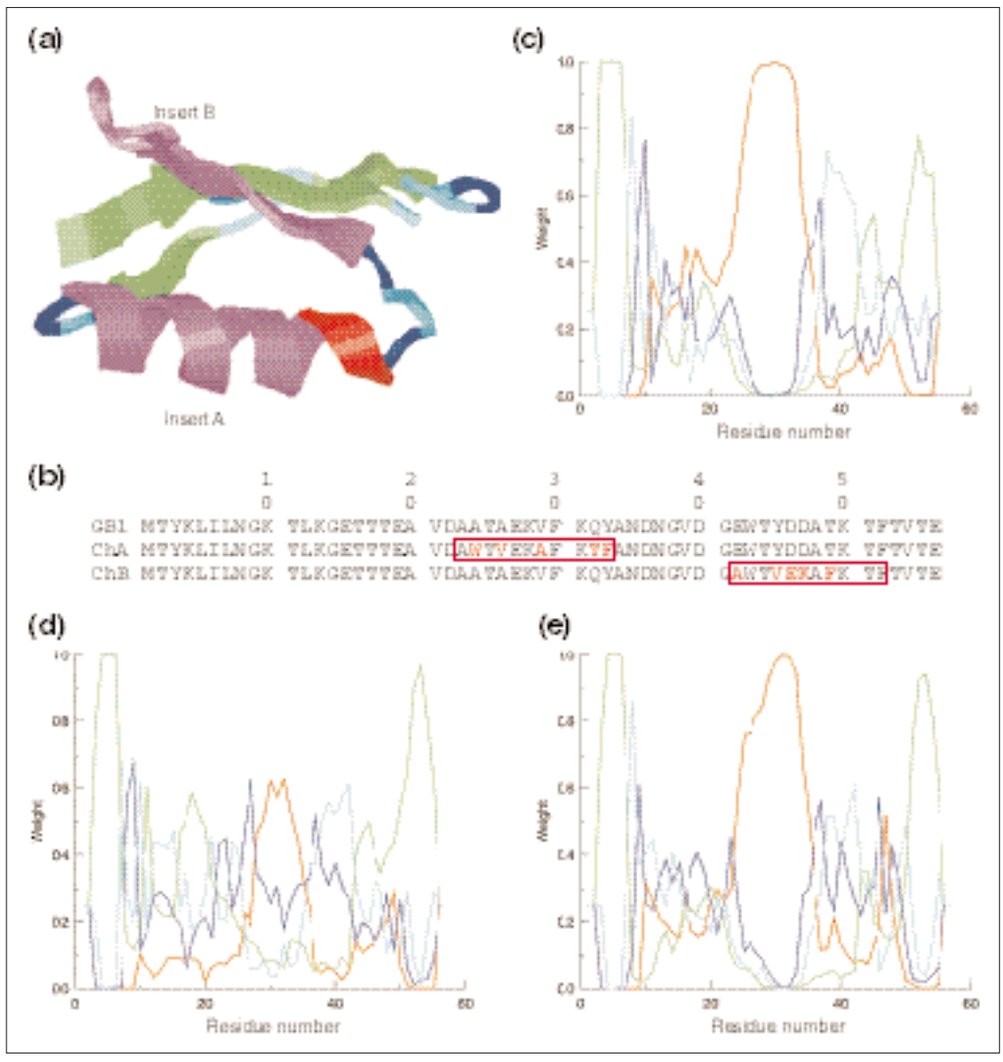


Figure 2

LINUS simulations[40] of protein G (1GB1) and the two chameleon variants[42]. Even in the absence of nonlocal interactions, pronounced conformational biases are evident throughout the polypeptide chain. Plots of both secondary structure and statistical biases are colored according to the following code: helix, red; strand, green; turn, blue; coil, cyan. **(a)** Rasmol cartoon of the X-ray structure of native protein G (1PGB). Each 11-residue chameleon insert is shown in magenta. Insert A is within the central helix (residues 23–35); insert B (residues 42–52) includes the penultimate strand, the last turn and part of the last strand. **(b)** Sequence alignment of native protein G and the two chameleon variants (ChA and ChB). In each variant, the 11-residue chameleon sequence is boxed, and mutated residues are indicated. Note that each variant differs from the parent sequence at five positions, not 11. **(c)** Conformational biases for native protein G. **(d)** Conformational biases for ChA. **(e)** Conformational biases for ChB. Biases range from 0 to 1. In each case, a pronounced bias towards helix is observed only in the segment that corresponds to the actual helix in the NMR structure. Within this segment, turn bias is increased in ChA, but strand bias remains negligible; in fact, the substitution of a helical turn by a β-turn at the helix N-terminus is not inconsistent with experimental data[42]. In the native sequence, only the first and last strands have high strand weights, and these persist in both variants.

Tests of hierarchic folding

The observation that segments of identical sequence can adopt different conformations in different proteins[41] challenges the hypothesis that secondary structure in proteins is determined by local interactions. This observation has been interpreted, by some, to mean that the formation of native secondary structure is a consequence of tertiary interactions, particularly the hydrophobic burial of side chains.

A provocative experiment reported by Minor and Kim[42] provides an extreme example. The authors devised an 11-residue sequence – dubbed the chameleon sequence – which they substituted at either of two sites (A and B) in protein G. The 56-residue protein comprises a central helix (residues 23–35) and a four-stranded β-sheet (see Fig. 2a). Site A (residues 23–33) lies within the helix; site B (residues 42–52) overlaps the penultimate β-strand [actually a strand, a turn and part of another strand (which are evident from the X-ray structure); see Fig. 2b]. When situated at site A, the chameleon sequence adopts a helical conformation; however, when situated at site B, it adopts a strand-turn-strand conformation (Fig. 2a). Thus, the chameleon sequence is aptly named. The conformation it assumes appears to be determined by its context within the total protein, not by local interactions.

The chameleon experiment was simulated by using LINUS (see Fig. 2c–e) with long-range interactions suppressed in order to disclose local effects on conformational bias. In the context of the entire protein, at site A the chameleon sequence retains high helix weights (Fig. 2d), whereas at site B the same sequence exhibits negligible helix weights (Fig. 2e). Thus, interactions that are local but extend beyond the boundaries of the chameleon sequence itself are

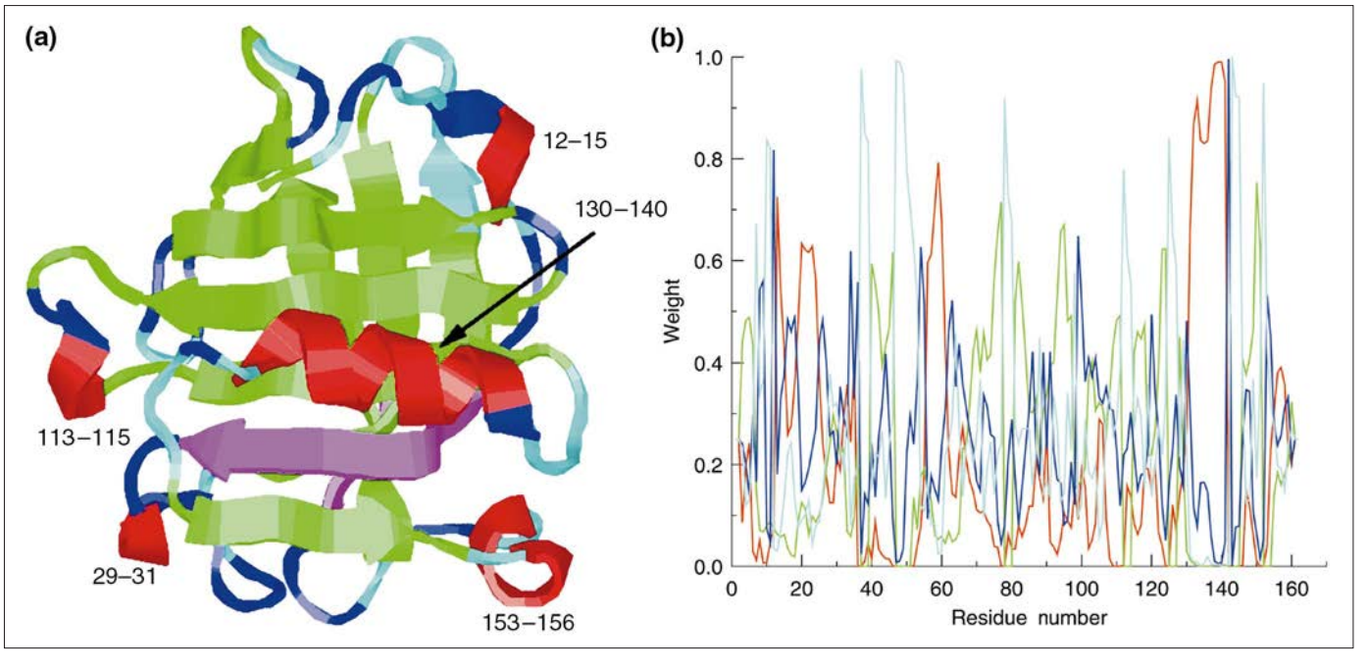


Figure 3

LINUS simulation of β-lactoglobulin (β-LG)[40]. Even in the absence of non-local interactions, pronounced conformational biases are seen throughout the polypeptide chain. Plots of both secondary structure and statistical biases are colored according to the following code: helix, red; strand, green; turn, blue; coil, cyan. **(a)** Rasmol cartoon of the β-LG X-ray structure (1BEB). The two magenta regions have high helix weights, but they are not helical in the X-ray structure. **(b)** Conformational biases for β-LG. Helical structure is of particular interest[44]. Five helical regions are evident in the X-ray structure: a long α-helix (residues 130–140) and four helical turns (residues 12–15, 29–31, 113–115 and 153–156). Four correspond to regions that exhibit high helix weights (residues 131–141, 13–15, 30–33 and 154–158). Also, a modest helix weight is apparent near the remaining region (residues 112–114). Additional high helix weights are evident in two non-helical regions [residues 19–27 and 58–63; shown in magenta in (a)]. Whether these regions correspond to the 'excess helix' predicted by circular dichroism[44] is not yet known. Nine regions exhibit high strand weights (residues 3–6, 40–46, 53–57, 66–77, 81–85, 92–104, 108–111, 115–124 and 146–151). With the exception of residue 115, all of these residues are in extended regions of the structure, many in actual β-strands. Conversely, most – but not all – of the β-strands observed in the X-ray structure are covered by these nine regions. The only exceptions are a single remaining strand and part of another, which have high helix weights. Finally, regions that exhibit high turn weights (residues 8–9, 12–13, 17–19, 27–28, 31–36, 62–64, 79–80, 85–86, 89–91, 99–100, 106–107, 112–113, 130, 142 and 159–160), except for residue 91 and residues 106–107, all correspond to actual reverse turns or helical turns in the X–ray structure.

sufficient to account for the observed position-dependent differences in conformational preference of this 11-residue sequence.

β-lactoglobulin (β-LG) is another provocative example. The protein is a 162-residue 'clam' that contains two opposing slabs of antiparallel β-sheet and a single 11-residue helix. Circular dichroism (CD) studies indicate that a compact early folding intermediate possesses non-native helical structure[43,44]. According to its CD spectrum, this burst phase (i.e. within the dead time of the measurement) intermediate, which is formed upon refolding in 3 M urea, contains 34 ± 15 helical residues[44], whose locations are not yet known.

Again, we have simulated β-LG, using LINUS with all long-range interactions suppressed. Local conformational biases reflect the actual secondary structure closely but not perfectly (see Fig. 3). The 'excess helix' is particularly interesting[44]. Examination of the X-ray structure (Fig. 3a) determined by Brownlow *et al.* (1BEB in the Protein Data Bank) reveals that the protein contains four short helices, in addition to the single long α-helix. Conformational biases (Fig. 3b) capture all five native helical regions and two additional non-native regions (residues 19–27 and 58–63). In total, 38 residues have high helix weights; this value is in good agreement with the value of 34 ± 15 residues estimated by CD (Ref. 44). The LINUS simulation predicts that residues 19–27 (a strand of β-sheet in the native structure) and 58–63 (a partial strand and adjacent turn) are the main non-native contributors to the helical CD spectrum.

Tendamistat, a 74-residue protein that possesses two disulfide bonds, provides yet another β-sheet example. Characterization of a partly folded equilibrium form[45] shows that its structure is consistent with hierarchic folding but contains some non-native helical structure (≥25%, according to the CD spectrum). This species forms at pH 2–3 in the presence of 3–6 M trifluoroethanol, which probably strengthens peptide hydrogen bonds. NMR data for nuclear Overhauser effects and protected peptide-NH protons show that a major part of the native β-sheet structure is present. However, the near-UV CD spectrum and chemical-shift dispersion indicate that the side chains remain flexible. The helical segments probably occur in loops of the native structure and are not very stable, given that no new protected NH protons were detected. In a hierarchic mechanism, folding begins by forming regions of native secondary structure. The existence of isolated, non-native secondary structure in other regions need not invalidate the hierarchic mechanism – unless a non-native component interacts productively with other local structures, and higher-order folding intermediates are evolved.

To test whether a unique nucleus dominates the rate-limiting step in folding of the α-spectrin SH3 domain, Viguera *et al.*[46] made two circularly permuted constructs, and determined the X-ray structures of the wild-type protein and both constructs. They then made

eight point mutations in each of the three proteins, and measured their folding and unfolding kinetics. The overall fold is conserved in all cases, but the residues with the largest ϕ_u values change, which indicates that the structure of the transition state changes upon circular permutation. The authors concluded that the transition state does not have a unique tertiary nucleus, although later experiments[33] lead them to doubt this conclusion and to believe, instead, that there are unexpected subtleties to finding the structure of a transition state from ϕ values.

Additional sources of complexity in folding patterns can arise during late folding stages in multidomain and oligomeric proteins, after some level of early organization has already been established. For example, the dimeric Trp repressor forms a monomeric helical intermediate early in folding, and the structure of the native protein shows that the two helical monomers must become intertwined as folding proceeds[47]. Issues such as this bear on the question of whether folding is hierarchic or nonhierarchic, but emerge late in the folding process and are beyond the scope of this review.

The information necessary for folding is highly dispersed

Mutagenesis studies support the inference that conformational specificity and folding stability are decoupled[48]. Mutations often reduce stability, but only rarely do they alter the overall fold. Sufficiently destabilizing mutations result in unfolding, not alternative folding. Recent results of wholesale mutagenesis of some proteins[14,35,37] bolster this view. In an extreme example, in which Dalal and co-workers[49] deliberately altered the tertiary fold of a protein, they had to change ~50% of the sequence. Protein folding is robust because the conformation is overdetermined by information spread throughout the sequence. Mutagenesis experiments (including those performed by nature) and LINUS simulations (Figs 1c, 2b and 3b) concur in this conclusion. In our view, the hierarchic mechanism, in which folding begins with local structures, is rooted in the dispersal of folding information throughout the polypeptide chain.

Concluding comments

Controversy over whether class I folding reactions are hierarchic has centered on whether the intermediates are an integral part of the mainstream folding process or an isolated offshoot in which further folding is arrested. Mounting evidence favors hierarchic folding: helices are native-like, and residues from the hydrophobic core are partly desolvated. In more-recent work, native-like tertiary properties, such as highly cooperative folding, the presence of a native tertiary fold, and the existence of stabilizing, native-like packing interactions, have been found as well. Taken together, these examples argue strongly that the intermediates are produced by the authentic folding process.

Characterization of class II folding reactions, which lack intermediates, necessarily is limited to the structures and properties of their transition states. Evidence from several studies shows that the position of the transition state is variable, and it can be moved by mutations. This fact argues for an incremental assembly process and is consistent with some recent folding simulations.

We suggest that the difference between class I and II folding reactions lies not in the mechanism of their folding but only in the stability of their intermediates.

Acknowledgements

We thank our co-workers, past and present, for their discussions of these issues, and we apologize to our colleagues whose contributions could not be cited because of the limit on references. We thank the NIH for support (grant GM 19988 to R. L. B. and grant GM 29458 to G. D. R.).

References

1 Baldwin, R. L. and Rose, G. D. (1999) *Trends Biochem Sci.* 24, 26–33
2 Kaiser, E. T. and Kézdy, F. J. (1984) *Science* 223, 249–255
3 Nozaki, Y. and Tanford, C. (1971) *J. Biol. Chem.* 246, 2211–2217
4 Kamtekar, S. *et al.* (1993) *Science* 262, 1680–1685
5 Hughson, F. M., Wright, P. E. and Baldwin, R. L. (1990) *Science* 249, 1544–1548
6 Jennings, P. A. and Wright, P. E. (1993) *Science* 262, 892–896
7 Chamberlain, A. K. and Marqusee, S. (1997) *Structure* 5, 859–863
8 Roder, H., Elöve, G. A. and Englander, S. W. (1988) *Nature* 335, 700–704
9 Jeng, M-F. *et al.* (1990) *Biochemistry* 29, 10433–10437
10 Yeh, S-R. and Rousseau, D. L. (1998) *Nat. Struct. Biol.* 5, 222–228
11 Eliezer, D., Yao, J., Dyson, H. J. and Wright, P. E. (1998) *Nat. Struct. Biol.* 5, 148–155
12 Ikeguchi, M., Kuwajima, K., Mitani, M. and Sugai, S. (1986) *Biochemistry* 25, 6965–6972
13 Schulman, B. A., Kim, P. S., Dobson, C. M. and Redfield, C. (1997) *Nat. Struct. Biol.* 4, 630–634
14 Dalby, P. A., Oliveberg, M. and Fersht, A. R. (1998) *J. Mol. Biol.* 276, 625–646
15 Bryngelson, J. D., Onuchic, J. N., Socci, N. D. and Wolynes, P. G. (1995) *Protein Struct. Funct. Genet.* 21, 167–195
16 Dill, K. A. and Chan, H. S. (1997) *Nat. Struct. Biol.* 4, 10–19
17 Wolynes, P. G. (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94, 6170–6175
18 Pande, V. S., Grosberg, A. Y., Tanaka, T. and Rokhsar, D. S. (1998) *Curr. Opin. Struct. Biol.* 8, 68–79
19 Lazaridis, T. and Karplus, M. (1997) *Science* 278, 1928–1931
20 Marmorino, J. L., Lehti, M. and Pielak, G. J. (1998) *J. Mol. Biol.* 275, 379–388
21 Kay, M. S. and Baldwin, R. L. (1996) *Nat. Struct. Biol.* 3, 439–445
22 Luo, Y., Kay, M. S. and Baldwin, R. L. (1997) *Nat. Struct. Biol.* 4, 925–930
23 Wu, L. and Kim, P. S. (1998) *J. Mol. Biol.* 280, 175–182
24 Peng, Z. and Kim, P. S. (1994) *Biochemistry* 33, 2136–2141
25 Mitraki, A., Danner, M., King, J. and Seckler, R. (1993) *J. Biol. Chem.* 268, 20071–20075
26 Cook, K. H., Schmid, F. X. and Baldwin, R. L. (1979) *Proc. Natl. Acad. Sci. U. S. A.* 76, 6157–6161
27 Schmid, F. X. and Blaschek, H. (1981) *Eur. J. Biochem.* 114, 111–117
28 Colón, W. and Roder, H. (1996) *Nat. Struct. Biol.* 3, 1019–1025
29 Zaidi, F. N., Nath, U. and Udgaonkar, J. B. (1997) *Nat. Struct. Biol.* 4, 1016–1024
30 Wildegger, G. and Kiefhaber, T. (1997) *J. Mol. Biol.* 270, 294–304
31 Munoz, V., Thompson, P. A., Hofrichter, J. and Eaton, W. A. (1997) *Nature* 390, 196–199
32 Hill, T. L. (1976) *Proc. Natl. Acad. Sci. U. S. A.* 73, 679–683
33 Doyle, R., Simons, K., Qian, H. and Baker, D. (1997) *Protein Struct. Funct. Genet.* 29, 282–291
34 Matouschek, A. and Fersht, A. R. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 7814–7818
35 Milla, M. E., Brown, B. M., Waldburger, C. D. and Sauer, R. T. (1995) *Biochemistry* 34, 13914–13919
36 Oliveberg, M., Tan, Y-J., Silow, M. and Fersht, A. R. (1998) *J. Mol. Biol.* 277, 933–943
37 Itzhaki, L. S., Otzen, D. E. and Fersht, A. R. (1995) *J. Mol. Biol.* 254, 260–288
38 Grantcharova, V. P., Riddle, D. S., Santiago, J. V. and Baker, D. (1998) *Nat. Struct. Biol.* 5, 714–720
39 Martinez, J. C., Pisabarro, M. T. and Serrano, L. (1998) *Nat. Struct. Biol.* 5, 721–729
40 Srinivasan, R. and Rose, G. D. (1995) *Protein Struct. Funct. Genet.* 22, 81–99
41 Kabsch, W. and Sander, C. (1984) *Proc. Natl. Acad. Sci. U. S. A.* 81, 1075–1078
42 Minor, D. L., Jr and Kim, P. S. (1996) *Nature* 380, 730–734
43 Kuwajima, K. *et al.* (1987) *FEBS Lett.* 221, 115–118
44 Hamada, D., Segawa, S. and Goto, Y. (1996) *Nat. Struct. Biol.* 3, 868–873
45 Schönbrunner, N. *et al.* (1996) *J. Mol. Biol.* 260, 432–445
46 Viguera, A. R., Serrano, L. and Wilmanns, M. (1996) *Nat. Struct. Biol.* 3, 874–880
47 Shao, X. and Matthews, C. R. (1998) *Biochemistry* 37, 7850–7858
48 Lattman, E. E. and Rose, G. D. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 439–441
49 Dalal, S., Balasubramanian, S. and Regan, L. (1997) *Nat. Struct. Biol.* 4, 548–552