The fundamentals of protein folding: bringing together theory and experiment

Christopher M Dobson* and Martin Karplus†

Experimental and theoretical studies together are providing insights into the mechanism by which proteins fold. Our present knowledge of the essential aspects of the folding reaction is outlined and some approaches, both theoretical and experimental, that are being developed to obtain a more detailed understanding of this complex process are described.

Addresses
*Oxford Centre for Molecular Sciences, New Chemistry Laboratory, University of Oxford, South Parks Road, Oxford, OX1 3QT, UK; e-mail: chris.dobson@chem.ox.ac.uk
†Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA and Laboratoire de Chimie Biophysique, Institut Le Bel, Universite Louis Pasteur, 67000 Strasbourg, France; e-mail: marci@tammy.harvard.edu

Introduction

There are nearly 100,000 protein sequences in the human genome. To become biologically active, the large majority of these must fold to a unique stable structure. There are at least 1000 fundamentally different folds adopted by natural proteins and many variants within these. The question of how individual protein sequences efficiently and reliably achieve their native state following synthesis on the ribosome is one of the most intriguing problems in structural biology. In a cell, folding takes place within a complex environment containing high concentrations of a wide variety of molecules and ions. It is well established that many factors are associated with the cellular folding process, including molecular chaperones and folding catalysts. Our present knowledge of these molecules is discussed in the parallel 'overview' article by Ellis and Hartl (pp 102–110). The various factors are involved in a wide range of control and localisation processes, but do not provide conformational information for the polypeptide chains with which they interact. The evidence gathered over many years supports the fundamental principle, formulated initially by Anfinsen and others, that the code for folding resides within the amino acid sequence [1].

The fundamental question is, therefore, how the sequence codes for the fold. Two features of proteins make this question particularly intriguing. First, since the mainchains of all proteins have an identical composition, how do the sidechains dictate the overall fold? Second, since the number of possible conformations of a polypeptide chain is astronomically large, how does a given sequence find its specific native structure in a finite time? The latter problem has come to be known as the Levinthal paradox and has dominated discussions of folding for 30 years [2]. One for-mulation of the problem is as follows. If a protein is made up of a polypeptide chain of 100 residues and we assume there are only two possible configurations for each residue, there are of about 10^{30} possible conformations. If only 10^{-11} s is required to convert one configuration into another, a random search of all conformations would require some 10^{19} s or 10^{11} year. As the doubling time of bacteria can be less than 30 min, it is clear that evolution has found an effective solution to this combinatorial problem.

The initial suggestions as to the nature of this solution involved the proposal that there exist specific pathways for folding [3]. On these pathways, the protein molecules would pass through well-defined partially structured states, some of which could be transient, but others would be populated significantly. The folding mechanism was thus envisaged as being similar to the reactions of small molecules. If these pathways were specific enough, only a small region of conformational space would be sampled and the Levinthal paradox would thereby be avoided. This view was apparently supported by the fact that experiments provide clear evidence for the existence of partially folded intermediates formed both during folding and under partially denaturing conditions. Recent experimental studies paint a more complex picture of folding, in which the behaviour of different proteins often appears quite distinct. The folding of some proteins is known to involve well-defined compact intermediates, whilst the folding of others has been found to be effectively a two-state reaction. A rather different ('new') view of protein folding has emerged recently from a combination of theoretical and experimental approaches, as we describe in the next section.

## Energy surfaces and energy landscapes

The 'new view' [2,4–6,7*] is based on a description of protein folding in terms of statistical ensembles and emphasises the differences between the folding reaction, in which the reorganisation of a very large number of weak noncovalent interactions determines the outcome, and a small molecule reaction, in which one or, at most, a small number of strong covalent bonds are broken or made. A major distinguishing feature of protein folding is the extreme heterogeneity of the reaction and the complex interplay between the entropic and enthalpic contributions to the free energy of the system during the reaction. From an experimental point of view, some aspects of this heterogeneity are very apparent. The starting point of folding in the laboratory is a protein unfolded in a denaturant such as urea or guanidinium chloride. It is well established that the denatured protein usually resembles a 'random coil', in which local interactions dominate the conformational behaviour [8]. As a result,

the denatured state is extremely heterogeneous, both globally and at the level of individual residues. Indeed, the number of accessible conformers may approach the number of possible conformations discussed above in the context of the Levinthal paradox. Given that any individual folding experiment will involve at most about 10^{18} molecules, every molecule in the solution being studied is likely to have a different conformation in the random-coil state. The folding process requires that all of these different conformations convert to the native state, the state of lowest energy by the thermodynamic hypothesis of protein stability [1].

An important aspect of the 'new view' of protein folding is that it provides a simple way of understanding why the Levinthal paradox is not a real problem. As there is a sizeable difference between the enthalpies of the denatured and the folded state (on the order of 30–100 kcal mol^{-1}), all that is needed for finding the latter is that this enthalpy difference provides a sufficient bias of the conformational space to avoid the need to search through an impossibly large number of configurations. This aspect of the 'new view' has been supported by the results of calculations on models that are simple enough to treat the process of folding in detail with current computational power and yet are sufficiently complex to include the key elements of the folding process, such as the existence of a Levinthal paradox. One such model, often referred to as a 'toy protein' because one can 'play' with it, represents the residues as beads positioned on an infinite cubic grid (the 'lattice'), with only nearest neighbour interactions that depend on the nature of the residues involved. The folding reaction is simulated by starting with a denatured (random coil) chain and making local Monte Carlo moves of the residues on the lattice until the native (lowest energy) configuration is reached.

Although the form of the potential and the use of a lattice leads to a highly oversimplified model, the results of such simulations have provided important information on possible protein folding scenarios [2,4–6,7•,9]. From simulations of a 27-mer [9] whose native state is a $3 \times 3 \times 3$ cube, effective free energy surfaces, or 'free energy landscapes', have been determined. One example of such a surface is illustrated in Figure 1. This surface represents the free energy of the system, including implicit solvation, as a function of the coordinates ('progress variables') chosen to describe the essential features of the folding reaction. Energy surfaces play an essential role in any reaction, whether that of small molecules or of proteins, but the complexity of the protein folding reaction, involving thousands of degrees of freedom versus three to ten or so for small molecule reactions, means that an immense 'compression' of the trajectory details is necessary to obtain a useful description [7•]. The progress variables used in Figure 1 are the parameter Q_0, which corresponds to the fraction of native state contacts, and the parameter C, which corresponds to the total number of contacts,
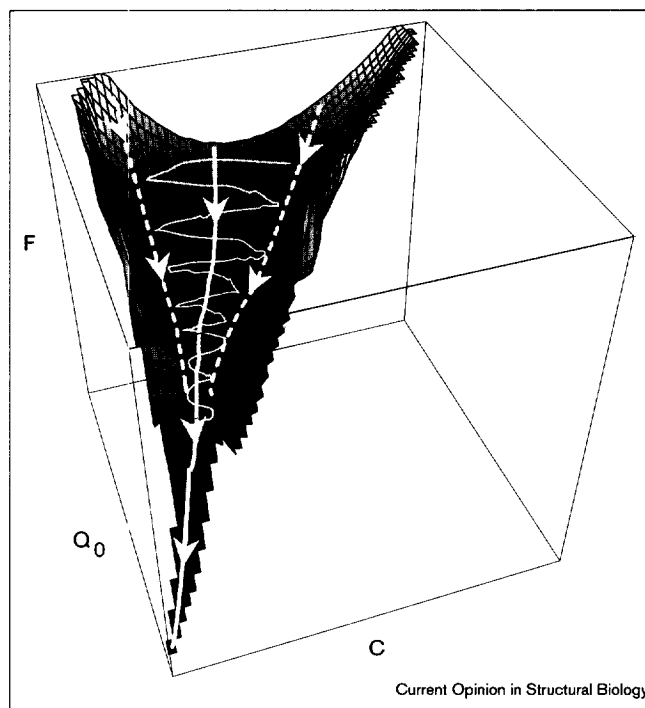
Free energy (F) surface of a protein at a temperature T (T < T$_m$) of a fast-folding 27-mer as a function of the fraction of native contacts (Q$_0$) and the total number of (native and non-native) contacts (C). The thin solid line shows a trajectory traced by the last structure at each value of Q$_0$ [C(Q$_0$)] in a typical Monte Carlo trial; the bold solid line shows the average trajectory for 1000 trials [<C(Q$_0$)>] and the dashed lines show a range of two standard deviations around the average. If the distribution of trajectories is close to a Gaussian in C for each value of Q$_0$, then the dashed lines would include about 95% of the trajectories.

native or otherwise, for each conformation. Q_0 varies from 0 to 1; the value 1 corresponding to the 28 contacts present in the native state.

The surface shown in Figure 1 illustrates that at the beginning of the folding reaction ($Q_0 \leq 0.2$), there are many conformations of similar free energy, so that the accessible surface is very broad. As folding progresses under conditions in which the native state is stable (i.e. below T_m in the absence of denaturants), the energy of the system decreases with the formation of native contacts that are generally more stabilising than non-native ones. In addition, the number of accessible conformations decreases as the energy difference between those with and those without favourable interactions increases. Thus, the entropy decreases as the native state is approached. It is the importance of the balance between the decrease in energy and entropy that makes it essential to consider the free energy surface shown in Figure 1, instead of the energy surfaces generally used for small molecule reactions, where entropic effects play a much less important role. The accessible free energy surface shown in Figure 1 has a 'funnel-like'

shape that guides the system along increasing Q_0 and C as it progresses towards the unique (lowest energy) native conformation. The ranges of the calculated trajectories shown on the surface illustrate the heterogeneity of the ensemble of structures sampled during the folding process.

As suggested by the 'toy protein' calculations, it is apparent from experimental studies that intermediate species detected during folding and unfolding are highly heterogeneous. The most intensively discussed of these are called 'molten globules' [10], which generally are relatively close to the native state. In these globules, which have been observed for many proteins under both equilibrium and nonequilibrium conditions, considerable native-like character can exist in terms of secondary structure and the overall fold, but there is generally extensive disorder in the sidechains and the global structural fluctuations are much greater than those of the native state. Apparently the overall fold of a protein can exist in the absence of close packing, that is, without the formation of many of the specific atomic interactions that stabilise the unique native structure [11]. This suggests, in accord with protein design experiments [12], that only the general nature of the amino acid residues (e.g. their hydrophobicity and electrostatic character) and their distribution in the sequence is involved in folding from the astronomical number of structures in the random-coil state to a restricted region of conformational space in the vicinity of the native fold. In the cases in which molten globules are populated, the final transition to the native state is often the slowest step in folding [10]. The Levinthal paradox has been resolved early in the folding process, however, and the rate is determined by barriers to the reorganisation involved in forming the native structure. By contrast, for fast folding proteins without intermediates, the search for a core or nucleus that involves only a small fraction of the native contacts is likely to be the rate-determining step; once the core is formed, folding to the native state is fast [13]. Although these patterns of folding behaviour appear to be very different, they are aspects of the same overall mechanism, modulated by a different balance of energetic and entropic contributions to the free energy of folding [7•].

A unified mechanism of folding

An outline mechanism of protein folding can be developed by considering the free energy surfaces for the reaction, such as that shown in Figure 1. It provides immediate insight into how the Levinthal paradox is overcome. Although the accessible conformation space is large, an individual molecule, as shown in the folding trajectories, samples only a small portion of this space. Each folding trajectory is different, depending both on the starting point (i.e. which of the 'random-coil' conformations a given molecule occupies at the initiation of folding) and on the stochastic nature of the folding process. The ensemble of molecules thus covers the accessible surface shown in Figure 1. It is important to note that the bias in the free energy surface already mentioned makes the accessible

surface much smaller than the full surface envisioned in the Levinthal paradox. The diagram makes clear the fundamental difference between protein folding and small molecule reactions, in which the trajectories of individual molecules on a surface dominated by strong interactions tend to be similar and stay in the neighbourhood of the minimum energy path.

Another suggestion from 'toy protein' simulations is that the overall folding behaviour (i.e. that which would be observed experimentally) can be changed drastically by relatively small changes in the model parameters. Many of the complexities of experiments have been simulated by changing the balance of forces in the simulations [14,15], particularly those that vary the relative importance of the entropy and the enthalpy. As already mentioned, fast two-state folding can occur when collapse involves only a small subset of highly stabilising native contacts in a core region or nucleus [16,17], so that the entropic penalty for reaching that state is not too great. In this context, it has been stated [18] that the distributions of local and nonlocal contacts are important factors in determining the ease with which proteins fold, although their exact role may vary depending on the protein architecture. In fact, more extensive data than those used in the original correlation suggest that, although helical proteins tend to fold faster than β-sheet proteins, within a class of proteins wide variations not related to contact numbers occur. Also, it has been found in some lattice simulations of larger proteins that long-range contacts are important, but in others that a mixture of short-range contacts for initiation and long-range contacts for cooperativity lead to efficient folding [19]. It has been proposed that these core residues are particularly important for defining both the protein fold and the folding rate and, hence, that they have been conserved during evolution [16,17,20]. One is beginning to find general features of folding that might constitute 'rules' by which apparently disparate experimental observations for different structures can be brought together in a unified mechanism.

If there is more uniform (hydrophobic) attraction between residues, rapid collapse to a disorganised globule may occur, with the slow step in folding corresponding to reorganisational events within a compact ensemble of states. Simulations with larger lattices indicate that such collapse processes are more likely to occur as a protein increases in size [19]. Moreover, the formation of a core or nucleus in larger systems may occur independently in different regions of the structure, resulting in additional complexities in folding, including the formation of partially structured intermediates and the possibility of extreme heterogeneity in the folding kinetics of a population of molecules [21]. Such predictions correlate in general terms with the findings of experimental studies of larger proteins [22]. The macroscopically different folding behaviour observed experimentally can thus be encompassed in a single

overall microscopic description (a unified mechanism of folding) when one takes account the delicate balance among the large number of weak interactions that are involved in the folding process [7•].

The ensemble view of folding involves the existence of considerable heterogeneity at all but the final stages of folding and there is now convincing evidence that this is an important aspect of the interpretation of experimental measurements [8]. The concept that the transition state for a folding reaction involves a distribution of structures, rather than a single well-defined structure as in a small molecule reaction, is in accord with recent interpretations of protein engineering experiments [13,23,24]. These experiments provide a unique insight into the role of individual residues in the transition states and have provided crucial evidence for the existence and nature of folding nuclei [13]. Recent findings show that the transition states of some proteins, such as the helical lambda repressor protein, can be altered substantially by mutations [25,26]. This is consistent with the idea of rather flat free energy surfaces resulting from the interplay of many weak interactions and the maintenance of a balance between entropic and energetic contributions over most of the folding reaction before reaching the native state, which corresponds to a deep energy well with minimal entropy. Studies of all-β SH3 domains, however, indicate that some interactions may be obligatory to off-set the entropic cost of folding [27,28]. Moreover, these studies suggest that the locations of the folding nucleus in the structures of different SH3 domains are similar, even when their sequences are not conserved. Drastic mutations, such as circular permutations, can, however, reveal alternative transition states that may themselves have little conformational variability [27]. It may well be that there are significant differences in the requirements for a well-defined nucleus in proteins with different topological features; for example, helical proteins may have much greater plasticity in this regard than sheet proteins [23,27,28]

Recent experimental developments
The 'new view' of protein folding leads to a unified mechanism that incorporates the features established from a wide range of experimental studies of folding and places them in the context of an overall theoretical framework. The challenge now is to define the specific folding events for individual proteins and to determine how these (and the overall protein fold) are encoded in the amino acid sequence. In the new view, this is phrased as the question of how the energy surface or 'landscape' is determined by the interactions of the amino acid residues in different portions of the polypeptide chain. Answering this question requires the development of improved experimental and theoretical methods for determining the nature of the conformational ensembles that contribute at various stages of the folding reaction. From such analyses, it is likely that an interpretation that integrates many of the earlier conceptual models of protein folding will be developed.

One aspect of such a procedure corresponds to mapping out the free energy surface for the folding of individual proteins, given that dynamical effects are not dominant in determining the distributions. Most of the present experiments are on the millisecond timescale or slower, so that the available observations of folding are largely limited to the stage after the Levinthal paradox has been 'solved'. It is, therefore, crucial to obtain information about the earlier stages of folding by improving the time resolution. It is equally important to increase the spatial resolution of experiments, so as to be able to define interactions at the level of individual residue contacts rather than just at a global level. Moreover, because of the heterogeneity of the ensemble of folding trajectories, it is essential to be able to define the behaviour of the various members within an ensemble, that is, to explore the nature of the folding trajectories of individual molecules. As much of this information will be difficult to obtain from experimental data alone, a parallel development is required in simulation techniques and other theoretical approaches, so as to be able to obtain a realistic, detailed description of the folding reaction of specific proteins. Thus, it is a combination of experimental and theoretical approaches that will serve to provide the information necessary for a full understanding of protein folding.

Experimental methods are currently being developed to probe the early events in folding by using fast mixing devices or by initiating the reactions *in situ* in a variety of ways [29•]. This is particularly important because, as discussed above, the generation of a native-like fold from a highly disordered denatured ensemble may be very fast (on the millisecond timescale or less). It is also the part of the folding reaction in which the statistical nature of the new view is most important and in which the multiplicity of routes inherent in the solution of the Levinthal paradox is expected to be most evident. Important advances in this area include rapid mixing devices which utilise innovative designs to reduce substantially the millisecond dead times of conventional stopped and quenched flow methodologies [30,31], and temperature jump techniques, which are able, in principle, to probe events on timescales as short as picoseconds. Of particular interest here are experiments that are initiated from cold denatured states and that generate refolding conditions by rapid heating, particularly using laser techniques [32,33].

Other important developments are methods for initiating reactions without perturbing the solution conditions. These include optical triggering of photochemical processes [34] and electron injection techniques, which change the stability of redox proteins by changes in oxidation state [35]. These various methods are beginning to reveal events in protein folding on timescales of 10–100 μs that are likely to be associated with not only the formation of local structure but also with global collapse [36,37]. In isolated peptides, the formation of secondary structural motifs, including α helices and β hairpins, has been seen

on timescales as short as 100 ns [38•]. At present, few structural techniques have been used to follow such events, the most common being fluorescence, and it is important that other methods are brought to bear on these fast processes [29•,39] Of particular interest will be a determination of the extent of the deviations from simple kinetic behaviour that arise from the heterogeneity of the unfolded ensembles. This will provide information that can be related to theoretical results on the rates of equilibration among structures sampled along the reaction paths, relative to the progress towards the native state [40].

The methods necessary to study most aspects of the kinetic steps in protein folding are limited in their ability to determine the specific structural changes that occur during a folding reaction. By combining the results of complementary techniques, however, it is feasible to begin to define the process in much more detail than is possible from any individual measurement [41]. The technique that is most appropriate to extend such studies to higher resolution in a spatial sense is NMR spectroscopy and a number of recent studies have highlighted the potential of this approach [42,43]. Three strategies have been particularly important. The first involves the study of the folding and unfolding process under equilibrium conditions. The ability of NMR spectroscopy to probe the interconversion of molecular species by their effects on nuclear relaxation processes is the foundation of this method [25]. One of the advantages of this approach is that very fast processes can be studied and the limitations of strategies in which the folding process must be initiated to generate a nonequilibrium reactive state are avoided.

A second approach is to investigate stable analogues of kinetic intermediates [42]. This area of study has proved to be very productive and a recent study of myoglobin illustrates the wealth of detail that can be obtained, particularly about the secondary structure present at different stages of folding [44]. It has proved very difficult, however, to determine any information about the tertiary interactions. Such information has emerged in related studies in which the persistent structure in partially folded states has been perturbed systematically by progressive folding in chemical denaturants [45]. This has revealed regions of local cooperativity in α-lactalbumin and has allowed the stability of different interactions in maintaining the native-like fold to be measured. An additional and important development has been the use of paramagnetic spin-label probes in NMR experiments for mapping the structural distributions in compact denatured states [46]. This method utilises the same basic principles as those established for native structures, but exploits the larger spectral perturbations of paramagnetic systems to probe the much more disordered ensembles inherent in non-native states. Application of this approach to staphylococcal nuclease confirms that in its compact denatured state its topology is native-like in the absence of tight packing. The third general approach involves the application of NMR spectroscopy in real time

to follow the folding reaction [43]. A particularly interesting example showing the power of NMR in probing structural transitions at the level of individual residues has involved the study of the refolding of a photoactive protein following an unfolding step in its photocycle [47]. As NMR experiments develop in resolution and sensitivity, they will offer a very powerful general approach to the detailed mapping of the energy surfaces of folding.

The trajectories of individual molecules on the energy surfaces are expected to be very different from each other, particularly in the early stages of the reaction. The development of techniques capable of detecting the behaviour of individual molecules can therefore play a fundamental role in characterising the distributions of molecular properties at different stages of the folding reaction. Although such methods have yielded only very preliminary information on folding so far, interesting results are being obtained by the use of atomic force microscopy and laser tweezer experiments to unfold protein constructs that are composed of a small number of molecular domains [48,49]. These data, when combined with simulations of the unfolding process, provide information concerning the forces involved along the unfolding pathway [50,51].

Recent theoretical developments

In concert with these continuing developments in experimental methods, there have been advances in theoretical approaches designed to permit more realistic simulations of protein folding reactions, that is simulations that can help to define what actually happens with specific proteins, rather than establishing possible folding scenarios. These go beyond the lattice simulations that have played such an important role up to this point (for reviews, see [14,52]). One approach makes use of off-lattice models (i.e. models in which the polypeptide chain moves in continuous space rather than on a lattice), but keeps a residue-based description with simplified interactions. A second approach uses all-atom protein models with explicit or implicit solvent to study the folding thermodynamics and the unfolding dynamics of specific proteins. Finally, simulations attempting to fold peptide fragments or entire proteins are beginning to play a role. In what follows, we describe some of this recent work. We select the references that exemplify the relation either between different theoretical approaches or between theoretical approaches and experiment.

All-atom simulations, both with implicit and explicit solvent, are becoming more focused on specific problems. A recent study has examined the high temperature unfolding of the small protein CI2 with an all-atom representation of the protein and an implicit description of the solvent [24]. Use of the latter speeded up the simulations so that a sufficient number of trajectories could be calculated to obtain meaningful statistics concerning the ensembles involved in the reaction. The protein CI2 was selected because the results could be validated by

comparison with unfolding simulations in explicit solvent [53,54] and could be compared with detailed protein engineering experiments of the type mentioned above [55,56]. It was concluded that the transition state occurred early, with only 25% of the native contacts, and that the transition-state ensemble had contributing structures with root mean square deviations as large as 15 Å. Nevertheless, when the trajectories are analysed in terms of native-like contacts, there is a statistically preferred folding pathway in which the helix and a two-stranded β sheet are the essential elements of a folding core. A strong preference for a certain order of events, determined by the amino acid sequence, is thus compatible with a funnel-like (single basin of attraction) average energy surface. It should be noted, however, that, as indicated by lattice simulations, the high temperature sampling may emphasise the portion of the surface that has rapidly decreasing energies to counterbalance the large entropy loss on folding. A combined all-atom simulation in explicit solvent and protein engineering study has provided additional information on the folding energy surface for CI2 [57].

Barnase, one of the first proteins that was studied by all-atom high temperature unfolding simulations in explicit solvent [58], has been investigated recently in a solvent composed of water and urea molecules, the latter at approximately 8 M concentration ([59]; A Caflish, M Karplus, unpublished data). Although urea has been used for many years to denature proteins, the molecular mechanism of urea-induced unfolding is not understood. The simulation results (A Caflish, M Karplus, unpublished data) demonstrate that an aqueous urea solution leads to better solvation of a polypeptide chain than pure water. Urea molecules interact more favourably with nonpolar groups than water and the presence of urea improves the interactions of water molecules with the hydrophobic groups of proteins. These results indicate that urea denaturation involves effects on both nonpolar and polar groups in proteins. More experimental data, particularly on the structural aspects of the interactions of urea with proteins and polypeptides, such as those available from X-ray [60] and NMR [61] studies, are clearly needed to refine our understanding of this complex process.

Another approach is to use molecular dynamics simulations of all-atom models with explicit solvent to map out the free energy surface for the protein folding reaction. The method has been applied to a three-helix bundle fragment of the *Staphylococcus aureus* protein A [62,63] and to the small α/β protein G [64]. For the three-helix bundle protein, the equilibrium surface was interpreted as involving an initial collapse, with the formation of 50 to 70% of the helical content, but only about 30% of the native tertiary structure. Although no explicit comparison with experiment can be made in this case, these values are comparable to estimates of these parameters for a range of small proteins [63]. The study of protein A is of particular interest

because it can be compared with kinetic simulations with a new type of off-lattice model. By use of a square well contact potential and discrete molecular dynamics, it has become possible to do extensive simulations of folding thermodynamics and kinetics in three-dimensional space, that is, several hundred trajectories have been run for a series of models of the same three-helix bundle protein ([65]; Y Zhou, M Karplus, unpublished data). Although the surface has a simple funnel-like shape, in correspondence with that calculated previously [62,63], the simulations of the kinetics have shown that there are alternative folding pathways, some of which involve metastable intermediates. This illustrates the importance of being able to carry out both equilibrium and kinetic simulations to obtain a more complete description of the protein folding reaction. Moreover, depending on a parameter determining the stability of the native state, the folding mechanism changes from one that corresponds to the diffusion-collision model [66] to one that is dominantly a random collapse with a subsequent search for the native state.

An important area that is receiving renewed emphasis, both experimentally and theoretically, is concerned with peptide fragments that have marginally stable structures corresponding to either α helices or β hairpins. The simplicity of such systems has made it possible to trigger their folding reactions and analyse the folding kinetics in some detail [29•,38•]. Although the kinetics of the folding of such peptides have not yet been studied by molecular dynamics or Monte Carlo simulations, a number of recent simulations have been concerned with the folding thermodynamics of peptides. The most extensive of these deal with the artificial β peptides of six or seven residues that form stable helices in methanol [67]. Simulations for 30 ns of the peptides with an explicit solvent model resulted in several folding and unfolding transitions from which thermodynamic parameters for the transition could be estimated. An alternative approach, which uses an implicit solvent model to speed up the calculations and adaptive umbrella sampling to improve the coverage of the conformational space, has been used in similar studies of a 13 residue peptide that forms an α helix in solution and a 12 residue peptide designed to form a stable β hairpin [68]. Excellent agreement with experiment was found for the stability of both peptides. Also, of considerable interest is the result that misfolded conformations (e.g. a β hairpin for the α helix) occur with significant probability. A new method, called 'self guided' molecular dynamics, appears to improve the motion through conformational space and has been applied to folding in a vacuum [69].

Folding a protein on a computer with a full-atom model in explicit solvent has been termed the 'holy grail' of the protein folding problem [70]. Recently, a 1 μs trajectory, a considerable extension of previous simulation times, was reported for the protein villin, a 36 residue, three-helix bundle [71]. Starting with a partly folded conformation (i.e. it included the correct turn topology), the two main

Figure 2
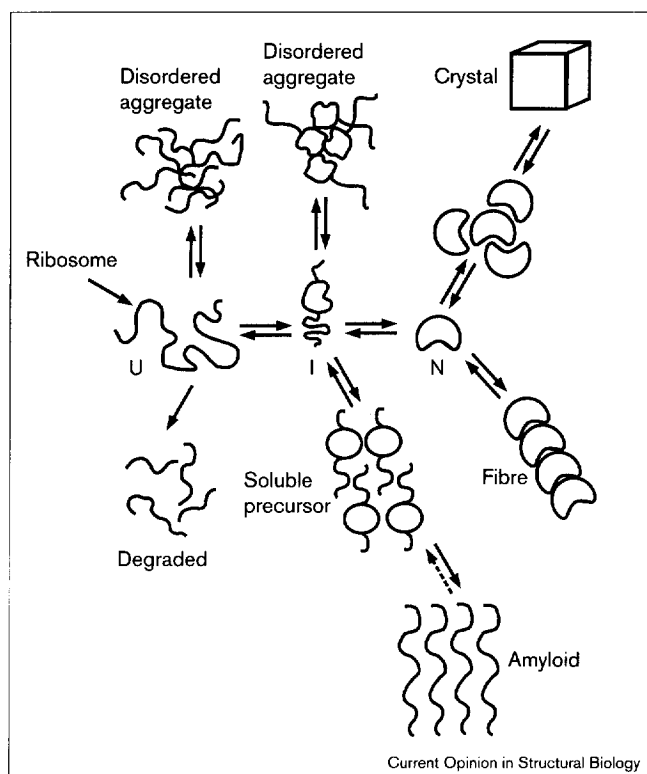
Schematic representation of some of the states accessible to a polypeptide chain following its biosynthesis. In its monomeric state, the protein is assumed to fold from its highly disordered unfolded state (U) through a partially structured intermediate (I) to a globular native state (N). The native state can form aggregated species, the most ordered of which is a three-dimensional crystal, whilst preserving its overall structure. The unfolded and partially folded states can form aggregated species that are frequently highly disordered, but amyloid fibrils can form through a nucleation and growth mechanism. There is evidence that this process occurs most readily from partially folded intermediate states of proteins.

helices formed in part, although not with the relative positions corresponding to the native state. A metastable 'intermediate' that lasted for 150 ns formed early, but there was no sign that the trajectory was approaching the native state in the remainder of the simulation. Clearly, if a single trajectory did lead to the native state, it would be of considerable interest as a 'tour de force', but a minimum of 20 or so such trajectories will be needed to obtain meaningful information for analysis [24]. Like the 'real' holy grail, many aspects of folding are still shrouded in mystery and complementary approaches, some of which are described above, are essential to obtain a detailed understanding of the events occurring during the folding of even a simple protein. As discussed in an analysis of the 'new view' [2], understanding the folding of specific proteins is likely to make use of the phenomenological models introduced earlier to suggest ways of circumventing the Levinthal paradox. One of these, the diffusion-collision model [66], has recently been implemented with the AGADIR

program [72] to estimate the stability of individual helices and applied successfully to explain the effect of mutations on the folding rate of the λ repressor [26].

The multiple states of proteins in biology

As our understanding of the principles of folding has increased, so has the realisation that folding and unfolding play an important role in the mechanism and control of a broad spectrum of cellular processes. These range from the translocation of proteins across membranes to their appropriate compartments, to the regulation of events in the cell cycle. The failure of such control processes can lead to cellular malfunctions and to disease [73]. It is important, therefore, to recognise that proteins in biological environments can exist in a variety of different states and that the state of a given protein under particular conditions depends on a complex series of thermodynamic and kinetic factors. This is illustrated in a highly schematic manner in Figure 2 [74]. Protein folding not only generates a biologically active structure, but also protects the protein from degradation by proteases and reduces the probability that aggregation will occur. Unfolding of a native protein exposes it to such possibilities. The process of folding and unfolding is, in some cases, directly coupled with function. Examples of this include a number of proteins associated with the regulation of protein synthesis and protein–protein interactions [75], and proteins, such as titin, that are involved in muscle action [76]. Titin is a highly elongated protein that consists of a number of domains, some of which may be unfolded under extreme tension (see above); the refolding of titin appears to be the driving force for muscle contraction after stress.

One of the fundamental problems of protein folding in the cell is the high concentration of material in the medium that can lead to aggregation prior to folding (see Ellis and Hartl, pp 102–110). Minimising aggregation is undoubtedly one of the major roles of the molecular chaperones within cellular compartments and it is important that more information is collected about the fundamental nature of protein aggregation. An important characteristic of aggregation is that it is usually very slow and it frequently requires specific nucleation processes [77]. Thus, the ability of proteins to fold rapidly is an important evolutionary development that can minimise competition with aggregation. In a bacterial cell, for example, fewer than 15% of proteins interact with GroEL [78] and it is likely that those that do are the ones that fold slowly through particularly aggregation-prone intermediates. Even after the intrinsic folding process has been completed, aggregation again becomes a possibility if proteins find themselves in an environment in which they are unfolded for prolonged periods of time. This is likely to be a critical feature of the family of diseases associated with the appearance of amyloid fibrils and plaques; they include the prion-associated spongiform encephalopathies and Alzheimer's disease [79•].

In accord with this view, recent *in vitro* experiments suggest that the ability to form amyloid fibrils is not limited to a small range of disease-related proteins, but could be a generic property of polypeptide chains [80]. Regardless of the configurational tendencies of an isolated polypeptide chain, intermolecular interactions favour β-sheet structures, at least some of which are highly ordered [81*]. In many cases, the conversion to amyloid structures is accompanied by a helix to sheet conversion and the fundamental nature of the fibrils from different proteins associated with clinical amyloidoses appears to be similar [82]. This can be attributed to the fact that the formation of fibrils is associated with the polypeptide mainchain, which is common to all proteins. The discovery that simple proteins can be converted under carefully controlled laboratory conditions into amyloid fibrils with the characteristics of those observed in disease states [80,83], is beginning to produce glimpses of the structures of such fibrils at a molecular level and their relationship to the conformations of globular proteins [84].

That ordered intermolecular aggregates are not seen in general in living systems is, at least in part, a consequence of the cooperativity of protein folding. This prevents the existence, under normal conditions, of significant concentrations of partially unfolded proteins, which would have a tendency to associate. Biology has generated a means of satisfying the intrinsic bonding capabilities of a polypeptide chain with intramolecular instead of intermolecular interactions. The significantly hydrophilic character of native proteins surfaces, involving a number of charged sidechains, then acts to inhibit aggregation. Aggregation of proteins to fibrils does not necessarily require a particular 'alternate' conformation of a soluble protein. The formation of aggregated proteins in diseases may simply be the result of aberrant conditions that lead to unfolding in an environment appropriate to nucleation and growth of well-ordered structures [83]. Seeding by preformed aggregates appears to be an important aspect of the rapid formation of such aggregate structures. This may be a key feature of the prion associated diseases, in which the nature of infectivity is of paramount significance [85,86].

Conclusions

The field of protein folding is at a stage at which many of the fundamental issues are becoming clear, at least in outline. The level of detail in our understanding is not yet sufficient to allow us to predict how specific sequences will fold or to design new sequences that fold to stable proteins of defined architecture. But progress in these directions is being made and the establishment of the fundamental rules for folding by experimental and theoretical means is an essential part of the efforts towards these goals. The mechanism of folding, as we have discussed, requires a balance to be preserved during the folding reaction between a very large number of weak interactions. It has evolved so that almost all the molecules are able to find their lowest energy state and avoid

stable minima that can act as traps to slow the folding process or to encourage aggregation. The difficulty in simulating the folding of a given sequence and, indeed, in predicting the structure that corresponds to its lowest energy state is a result of the complexity of the energetic and entropic balance involved. The multiplicity of states at a molecular level is compounded substantially when intermolecular interactions are considered. An important challenge for the future is to find ways in which these additional complexities can be explored through both theory and experiment.

Acknowledgement

References and recommended reading

In this article, we have tried to give the background to our present understanding of the mechanism of protein folding and to indicate some important areas of current research. We indicate below several review articles published during the past year that will enable the reader to explore specific topics mentioned here in greater detail.

1. Anfinsen CB: **Principles that govern the folding of protein chains.** *Science* 1973, **181**:223-230.

2. Karplus M: **The Levinthal paradox: yesterday and today.** *Fold Des* 1997, **2**:S69-S75.

3. Levinthal C: **Are there pathways for protein folding?** *J Chim Phys* 1968, **65**:44-45.

4. Baldwin RL: **Matching speed and stability.** *Nature* 1994, **369**:183-184.

5. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG: **Funnels, pathways and the energy landscape of protein folding: a synthesis.** *Proteins* 1995, **21**:167-195.

6. Dill KA, Chan HS: **From Levinthal to pathways to funnels.** *Nat Struct Biol* 1997, **4**:10-19.

7. Dobson CM, Šali A, Karplus M: **Protein folding: a perspective from**
•   **theory and experiment.** *Angew Chem Int Ed* 1998, **37**:868-893.
    A comprehensive article that discusses the nature of the protein folding 'reaction' and brings together the results of a wide range of theoretical and experimental studies in terms of a unified mechanism of protein folding.

8. Smith LJ, Fiebig K, Schwalbe H, Dobson CM: **The concept of a random coil. Residual structure in peptides and denatured proteins.** *Fold Des* 1996, **1**:95-106.

9. Šali A, Shakhnovich E, Karplus M: **Kinetics of protein folding: a lattice model study of the requirements for folding to the native state.** *J Mol Biol* 1994, **235**:1614-1636.

10. Ptitsyn OB: **Molten globule and protein folding.** *Adv Protein Chem* 1995, **47**:83-229.

11. Wu LC, Peng Z-Y, Kim PS: **Bipartite structure of the α-lactalbumin molten globule.** *Nat Struct Biol* 1995, **2**:281-286.

12. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH: **Protein design by binary patterning of polar and non-polar amino acids.** *Science* 1993, **262**:1680-1685.

13. Fersht AR: **Nucleation mechanisms in protein folding.** *Curr Opin Struct Biol* 1997, **7**:3-9.

14. Shakhnovich E: **Theoretical studies of protein-folding thermodynamics and kinetics.** *Curr Opin Struct Biol* 1997, **7**:29-40.

15. Chan HS, Dill KA: **Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics.** *Proteins* 1998, **30**:2-33.

16. Shakhnovich E, Abkevich V, Ptitsyn O: **Conserved residues and the mechanism of protein folding.** *Nature* 1996, **379**:96-98.

17. Mirny LA Abkevich VI, Shakhnovich E: **How evolution makes proteins fold quickly.** *Proc Natl Acad Sci USA* 1998, **95**:4976-4981.

18. Plaxco KW, Simons KT, Baker D: **Contact order, transition state placement and the refolding rates of single domain proteins.** *J Mol Biol* 1998, **277**:985-994.

19. Dinner A, Šali A, Karplus M: **The folding mechanism of larger proteins.** *Proc Natl Acad Sci USA* 1996, **93**:8356-8361.

20. Ptitsyn OB: **Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes?** *J Mol Biol* 1998, **278**:655-666.

21. Dinner A, Karplus M: **The thermodynamics and kinetics of protein folding. A lattice model analysis of multiple pathways with intermediates.** *J Mol Biol* 1999, in press.

22. Dobson CM, Evans PA, Radford SE: **Understanding protein folding - the lysozyme story so far.** *Trends Biol Sci* 1994, **19**:31-37.

23. Onuchic JN, Socci ND, Luthey-Schulten Z, Wolynes PG: **Protein folding funnels: the nature of the transition state ensemble.** *Fold Des* 1996, **1**:441-450.

24. Lazaridis T, Karplus M: **"New view" of protein folding reconciled with the old through multiple unfolding trajectories.** *Science* 1997, **278**:1928-1931.

25. Burton RE, Huang GS, Daugherty MA, Calderone TL, Oas TG: **The energy landscape of a fast-folding protein mapped by Ala→Gly substitutions.** *Nat Struct Biol* 1997, **4**:305-310.

26. Burton RE, Meyers JK, Oas TG: **Protein folding dynamics: quantitative comparison between theory and experiment.** *Biochemistry* 1998, **37**:5337-5343.

27. Martinez JC, Pisabarro MT, Serrano L: **Obligatory steps in protein folding and the conformational diversity of the transition state.** *Nat Struct Biol* 1998, **5**:721-729.

28. Grantcharova VP, Riddle DS, Santiago JV, Baker D: **Important role of hydrogen bonds in the structurally polarised transition state for folding of the src SH3 domain.** *Nat Struct Biol* 1998, **5**:714-720.

29. Callender RH, Dyer RB, Gilmanshin R, Woodruff WH: **Fast events in**
• **protein folding.** *Annu Rev Phys Chem* 1998, **49**:173-202.
An in-depth review of new approaches for monitoring fast events in protein folding (down to picoseconds) and the results that have been obtained for proteins and peptides.

30. Shastry MCR, Luck SD, Roder H: **A continuous–flow capillary mixer to monitor reactions on the microsecond timescale.** *Biophys J* 1998, **74**:2714-2721.

31. Bökenkamp D, Desai A, Yang X, Tai Y-C, Marziuff EM, Mayo SL: **Microfabricated silicon mixers for submillisecond quench-flow analysis.** *Anal Chem* 1998, **70**:232-236.

32. Nölting B, Golbik R, Fersht AR: **Submillisecond events in protein folding.** *Proc Natl Acad Sci USA* 1995, **92**:10668-10672.

33. Gruebele M, Sabelko J, Ballew R, Ervin J: **Laser temperature jump induced protein refolding.** *Accounts Chem Res* 1998, **31**:699-707.

34. Jones CM, Henry ER, Hu Y, Chan C-K, Luck SD, Bhuyan A, Roder H, Hofrichter J, Eaton WA: **Fast events in protein folding initiated by nanosecond laser photolysis.** *Proc Natl Acad Sci USA* 1993, **90**:11860-11864.

35. Telford JR, Wittung-Stafshede P, Gray HB, Winkler JR: **Protein folding triggered by electron transfer.** *Accounts Chem Res* 1998, **31**:755-763.

36. Hagen SJ, Hofrichter J, Szabo A, Eaton WA: **Diffusion-limited contact formation in unfolded cytochrome c: estimating the maximum rate of protein folding.** *Proc Natl Acad Sci USA* 1996, **93**:11615-11617.

37. Shastry MCR, Roder H: **Evidence for barrier-limited protein folding kinetics on the microsecond time scale.** *Nat Struct Biol* 1998, **5**:385-392.

38. Eaton WA, Muñoz V, Thompson PA, Henry ER, Hofrichter J: **Kinetics**
• **and dynamics of loops, α-helices, β-hairpins and fast-folding proteins.** *Accounts Chem Res* 1998, **31**:745-753.
A survey of the results of applying fast reaction methods to the study of the folding of peptides and proteins, and the conclusions that can be drawn about the timescales of the formation of different fundamental elements of structure. Other articles in the same issue of Accounts of Chemical Research review a range of applications of biophysical techniques to the study of folding.

39. Eaton WA, Thomson PA, Chan CK, Hagen SJ, Hofrichter J: **Fast events in protein folding.** *Structure* 1996, **4**:1133-1139.

40. Karplus M: **Protein reactions and conformational change: simple or complex or both?** In *Simplicity and Complexity in Proteins and Nucleic Acids.* Edited by Frauenfelder H, Deisenhofer J, Wolynes PG. Berlin: Dahlem University Press; 1999:in press.

41. Plaxco KW, Dobson CM: **Time-resolved biophysical methods in the study of protein folding.** *Curr Opin Struct Biol* 1996, **6**:630-636.

42. Dyson HJ, Wright PE: **Equilibrium NMR studies of unfolded and partially folded proteins.** *Nat Struct Biol* 1998, **5**:499-503.

43. Dobson CM, Hore PJ: **Kinetic studies of protein folding using NMR spectroscopy.** *Nat Struct Biol* 1998, **5**:504-507.

44. Eliezer D, Yao J, Dyson HJ, Wright PE: **Structural and dynamic characterisation of partially folded states of apomyoglobin and implications for protein folding.** *Nat Struct Biol* 1998, **5**:148-155.

45. Schulman B, Kim PS, Dobson CM, Redfield C: **A residue specific NMR view of the non-cooperative unfolding of a molten globule.** *Nat Struct Biol* 1997, **4**:630-634.

46. Gillespie JR, Shortle D: **Characterisation of long-range structure in the denatured state of staphylococcal nuclease. II. Distance restraints from paramagnetic relaxation and calculation of an ensemble of structures.** *J Mol Biol* 1997, **268**:170-184.

47. Rubinstenn G, Vuister GW, Mulder FAA, Düx PE, Boelens R, Hellingwerf KJ, Kaptein R: **Structural and dynamic changes of photoactive yellow protein during its photocycle in solution.** *Nat Struct Biol* 1998, **5**:568-570.

48. Kellenmeyer MZS, Smith SB, Granzier HL, Bustamante C: **Folding-unfolding transition in single titin molecules characterized with laser tweezers.** *Science* 1997, **276**:1112-1116.

49. Rief M, Fernandez JM, Gaub HE: **Elastically coupled two-level systems as a model for biopolymer extensibility.** *Phys Rev Lett* 1998, **81**:4764-4767.

50. Lu H, Isralewitz B, Krammer A, Vogel V, Schulten K: **Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation.** *Biophys J* 1998, **75**:662-671.

51. Paci E, Karplus M: **Unfolding of fibronectin type 3 modules by biased molecular dynamics simulation.** *J Mol Biol* 1999, in press.

52. Karplus M, Šali A: **Theoretical studies of protein folding and unfolding.** *Curr Opin Struct Biol* 1995, **5**:58-73.

53. Li A, Daggett V: **Identification and characterisation of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations.** *J Mol Biol* 1996, **257**:412-429.

54. Li A, Daggett V: **Characterisation of the transition state of protein unfolding by use of molecular dynamics: chymotrypsin inhibitor 2.** *Proc Natl Acad Sci USA* 1994, **91**:10430-10434.

55. Fersht AR, Itzhaki LS, elMasry NF, Matthews JM, Otzen DE: **Single versus parallel pathways of protein folding and fractional formation of structure in the transition state.** *Proc Natl Acad Sci USA* 1994, **91**:10426-10429.

56. Itzhaki LS, Otzen DE, Fersht AR: **The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation condensation mechanism for protein folding.** *J Mol Biol* 1995, **254**:260-288.

57. Ladurner AG, Itzhaki LS, Daggett V, Fersht AR: **Synergy between simulations and experiment in describing the energy landscape of protein folding.** *Proc Natl Acad Sci USA* 1998, **95**:8473-8478.

58. Caflish A, Karplus M: **Acid and thermal denaturation of barnase investigated by molecular dynamics simulations.** *J Mol Biol* 1995, **252**:672-708.

59. Tirado-Rives J, Orozco M, Jorqensen WL: **Molecular dynamics simulations of the unfolding of barnase in water and 8M urea.** *Biochemistry* 1997, **36**:7313-7329.

60. Pike ACW, Acharya KR: **A structural basis for the interaction of urea with lysozyme.** *Protein Sci* 1994, **3**:706-710.

61. Liepinsh E, Otting G: **Specificity of urea binding to proteins.** *J Am Chem Soc* 1994, **116**:9670-9674.

62. Brooks CL III, Boczko EM: **First-principles calculation of the folding free energy of a three-helix bundle protein.** *Science* 1995, **269**:393-396.

63. Guo Z, Brooks CL III, Boczko EM: **Exploring the folding free energy surface of a three-helix bundle protein.** *Proc Natl Acad Sci USA* 1997, **94**:10161-10166.

64. Sheinermann FB, Brooks CL III: **Molecular picture of folding of a small α/β protein.** *Proc Natl Acad Sci USA* 1998, **95**:1562-1567.

65. Zhou Y, Karplus M: **Protein folding thermodynamics of a model three-helix-bundle protein.** *Proc Natl Acad Sci USA* 1997, 94:14429-14432.

66. Karplus M, Weaver DL: **Folding dynamics: the diffusion-collision model and experimental data.** *Protein Sci* 1994, 3:650-668.

67. Daura X, Jaun B, Seebach D, van Gunsteren WF, Mark AE: **Reversible peptide folding in solution by molecular dynamics simulation.** *J Mol Biol* 1998, 280:925-932.

68. Schaefer M, Bartels C, Karplus M: **Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model.** *J Mol Biol* 1998, 284:835-848.

69. Wu X, Wang S: **Self-guided molecular dynamics simulations for efficient conformational search.** *J Phys Chem B* 1998, 102:7238-7250.

70. Berendsen HJC: **A glimpse of the holy grail?** *Science* 1998, 282:642-643.

71. Duan Y, Kollman PA: **Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.** *Science* 1998, 282:740-749.

72. Muñoz V, Serrano L: **Development of the multiple sequence approximation within the AGADIR model of α-helix formation.** *Biopolymers* 1996, 41:495-509.

73. Thomas PJ, Qu BH, Pederson PL: **Defective protein folding as a basis of human disease.** *Trends Biochem Sci* 1995, 20:456-459

74. Dobson CM: **How do we explore the energy landscapes for folding?** In *Simplicity and Complexity in Proteins and Nucleic Acids.* Edited by Frauenfelder H, Deisenhofer J, Wolynes PG. Berlin: Dahlem University Press; 1999:in press.

75. Plaxco KW, Gross M: **The importance of being unfolded.** *Nature* 1997, 386:657-659.

76. Labeit S, Kolmerer B: **Titins: giant proteins in charge of muscle ultrastructure and elasticity.** *Science* 1995, 290:273-296.

77. Lomakin A, Chung DS, Benedek GB, Kirschner DA, Teplow DB: **On the nucleation and growth of amyloid β-protein fibrils: detection of nuclei and quantification of rate constraints.** *Proc Natl Acad Sci USA* 1996, 93:1125-1129.

78. Ewalt KL, Hendrick JP, Houry WA, Hartl FU: ***In vivo* observation of polypeptide flux through the bacterial chaperonin system.** *Cell* 1997, 90:491-500.

79. Kelly JW: **The alternative conformations of amyloid proteins and**
•   **their multi-step assembly pathways.** *Curr Opin Struct Biol* 1998, 8:101-106.
A survey of proteins associated with the variety of amyloid-based diseases and a discussion of possible mechanisms by which the soluble forms of these proteins are converted into amyloid fibrils.

80. Guijarro JI, Sunde M, Jones JA, Campbell ID, Dobson CM: **Amyloid fibril formation by an SH3 domain.** *Proc Natl Acad Sci USA* 1998, 95:4224-4228.

81. Fink A: **Protein aggregation: folding aggregates, inclusion bodies**
•   **and amyloid.** *Fold Des* 1998, 3:R9-R23.
A comprehensive review of the nature of different types of protein aggregates and the relationship of these to the solution states from which they are formed.

82. Sunde M, Blake CCF: **The structure of amyloid fibrils by electron microscopy and X-ray diffraction.** *Adv Protein Chem* 1997, 50:123-159.

83. Chiti F, Webster P, Taddei N, Clark A, Stefani M, Ramponi G, Dobson CM: **Designing conditions for *in vitro* formation of amyloid protofilaments and fibrils.** *Proc Natl Acad Sci USA* 1999, in press.

84. Jiménez JL, Guijarro JI, Orlova E, Zurdo J, Dobson CM, Sunde M, Saibil HR: **Cryo-electron microscopy structure of an SH3 amyloid fibril and model of the molecular packing.** *EMBO J* 1999, in press.

85. Harper JD, Lansbury PT: **Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequence of the time-dependent solubility of amyloid proteins.** *Annu Rev Biochem* 1997, 66:385-407.

86. Prusiner S: **Prion diseases and the BSE crisis.** *Science* 1997, 278:245-251.