Towards complete descriptions of the free-energy landscapes of proteins

By Michele Vendruscolo and Christopher M. Dobson

Department of Chemistry, University of Cambridge,
Lensfield Road, Cambridge CB2 1EW, UK (cmd44@cam.ac.uk)

In recent years increasingly detailed information about the structures and dynamics of protein molecules has been obtained by innovative applications of experimental techniques, in particular nuclear magnetic resonance spectroscopy and protein engineering, and theoretical methods, notably molecular dynamics simulations. In this article we discuss how such approaches can be combined by incorporating a wide range of different types of experimental data as restraints in computer simulations to provide unprecedented detail about the ensembles of structures that describe proteins in a wide variety of states from the native structure to highly unfolded species. Knowledge of these ensembles is beginning to enable the complete free-energy landscapes of individual proteins to be defined at atomic resolution. This strategy has provided new insights into the mechanism by which proteins are able to fold into their native states, or by which they fail to do so and give rise to harmful aggregates that are associated with a wide range of debilitating human diseases.

Keywords: protein folding; energy landscape; protein misfolding; protein aggregation; amyloid diseases; computer simulations

1. Introduction

Characterizing the nature of partially folded states of proteins is crucial for understanding the determinants of many aspects of their behaviour, including the stability of the native state, the kinetics of folding and the mechanisms of misfolding (Vendruscolo *et al.* 2003*a*). Extensively or partially unfolded species are also interesting in their own right, as in the cell they are often involved in processes such as signal transduction, translocation across membranes, transcriptional activation and cell cycle regulation (Uversky 2003).

Most approaches used for protein structure determination, including X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and electron microscopy, involve three major steps:

(i) a suitable technique is chosen and the results of the experimental measurements are interpreted in terms of parameters known to be related to molecular structures;

(ii) an appropriate model is chosen to represent the structure and the energy of the molecule;

(iii) an optimization method is defined that minimizes violations of the calculated conformations from the experimentally derived structural information.

This approach has proved extremely successful for native states of proteins and protein complexes. Examples of recent spectacular achievements include the determination of the structures of F_1-ATPase (Abrahams *et al.* 1994), the proteasome (Groll *et al.* 1997) and the ribosome (Ban *et al.* 2000; Wimberly *et al.* 2000). In principle, extensions of these methods will make it possible to determine the structures of yet larger and more complex systems, provided that the structures themselves are well defined (Aloy *et al.* 2004). However, as we mentioned above, a number of important aspects of protein behaviour involve non-native states, or more generally states that are only partially or transiently structured (Vendruscolo *et al.* 2003*a*) (see figure 1). The general problem posed by the determination of the structures of such species is that they must be described in terms of ensembles of conformations. Sometimes these ensembles are made up by relatively compact, native-like structures, but in other cases they may be rather heterogeneous. In this latter case, the conformational preferences may not differ greatly from those corresponding to a random-coil state (Lindorff-Larsen *et al.* 2004*a*). In addition, the dynamics within stable or metastable non-native states are rich and may span time-scales from picoseconds to seconds or more. Furthermore, it is often extremely important to characterize the dynamics of transitions between different states to understand complex biological processes such as protein folding. In such cases, kinetic measurements, particularly when coupled with protein engineering techniques, can provide structural information about the relevant transition states (Fersht 1999).

One major goal for theoreticians working in this area is to calculate protein structures, and the pathways that connect them, *ab initio* from the sequence of the polypeptide chain. In approaches of this type only step (ii) in the previous scheme is used, without assuming directly any experimental knowledge. The latter, however, may be used afterwards for the purpose of validating the structures. This approach has been used successfully in a number of applications (Fersht & Daggett 2002) and is based upon the availability of reliable force fields, as well as of a considerable amount of computer time. The latter issue is of major significance in the case of non-native states, where the ensembles may cover vast regions of conformational space.

Experimentalists, in contrast, strive to measure parameters that describe in the required detail the key structural features of a given state of a protein molecule. In the case of native states the procedures are relatively well developed, notably in both X-ray diffraction and NMR spectroscopy (Brunger *et al.* 1998). More recently, a wide range of experimental techniques has been developed to probe the structure and dynamics of non-native states of proteins (Ferguson & Fersht 2003; Vendruscolo *et al.* 2003*a*). In particular, new approaches involving NMR spectroscopy have been shown to be very powerful in providing such information at the atomic level, e.g. through the detection of dipolar couplings between atomic nuclei and their relaxation rates (Akke 2002; Kay 1998). Experimentalists so far have relied for the most part on using mainly step (i) in the previous scheme in the context of the structure determination of non-native states. The results are sometimes interpreted simply by relating them qualitatively (and thus implicitly carrying out step (iii)) to the
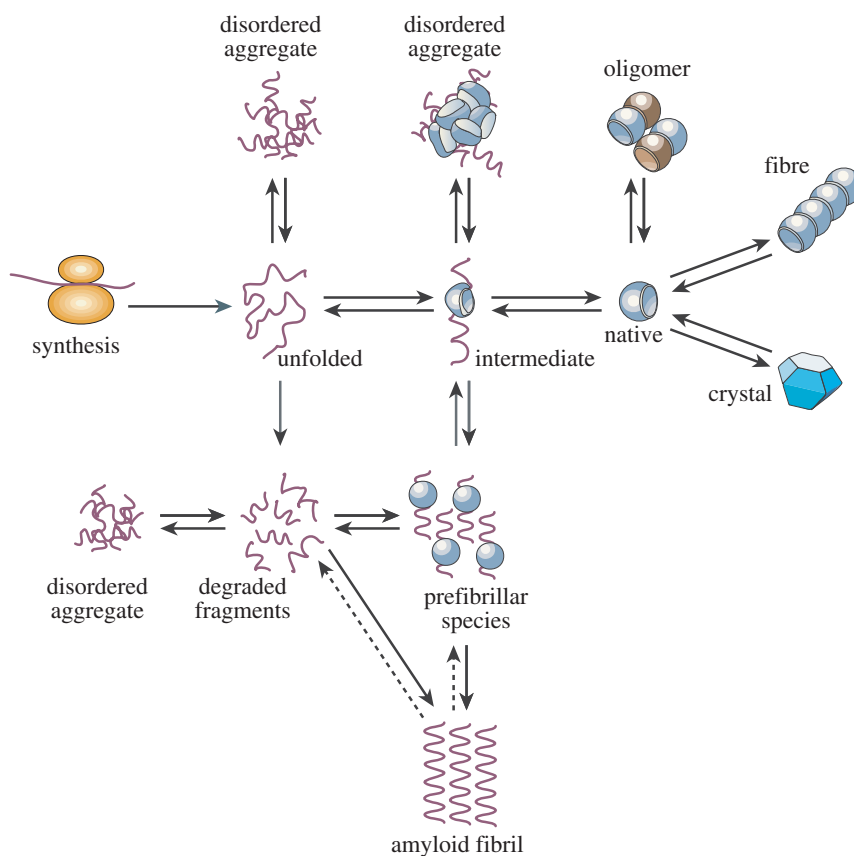
Figure 1. Some of the states accessible to a protein after its
biosynthesis in the ribosome (adapted from Dobson (2003*b*)).

structure of the native state (a procedure corresponding to step (ii)). Although this
strategy has provided valuable insights into the general organization of non-native
states, it becomes difficult to apply when the topology departs significantly from
the native one and often lacks any specific structural information and predictive
power.

In order to develop techniques for the quantitative determination of non-native
structures it is essential to address the issue that there will often be quite large
ensembles of conformations that minimize the violations from the experimental data
(Lindorff-Larsen *et al.* 2004*a*; Vendruscolo *et al.* 2001, 2003*b*). Indeed, a series of
questions arises about any approach: is the structural interpretation of the restraints
correct? Is the ensemble under-determined, i.e. are there too few restraints? How
well does the calculated ensemble correspond to the true one? Moreover, as exper-
imental techniques give time- and ensemble-averaged measurements, one should be
able to reconstruct a representative ensemble of conformations knowing only some
of its average properties (Lindorff-Larsen *et al.* 2004*a*; Vendruscolo *et al.* 2003*b*).
In the following sections we shall discuss methods by which these problems can be
approached.

2. Restrained simulations: applications

(*a*) *General approach*

We have recently developed an approach that combines directly computer-based simulations with experimental data, the latter being used as restraints in the sampling of conformational space. This approach aims at finding all the conformations that minimize a pseudo-energy function composed of two parts. The first part defines the physico-chemical properties of the polypeptide chains and their environment, and the second part penalizes deviations from the experimentally derived information (Paci *et al.* 2002; Vendruscolo *et al.* 2001). The simultaneous optimization of the two components of the pseudo-energy function ensures that the molecular models generated by the simulations have the properties of proteins and that they are compatible with the experimental measurements.

(*b*) *Native-state fluctuations*

Under native conditions, a globular protein exists for most of the time in a narrow ensemble of conformations, centred about a well-defined average structure. Large thermal fluctuations are, however, possible and a wide variety of experimental data, ranging from proteolytic susceptibility to fluorescence quenching measurements, indicate that the protein does occasionally visit substantially more unfolded conformations that can be of great biological significance, for example, in defining the susceptibility to degradation or to aggregation (Dobson 2003*a*, 2004). Such states are, however, difficult to study as they are populated only during very rare and transient events. Hydrogen-exchange experiments are uniquely suited to provide residue-specific information about such rare fluctuations (Dobson *et al.* 1998; Fersht 1999). These experiments monitor, through the use of NMR spectroscopy, the replacement of labile hydrogen atoms, notably those bonded to backbone amide nitrogen atoms, for solvent hydrogen atoms of a different isotope. Results are usually given in terms of *protection factors*, which are defined as the ratios of the rates of exchange observed for labile hydrogen atoms in the states of the protein under consideration to the rates of exchange of the same type of hydrogen atoms in a reference state representing a fully unfolded random-coil state.

We have recently developed a method that uses protection factors to bias the sampling of conformational space in order to define the structures of the protein from which the exchange takes place (Vendruscolo *et al.* 2003*c*). This approach is based on the assumption that protection factors are proportional to the free-energy differences between open (or *exchange competent*) and closed (or *exchange incompetent*) conformations (Fersht 1999). In the calculations, these free energies are approximated in terms of numbers of inter-atomic contacts and hydrogen bonds formed by a given labile hydrogen atom; these terms are chosen following the assumption that protection arises either from burial inside the core of a protein or from hydrogen bonding. An important technical aspect of this computational approach is that, as protection factors are measured as averages over a large number of molecules in solution (typically 10^{18}), no single molecule can be required to satisfy a given set of experimental protection factors. Since an exchange event requires a significant disruption of the structure, at least in the proximity of the amide group involved, the simultaneous enforcement of all the protection factors to be satisfied by a single molecule
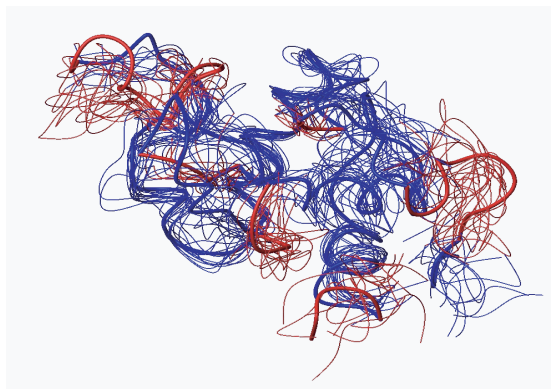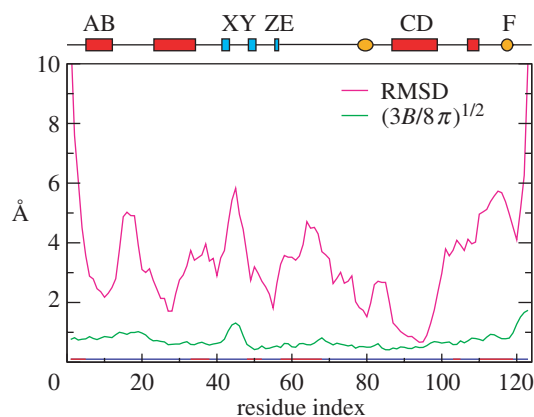
Figure 2. Representation of a series of structures of the protein α-lactalbumin that give rise to hydrogen exchange under native conditions (adapted from Vendruscolo *et al.* (2003*c*)). For comparison, conformation represented by a thick line represents the X-ray structure.

would always require global unfolding. Therefore, we impose the restraints as averages over a sufficient number of replicas of the molecules (typically 20), so that a single molecule at any given instant may maintain its overall folded conformation, except for one or very few local structural rearrangements.

We have initially applied this approach to determine the structures that give rise to hydrogen exchange from the native state of α-lactalbumin (Vendruscolo *et al.* 2003*c*). The ensemble of resulting structures, shown in figure 2, reveals that the large fluctuations that give rise to exchange may be more than five times larger than those measured by crystallographic B factors, which describe the average structural variability, and that some native α-helices may undergo sizeable reorientations. The use of this technique is not, however, limited to native states, and can be applied to any state, stable or metastable, for which some protection from hydrogen exchange is experimentally detectable. We anticipate that particularly interesting applications will involve the determination of the structures of molten globule states, of reaction intermediates and of amyloid fibrils.

Native states of proteins are also highly dynamic on much shorter time-scales, from picoseconds to nanoseconds, and in this case such fluctuations occur around the average structure. Motions of individual bond vectors on these time-scales are

accessible through NMR relaxation experiments, which are usually analysed using the so-called *model-free* approach (Lipari & Szabo 1982), where it is assumed that large-scale motions of the molecule take place on longer time-scales than the small fluctuations that influence the relaxation data. In these experiments, the amplitude of the motion of a bond vector (usually a backbone amide group or a side-chain methyl group (Choy *et al.* 2003)) is quantified in terms of an NMR order parameter, called S^2. We have recently proposed a method that uses S^2 values as restraints in computer simulations to obtain a model of the motions of the protein backbone and side-chains on the nanosecond time-scale (Best & Vendruscolo 2004). The use of replicas is in this case essential, as order parameters are by definition always equal to unity for a single molecule at any given time. In this computational technique, order parameter restraints act to define the *width* of the distribution of bond vectors. They are therefore complementary to other restraints, such as NOEs or residual dipolar couplings, where *average values* are enforced in the simulations.

Application of this approach to describe the native dynamics of the fibronectin type-III domain from human tenascin has provided a uniquely detailed description of the relative populations of the rotameric states of the side-chains of the protein and has been validated by the calculation of the average values of a set of three-bond (^{3}J) couplings that were compared with the corresponding experimental values (Best & Vendruscolo 2004). The successful prediction of these quantities, which were not used as a bias in the simulations, suggests that the ensemble of conformations determined through the use of S^2 order parameters as a bias is a faithful representation of the true native ensemble. We anticipate that the simultaneous use of restraints that define average values of molecular properties and their width should provide a more accurate determination of the ensemble of structures representing the native state of proteins and that the use of this approach will give new insight into the role of dynamic effects in the packing of hydrophobic cores of proteins and the changes in the dynamics upon mutation or upon ligand binding. Indeed, we have very recently obtained results that support this conjecture (Lindorff-Larsen *et al.* 2004*c*).

(*c*) *Unfolded states*

The conformations present in highly unfolded states of proteins may be expected to resemble those present in a random-coil ensemble, except for the effect of any residual preferential interaction between side-chain or main-chain atoms. The characterization of this residual structure is important, since any weak conformational preference may, for example, bias the system towards conformations involved in the initial events leading to folding and, in less benign cases, to misfolding. Furthermore, this type of structure is now recognized to be important in its own right, as a sizeable portion, perhaps up to one-third, of the proteins encoded in the genome of an organism may be, at least in some regions, intrinsically disordered, i.e. lacking a well-defined three-dimensional structure under physiological conditions unless in complex with their binding partners (Dunker *et al.* 2001).

NMR spectroscopy can be used, at least in favourable cases, to obtain structural information about such unfolded states, e.g. through NOEs, chemical shifts or residual dipolar couplings (Choy *et al.* 2003, 2001; Crowhurst & Forman-Kay 2003; Dyson & Wright *et al.* 2002; Kazmirski *et al.* 2001; Wong *et al.* 2000). Relaxation experiments may also be used to identify regions of reduced mobility and hence of partial

ordering; the existence of persistent and partially non-native hydrophobic clusters in the unfolded state of lysozyme has been, for example, demonstrated in this way (Klein-Seetharaman *et al.* 2002). Since dipolar interactions between unpaired electrons and atomic nuclei up to distances of 20 Å can be detected experimentally (compared with *ca.* 5 Å for intermolecular dipolar interactions) they may provide extremely valuable information about the structural ensemble describing such a state. We are using this type of experiment to characterize the residual structure in chemically unfolded states, e.g. the four-helix bundle protein ACBP (Lindorff-Larsen *et al.* 2004*a*) (see figure 2). These studies reveal that in the guanidine-induced denatured state of this protein there is very weak preferential sampling of a native-like topology, but no evidence of a persistent population of native-like structures. Most structures in the unfolded state ensemble of ACBP have random-coil-like characteristics, except for a small but significant increase in the probability of specific native-like interactions in the regions of the sequences corresponding to the second and the fourth α helices in the native state. Interestingly, the latter regions are important in the nucleation of the folding process. These results suggest that evolution may have resulted in the selection of polypeptide sequences that have a tendency in the unfolded state to form at least transiently clusters of residues with a native-like arrangement; the bias in the stochastic search towards specific interactions may thus increase the probability of formation of the folding nucleus.

(*d*) *Transition states for folding*

While most proteins are likely to fold through a complex process characterized by a multiplicity of intermediates formed on different time-scales (*multi-exponential kinetics*), it has been shown that some small proteins can be observed essentially in only two states: the native and the unfolded ones (Fersht 1999). During folding, the free-energy barrier between the unfolded and the folded states must be crossed, but it is extremely difficult, as for all reactions, to obtain any direct experimental information about the transition state for folding, because of its transient nature. The major technique currently available involves protein engineering (Fersht 1999), in which conservative mutations of particular side-chains are carried out and the effects on the kinetics of folding and unfolding are measured. In order to map the transition state, conservative mutations (e.g. the removal of a methyl group) are most readily interpreted. Experimental results are often given in terms of ϕ values; a ϕ value of unity indicates that the stability of interactions made by a particular side chain in the transition state is the same as in the native state. A ϕ value of zero, by contrast, indicates that no native-like interactions are formed in the transition state, and therefore that residue does not play an active role in determining the rate of folding.

In some respect, ϕ values are analogous to NOEs, as they provide structural information in terms of distance restraints (Daggett *et al.* 1996; Vendruscolo *et al.* 2001). An important difference, however, is that fractional ϕ values are difficult to interpret in structural terms. Nevertheless, armed with the assumption that ϕ values can be approximated to the fraction of native interactions formed in the transition state, at least in the cases of conservative mutations, we have shown that it is possible to use ϕ values as restraints in simulations to generate ensembles of structures that represent the transition state of proteins (Davis *et al.* 2002; Lindorff-Larsen *et al.* 2004*b*;
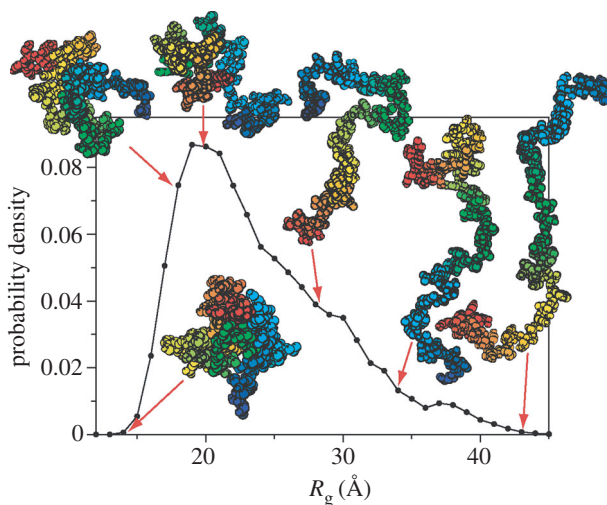
Figure 3. The unfolded state of the protein ACBP. This state consists of a large ensemble of conformations, which range from compact, native-like structures to highly extended ones (adapted from Lindorff-Larsen *et al.* (2004*a*)).

Paci *et al.* 2002, 2003, 2004; Vendruscolo *et al.* 2001). The most important insight that these calculations have given so far is that the topology of the transition state appears to be established as native-like when the ϕ values of just a small number of residues are specified. For example, using only three ϕ values, the topology of the native state is present in the large majority of the members of the transition state ensemble of AcP (Vendruscolo *et al.* 2001) and all the remaining ϕ values can be predicted with high precision (see figure 3). One of the consequences of this result is that the transition state ensemble is rather heterogeneous and yet its overall topology is native-like (Lindorff-Larsen *et al.* 2004*b*, 2005; Vendruscolo *et al.* 2001). We have demonstrated this conclusion in a quantitative way by showing that the overwhelming majority of the structures of the transition state ensemble matches the correct fold in the PDB in a structural alignment procedure (Lindorff-Larsen *et al.* 2004*b*, 2005).

These results imply that a careful ϕ value analysis provides an over-determined set of restraints, as far as the definition of the overall topology of the molecule is concerned, in that a few specific ϕ values are sufficient for its unambiguous definition. This surprising conclusion may be understood in the context of a network analysis of protein structures (Vendruscolo *et al.* 2002) (see figure 4). Given the constraints that result from the connectivity of the polypeptide chain, the overall fold of a protein is established when the interactions between a few key residues are created. In turn, this result provides a physical description of the mechanism of nucleation process that is thought to be principally responsible for efficient protein folding. It also provides a structural model of the critical nuclei that complements those obtained by techniques such as atomic force microscopy for nucleation processes involving much larger particles, such as the crystallization of proteins (Yau & Vekilov 2000) and of other colloidal systems (Russel 2003).

In terms of the finer details of the structures of the transition state ensemble the restraints obtained through ϕ value analysis will generally represent an under-
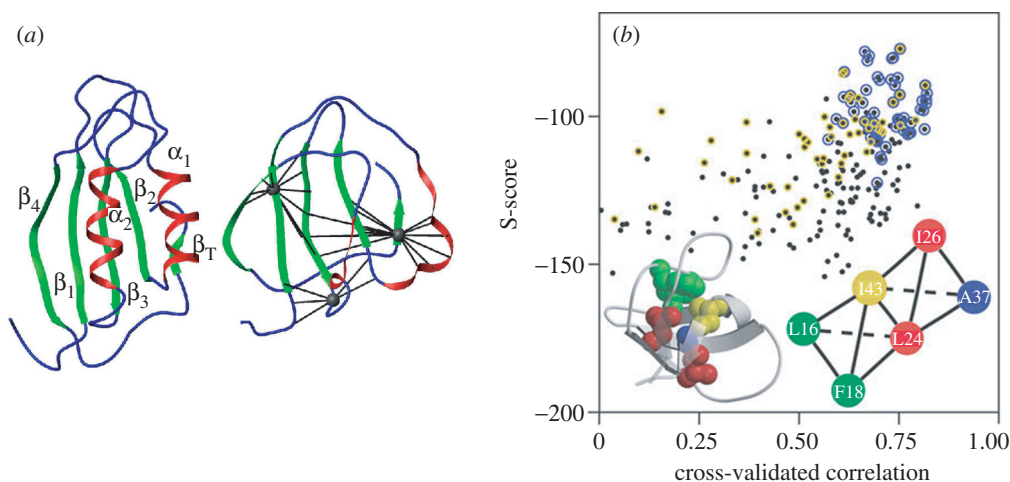
Figure 4. (*a*) The transition state for folding of the protein acylphosphatase. The network of interactions that stabilize this state is determined when the contacts of three key residues are specified (adapted from Vendruscolo *et al.* (2002)). (*b*) The critical network of interactions involved in the formation of the folding nucleus of the src SH3 domain is formed by any triplet of residues chosen among a set of six central residues (adapted from Lindorff-Larsen *et al.* (2004*b*)). When these interactions are formed, all the remaining ϕ values can be predicted (horizontal axis) and the topology of the ensemble resembles that of the native state (vertical axis).

determined set, as they may provide insufficient information to describe the structure accurately at a more local level. The most important question in this respect, however, is whether the calculated ensembles can be validated by using them to make predictions about the results of additional experimental measurements. In the case of the transition state of, for example, the protein barnase, the structures have been used to predict 'double-mutant' ϕ values that are in remarkably good agreement with experiments (Salvatella *et al.* 2004). This type of test is particularly stringent, as double-mutant experiments often report on the interactions between residue pairs at the surface of the molecule, i.e. far from the folding nucleus, which is the region where most structural information is provided by the single-mutant ϕ values.

Another important question is whether the structural interpretation that we have given of ϕ values in terms of the fraction of native contacts is appropriate. This interpretation is, however, supported by the fact that, as we mentioned above, we can use a subset of ϕ values as restraints to predict all the remaining ones (Lindorff-Larsen *et al.* 2004*b*; Paci *et al.* 2003, 2004; Vendruscolo *et al.* 2001). Since ϕ values are derived from ensemble-averaged experimental measurements, it is also important to examine how the use of replicas influences the results of the structure determination procedure. We have shown that, when parallel folding pathways are present, use of multiple replicas in the simulations restrained by the ϕ values is essential for making correct predictions (Davis *et al.* 2002). However, when folding proceeds through a dominant pathway, albeit involving broad ensembles of conformations in many cases, i.e. through a single transition state, calculations using a single replica are a suitable method. In most cases, proteins appear to fold through a single dominant pathway, perhaps as a consequence of the evolution of a quality-control mechanism that minimizes the tendency to misfold (Davis *et al.* 2002).

Table 1. *Example of experimental techniques suitable for providing structural information about native and non-native states of proteins*

(References are given in the cases when these techniques have been used to provide restraints for the type of simulations discussed in this paper.)

state	technique
native state	NMR order parameters (Best & Vendruscolo 2004; Lindorff-Larsen *et al.* 2004*c*)
	residual dipolar couplings (Lindorff-Larsen *et al.* 2004*c*)
	nuclear Overhauser effects (Lindorff-Larsen *et al.* 2004*c*)
	J couplings (Best & Vendruscolo 2004; Lindorff-Larsen *et al.* 2004*c*)
	hydrogen exchange protection factors (Vendruscolo *et al.* 2003*c*)
transition state	ϕ values (Lindorff-Larsen *et al.* 2004*b*; Paci *et al.* 2002, 2003, 2004; Vendruscolo *et al.* 2001)
	double mutant cycles (Salvatella *et al.* 2004)
	β Tanford values (Lindorff-Larsen *et al.* 2004*b*; Paci *et al.* 2002, 2003, 2004; Vendruscolo *et al.* 2001)
intermediate state	NMR chemical shifts (Korzhnev *et al.* 2004)
	hydrogen exchange protection factors
	nuclear Overhauser effects
molten globule	progressive unfolding detected by NMR techniques (Vendruscolo *et al.* 2003*b*)
	hydrogen exchange
unfolded state	spin-labelling NMR techniques (Lindorff-Larsen *et al.* 2004*a*)
	nuclear Overhauser effects
	NMR chemical shifts
	residual dipolar couplings
protein aggregate	solid-state NMR distance measurements (Jaroniec *et al.* 2002, 2004)
	NMR chemical shifts
	spin-labelling NMR techniques
	hydrogen exchange protection factors

(*e*) *Intermediates*

The free-energy landscapes of proteins appear to have evolved to minimize the presence of long-lived non-native intermediate states or to regulate their properties to optimize folding or to reduce the possibility of misfolding and aggregation (Dobson 2003*a*). Intermediate states must, however, always be populated to some degree. This expectation is based on a view of protein folding as a condensation process, akin to liquid–solid transitions. It is known that this type of process tends to progress through metastable states. According to this observation, the phase that appears first in the process need not be the most stable thermodynamically, but is the one
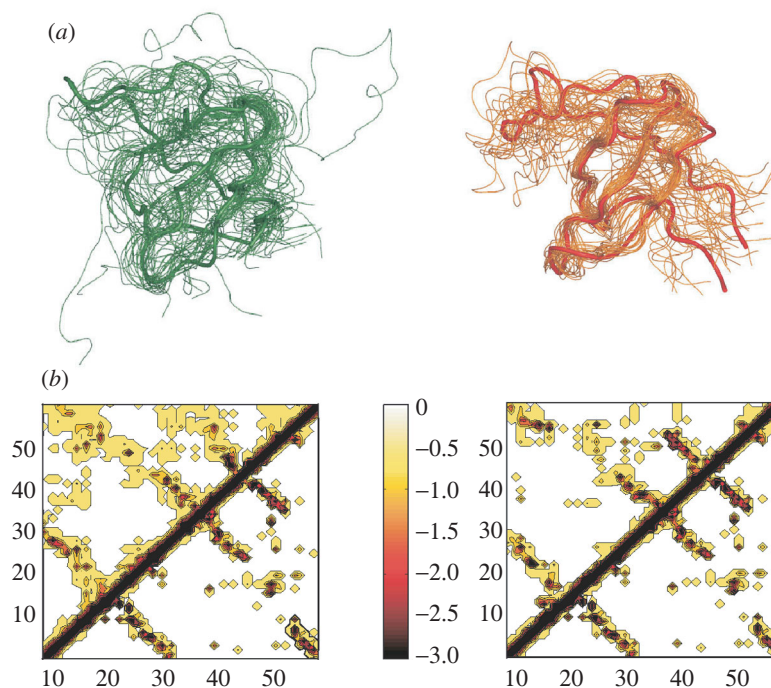
Figure 5. (*a*) Representation of the ensembles of structures making up the intermediate states of two mutational variants of the Fyn SH3 domain (left, G48M; right G48V). (*b*) Energy maps corresponding to the two intermediate states; for comparison the energy map of the native state is shown below the diagonal in each case. An energy map illustrates the interaction energy between two amino acids averaged over the ensemble of conformations representing a given state (Korzhnev *et al.* 2004).

closest in free energy to the initial state (Auer & Frenkel 2001; Ostwald 1897) or, more precisely, since nucleation is a dynamic process, the state that is separated by the lowest energy barrier from the initial state. The short lifetimes of many intermediates in protein folding, however, makes it difficult to extract experimental information about their structures. For the completion of the folding process—the crossing of the free-energy barrier between the intermediate and the native states—local interactions are probably responsible for the close packing of the side chains. More complex folding processes may involve the successive formation of several intermediates according to similar principles. In addition, large proteins are likely to fold in segments, or domains, with a final step in which such domains are closely packed together (Fersht & Daggett 2002; Vendruscolo *et al.* 2003*b*).

Structural information about intermediates can be obtained through ϕ value analysis, the measurement of NOEs, of NMR chemical shifts and of protection factors from hydrogen exchange. In principle, all these types of measurements can be used as restraints in computer simulations (see table 1). As one example, we have recently identified and characterized intermediates present at very low population (about 1%) in equilibrium with the native and unfolded states of two mutants of an SH3 domain (Korzhnev *et al.* 2004); in this case the restraints were based on the measurement of the chemical shift differences between the native and the intermediate states. The

latter were obtained by relaxation dispersion NMR experiments that extracted such data from the contributions of chemical exchange to the transverse relaxation rates. By repeating these experiments at different temperatures, not only structural information but also thermodynamic and kinetic parameters characterizing the folding process can be obtained. In these experiments, therefore, contributions to line widths of cross peaks in ^{1}H–^{15}N correlation spectra are interpreted in terms of the rates of interconversion between states (kinetics), their chemical shift differences (structure) and their populations (thermodynamics) (Akke 2002; Mulder *et al.* 2001). The results of the restrained simulations reveal that the SH3 intermediates share distinctive topological features with the corresponding transition state for folding (see figure 5). By taking this approach further, this study suggests a general strategy for an experimental determination of the free-energy landscape of a protein by applying this methodology to a series of mutants whose intermediates have different degrees of structural similarity to the native state.

(*f*) 'Molten globules'

The native states of some proteins that are relatively unstable with respect to perturbations such as the removal of a substrate or metal ion, the reduction of disulphide bridges or the lowering of pH can form 'molten globules'. In these cases, the tight native packing of side-chains is lost so as to allow the protein to convert into the so-called molten globule state, which is nearly as compact as the native state, despite having large structural fluctuations (Kuwajima 1996; McParland *et al.* 2002; Pande & Rokhsar 1998; Schulman *et al.* 1997). The dynamic nature of the molten globule state, however, makes it difficult to extract experimental information about its residual structure, as such species do not crystallize and NMR spectra are usually severely broadened by chemical exchange effects of the type discussed in the previous section. We have shown that it is possible to use NMR ^{1}H–^{15}N heteronuclear single quantum coherence (HSQC) spectra recorded at increasing concentrations of denaturant to monitor the relative stability of native-like interactions in the molten globule state of α-lactalbumin (Schulman *et al.* 1997). These NMR data can be translated into restraints for simulations, by defining the set of residues that become unstructured at different stages of the denaturation procedure. Analysis of the structural ensembles obtained in this manner reveals that, upon addition of urea, α-lactalbumin can undergo a structural transition in which the interface between the two structural domains of the protein becomes disordered, while the two domains themselves retain a significant degree of native-like order (Vendruscolo *et al.* 2003*b*) (see figure 6). The use of the restraints derived from experimental measurements at different concentrations of denaturant allowed extensive sampling of conformational space to be achieved and, as a result, a coarse-grained experimental free-energy landscape could be defined (Vendruscolo *et al.* 2003*b*). The result of this procedure suggests that evolution has created landscapes that are characterized by the presence of deep and extended basins that make them generally robust against perturbations, such as the addition of denaturant, changes in temperature, mutations or chemical modifications of the molecules (Vendruscolo & Paci 2003; Vendruscolo *et al.* 2003*b*).

The molten globule of α-lactalbumin represents a particularly interesting case for structural determination approaches using restrained simulations, as a range of different sets of experimental measurements is available (Kuwajima 1996; Vendruscolo
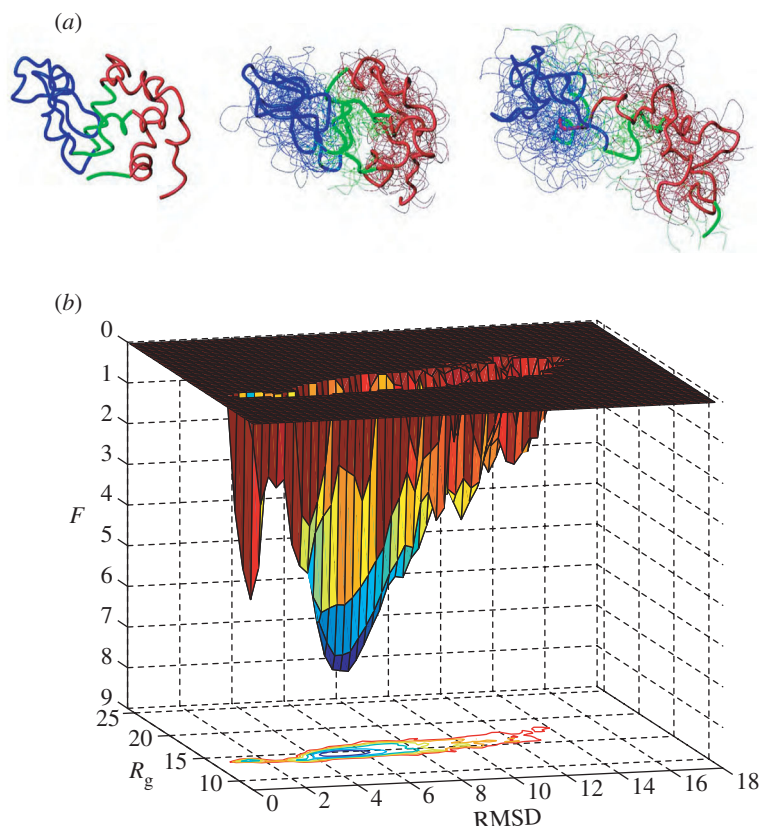
(a)

(b)

Figure 6. (a) Representation of the structures of α-lactalbumin that are populated in the native state (left), and in the 'molten globule' state at pH 2 (centre) and at pH 2 and 10 M urea (right). (b) Representation of free-energy landscape at pH 2, reconstructed by combining the ensembles of conformations corresponding to NMR ^{1}H–^{15}N HSQC spectra recorded at increasing concentrations of urea (adapted from Vendruscolo *et al.* (2003b)).

et al. 2003b). By repeating the determination of the ensemble of structures representing the molten globule state using different types of experimental data, it should be possible to validate the structural interpretation of the experimental measurements. In addition, combining the various experimental sets of data as restraints in the structure calculation should allow increasingly well-defined ensembles to be determined. Studies of this type are in progress.

(g) *Protein aggregates*

As figure 1 illustrates, it is important to characterize not only monomeric states, but also aggregated states, in order to provide a complete description of the conformations accessible to a protein molecule. Of particular interest is the recent recognition that the phase diagrams of polypeptide chains contain a second kind of highly ordered state, in addition to the native state. This state is characterized by ordered linear assemblies, known as an amyloid (Dobson 1999, 2003a). The formation of this type of assembly is generally harmful to a living system and a wide range of cellular

mechanisms have evolved to avoid it, or to counteract it when it happens (Cohen & Kelly 2003; Selkoe 2003). These defence mechanisms are, however, sometimes overwhelmed, and more than 20 diseases have so far been associated with the presence of amyloid aggregates in a variety of different tissues (Soto 2003; Stefani & Dobson 2003). One reason for this imperfect quality control is that the sequence determinants of the folding process are similar to those of the misfolding process, resulting in an inherent tendency for the latter to occur even under physiological conditions (Chiti *et al.* 2003).

Conventional techniques, such as X-ray crystallography and solution NMR spectroscopy, are not ideally suited for obtaining the detailed information needed for structural determination of amyloid fibrils. It has been recently recognized, however, that solid-state NMR techniques allow interatomic distances within amyloid fibrils to be measured with great accuracy (Jaroniec *et al.* 2002, 2004; Petkova *et al.* 2002; Tycko 2004). These distances can be used as restraints in computational approaches to structure determination. In this case, simulation techniques that are commonly used for the determination of native states of proteins (Brunger *et al.* 1998) may be suitable to define the narrow ensembles of conformations that characterize the rather rigid structure of amyloid fibrils.

Of very great interest in addition to the structure of amyloid fibrils is the determination of the intermediate states present along the pathways that lead to their formation. A question of general interest is the extent to which folding and misfolding intermediates resemble each other and to what degree folding and misfolding pathways share common traits. Studies of the aggregation of lysozyme, for example, suggested that the amyloidogenic intermediate is similar to species populated during the normal folding process (Dobson 2001). In addition, recent studies have, for example, identified amino acids, sometimes known as gatekeepers, that appear to play a crucial role in directing the correct folding (Chiti *et al.* 2002; Otzen & Oliveberg 1999). One exciting possibility is to use NMR-derived restraints (McParland *et al.* 2002) of the type described above to bias computer simulations in order to determine an ensemble of conformations representing these amyloidogenic intermediates. Such a strategy could provide a uniquely powerful approach to defining the molecular basis of the different types of misfolding diseases.

3. Conclusions

Structural biology is facing exciting new challenges and opportunities. On the one hand it is becoming possible to resolve at atomic resolution the structures of ever larger complexes, such as the ribosome (Ban *et al.* 2000; Wimberly *et al.* 2000); indeed it has been suggested that the structural properties of entire cells can be determined, although initially at low resolution (Medalia *et al.* 2002). On the other hand, it is also becoming clear that biological macromolecules, and the complexes that they form, exist in highly dynamical states. An understanding of their behaviour requires a knowledge of all the conformations that they can assume, as well as the interactions that stabilize them and permit their interconversion (Vendruscolo *et al.* 2003*a*) (see figure 1).

In this article we have outlined some results of our long-term programme to define the different states that are accessible to a set of representative molecules in the cell. In the approach that we have discussed here, experimental measurements are

translated into structural restraints, which are then used in computer simulations to determine ensembles of conformations at high resolution. The application of this approach to specific proteins has provided important insights into the general principles regulating the behaviour of these macromolecules, such as how nucleation mechanisms are established at an atomic level (Lindorff-Larsen *et al.* 2004*b*; Vendruscolo *et al.* 2001), and how free-energy landscapes of proteins that have been selected by biological evolution are robust to many types of perturbations (Vendruscolo & Paci 2003; Vendruscolo *et al.* 2003*b*).

Structural studies such as those that we have presented here will gradually provide complete descriptions of the behaviour of individual macromolecule as well as interacting macromolecules, including proteins, lipids and nucleic acids. These descriptions will require representations such as those we have discussed, such as probability distributions, energy plots and free-energy landscapes. We believe that the innovative combinations of experimental data and simulation techniques of the type that we have described here have a unique role to play in enabling these objectives to be achieved.

References

Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. 1994 Structure at 2.8-angstrom resolution of F1-Atpase from bovine heart-mitochondria. *Nature* **370**, 621–628.

Akke, M. 2002 NMR methods for characterizing microsecond to millisecond dynamics in recognition and catalysis. *Curr. Opin. Struct. Biol.* **12**, 642–647.

Aloy, P. (and 10 others) 2004 Structure-based assembly of protein complexes in yeast. *Science* **303**, 2026–2029.

Auer, S. & Frenkel, D. 2001 Suppression of crystal nucleation in polydisperse colloids due to increase of the surface free energy. *Nature* **413**, 711–713.

Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. 2000 The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution. *Science* **289**, 905–920.

Best, R. B. & Vendruscolo, M. 2004 Determination of protein structures consistent with NMR order parameters. *J. Am. Chem. Soc.* **126**, 8090–8091.

Brunger, A. T. (and 13 others) 1998 Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr.* D **54**, 905–921.

Chiti, F., Taddei, N., Baroni, F., Capanni, C., Stefani, M., Ramponi, G. & Dobson, C. M. 2002 Kinetic partitioning of protein folding and aggregation. *Nat. Struct. Biol.* **9**, 137–143.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M. 2003 Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424**, 805–808.

Choy, W. Y., Shortle, D. & Kay, L. E. 2003 Side chain dynamics in unfolded protein states: an NMR based ^{2}H spin relaxation study of Delta 131 Delta. *J. Am. Chem. Soc.* **125**, 1748–1758.

Choy, W. Y.& Forman-Kay, J. D. 2001 Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* **308**, 1011–1032.

Cohen, F. E. & Kelly, J. W. 2003 Therapeutic approaches to protein-misfolding diseases. *Nature* **426**, 905–909.

Crowhurst, K. A. & Forman-Kay, J. D. 2003 Aromatic and methyl NOES highlight hydrophobic clustering in the unfolded state of an SH3 domain. *Biochemistry* **42**, 8687–8695.

Daggett, V., Li, A. J., Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. 1996 Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol.* **257**, 430–440.

Davis, R., Dobson, C. M. & Vendruscolo, M. 2002 Determination of the structures of distinct transition state ensembles for a beta-sheet peptide with parallel folding pathways. *J. Chem. Phys.* **117**, 9510–9517.

Dobson, C. M. 1999 Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24**, 329–332.

Dobson, C. M. 2001 The structural basis of protein folding and its links with human disease. *Phil. Trans. R. Soc. B* **356**, 133–145.

Dobson, C. M. 2003*a* Protein folding and misfolding. *Nature* **426**, 884–890.

Dobson, C. M. 2003*b* Protein folding and disease: a view from the first Horizon Symposium. *Nat. Rev. Drug. Discov.* **2**, 154–160.

Dobson, C. M. 2004 In the footsteps of alchemists. *Science* **304**, 1259–1262.

Dobson, C. M., Sali, A. & Karplus, M. 1998 Protein folding: a perspective from theory and experiment. *Angew. Chem. Int. Ed.* **37**, 868–893.

Dunker, A. K. (and 19 others) 2001 Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59.

Dyson, H. J. & Wright, P. E. (eds) 2002 Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. In *Unfolded proteins*, Advances in Protein Chemistry, vol. 62, pp. 311–340.

Ferguson, N. & Fersht, A. R. 2003 Early events in protein folding. *Curr. Opin. Struct. Biol.* **13**, 75–81.

Fersht, A. R. 1999 *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding.* San Francisco, CA: Freeman.

Fersht, A. R. & Daggett, V. 2002 Protein folding and unfolding at atomic resolution. *Cell* **108**, 573–582.

Groll, M., Ditzel, L., Lowe, J., Stock, D., Bochtler, M., Bartunik, H. D. & Hüber, R. 1997 Structure of 20S proteasome from yeast at 2.4 angstrom resolution. *Nature* **386**, 463–471.

Jaroniec, C. P., MacPhee, C. E., Astrof, N. S., Dobson, C. M. & Griffin, R. G. 2002 Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril. *Proc. Natl Acad. Sci. USA* **99**, 16748–16753.

Jaroniec, C. P., MacPhee, C. E., Bajaj, V. S., McMahon, M. T., Dobson, C. M. & Griffin, R. G. 2004 High-resolution molecular structure of a peptide in an amyloid fibril determined by magic angle spinning NMR spectroscopy. *Proc. Natl Acad. Sci. USA* **101**, 711–716.

Kay, L. E. 1998 Protein dynamics from NMR. *Nat. Struct. Biol.* **5**, 513–517.

Kazmirski, S. L., Wong, K. B., Freund, S. M. V., Tan, Y. J., Fersht, A. R. & Daggett, V. 2001 Protein folding from a highly disordered denatured state: the folding pathway of chymotrypsin inhibitor 2 at atomic resolution. *Proc. Natl Acad. Sci. USA* **98**, 4349–4354.

Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T., Imoto, T., Smith, L. J., Dobson, C. M. & Schwalbe, H. 2002 Long-range interactions within a nonnative protein. *Science* **295**, 1719–1722.

Korzhnev, D. M., Salvatella, X., Vendruscolo, M., DiNardo, A., Davidson, A. R., Dobson, C. M. & Kay, L. E. 2004 Characterizing low populated folding intermediates of an SH3 domain by NMR spectroscopy. *Nature* **430**, 586–590.

Kuwajima, K. 1996 The molten globule state of alpha-lactalbumin. *Faseb J.* **10**, 102–109.

Lindorff-Larsen, K., Kristjansdottir, S., Teilum, K., Fieber, W., Dobson, C. M., Poulsen, F. M. & Vendruscolo, M. 2004*a* Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. *J. Am. Chem. Soc.* **126**, 3291–3299.

Lindorff-Larsen, K., Vendruscolo, M., Paci, E. & Dobson, C. M. 2004*b* Transition states for protein folding have native topologies despite high structural variability. *Nat. Struct. Molec. Biol.* **11**, 443–449.

Lindorff-Larsen, K., Best, R. B., Vendruscolo, M. & Dobson, C. M. 2004*c* Simultaneous determination of protein structure and dynamics. *Nature* (In the press.)

Lindorff-Larsen, K., Rogen, P., Paci, E., Vendruscolo, M. & Dobson, C. M. 2005 Protein folding and the organization of the protein topology universe. *Trends Biochem. Sci.* (In the press.)

Lipari, G. & Szabo, A. 1982 Model-free approach to the interpretation of nuclear magnetic-resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.* **104**, 4546–4559.

McParland, V. J., Kalverda, A. P., Homans, S. W. & Radford, S. E. 2002 Structural properties of an amyloid precursor of beta(2)-microglobulin. *Nat. Struct. Biol.* **9**, 326–331.

Medalia, O., Weber, I., Frangakis, A. S., Nicastro, D., Gerisch, G. & Baumeister, W. 2002 Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science* **298**, 1209–1213.

Mulder, F. A. A., Mittermaier, A., Hon, B., Dahlquist, F. W. & Kay, L. E. 2001 Studying excited states of proteins by NMR spectroscopy. *Nat. Struct. Biol.* **8**, 932–935.

Ostwald, W. 1897 Studien über die Bildung und Umwandlung fester Korper. *Z. Phys. Chem.* **22**, 289–302.

Otzen, D. E. & Oliveberg, M. 1999 Salt-induced detour through compact regions of the protein folding landscape. *Proc. Natl Acad. Sci. USA* **96**, 11746–11751.

Paci, E., Vendruscolo, M., Dobson, C. M. & Karplus, M. 2002 Determination of a transition state at atomic resolution from protein engineering data. *J. Mol. Biol.* **324**, 151–163.

Paci, E., Clarke, J., Steward, A., Vendruscolo, M. & Karplus, M. 2003 Self-consistent determination of the transition state for protein folding: application to a fibronectin type III domain. *Proc. Natl Acad. Sci. USA* **100**, 394–399.

Paci, E., Friel, C. T., Lindorff-Larsen, K., Radford, S. E., Karplus, M. & Vendruscolo, M. 2004 Comparison of the transition state ensembles for folding of Im7 and Im9 determined using all-atom molecular dynamics simulations with phi value restraints. *Proteins* **54**, 513–525.

Pande, V. S. & Rokhsar, D. S. 1998 Is the molten globule a third phase of proteins? *Proc. Natl Acad. Sci. USA* **95**, 1490–1494.

Petkova, A. T., Ishii, Y., Balbach, J. J., Antzutkin, O. N., Leapman, R. D., Delaglio, F. & Tycko, R. 2002 A structural model for Alzheimer's beta-amyloid fibrils based on experimental constraints from solid state NMR. *Proc. Natl Acad. Sci. USA* **99**, 16742–16747.

Russel, W. B. 2003 Condensed-matter physics: tunable colloidal crystals. *Nature* **421**, 490–491.

Salvatella, X., Dobson, C. M., Fersht, A. R. & Vendruscolo, M. 2004 Determination of the successive transition states on the folding of barnase: validation by double-mutant cycle experiments. (Submitted.)

Schulman, B. A., Kim, P. S., Dobson, C. M. & Redfield, C. 1997 A residue-specific NMR view of the non-cooperative unfolding of a molten globule. *Nat. Struct. Biol.* **4**, 630–634.

Selkoe, D. J. 2003 Folding proteins in fatal ways. *Nature* **426**, 900–904.

Soto, C. 2003 Unfolding the role of protein misfolding in neurodegenerative diseases. *Nat. Rev. Neurosci.* **4**, 49–60.

Stefani, M. & Dobson, C. M. 2003 Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med.* **81**, 678–699.

Tycko, R. 2004 Progress towards a molecular-level structural understanding of amyloid fibrils. *Curr. Opin. Struct. Biol.* **14**, 96–103.

Uversky, V. N. 2003 Protein folding revisited. A polypeptide chain at the folding–misfolding–nonfolding crossroads: which way to go? *Cell. Mol. Life Sci.* **60**, 1852–1871.

Vendruscolo, M. & Paci, E. 2003 Protein folding: bringing theory and experiment closer together. *Curr. Opin. Struct. Biol.* **13**, 82–87.

Vendruscolo, M., Paci, E., Dobson, C. M. & Karplus, M. 2001 Three key residues form a critical contact network in a protein folding transition state. *Nature* **409**, 641–645.

Vendruscolo, M., Dokholyan, N. V., Paci, E. & Karplus, M. 2002 Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev.* E **65**, 061910.

Vendruscolo, M., Zurdo, J., MacPhee, C. E. & Dobson, C. M. 2003*a* Protein folding and misfolding: a paradigm of self-assembly and regulation in complex biological systems. *Phil. Trans. R. Soc. A* **361**, 1205–1222.

Vendruscolo, M., Paci, E., Karplus, M. & Dobson, C. M. 2003*b* Structures and relative free energies of partially folded states of proteins. *Proc. Natl Acad. Sci. USA* **100**, 14 817–14 821.

Vendruscolo, M., Paci, E., Dobson, C. M. & Karplus, M. 2003*c* Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange. *J. Am. Chem. Soc.* **125**, 15686–15687.

Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T. & Ramakrishnan, V. 2000 Structure of the 30S ribosomal subunit. *Nature* **407**, 327–339.

Wong, K. B., Clarke, J., Bond, C. J., Neira, J. L., Freund, S. M. V., Fersht, A. R. & Daggett, V. 2000 Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding. *J. Mol. Biol.* **296**, 1257–1282.

Yau, S. T. & Vekilov, P. G. 2000 Quasi-planar nucleus structure in apoferritin crystallization. *Nature* **406**, 494–497.

Discussion

A. H. ZEWAIL (*Laboratory for Molecular Sciences, Arthur Amos Noyes Laboratory of Chemical Physics, California Institute of Technology, USA*). What about the matrix in which folding, unfolding and misfolding take place, namely water? Do you think that your idea of aggregation—rather than conformational change—can explain the prion disease mechanism?

C. M. DOBSON. Firstly, water is an extremely important factor that contributes very significantly to the thermodynamics and kinetics of folding and misfolding and one that is implicitly or explicitly included in the approach that we discussed. Secondly, I think that the existence of the prion diseases can be rationalized on essentially the same basis as the other diseases we have discussed, although the mechanism of transmission is a unique feature of these conditions and it is not yet understood in detail. It may well be that the phenomenon of seeding is a crucial factor, and indeed recent work from Prusiner's laboratory at the University of California, San Francisco (Legname *et al*. 2004), has strengthened the link with amyloid formation.

P. F. MCMILLAN (*Department of Chemistry, University College London, UK*). With reference to Angell *et al*. (2005) and your picture of folding evolving within a highly crowded environment with the cell, are these configurational landscapes determined by constant pressure or constant volume conditions?

C. M. DOBSON. The free-energy landscape of a protein is greatly modified in a crowded environment since the conformational space is restricted by the presence of other molecules (Minton 2000). Since this restriction is greater for more extended conformations, the free energy of the unfolded state is increased, and therefore the stability of the folded state is also increased by crowding. A proper description of these effects should involve the explicit treatment of the crowding agent, rather than simulations at constant volume or constant pressure.

J. F. DAVIES (*Wirral, UK*). When we cook a protein, we cause denaturation. Do we also cause aggregation or fibrillation that resists breakdown in the digestive process, and can accumulate, e.g. in the gut, and be a source of disease?

C. M. DOBSON. It is extremely likely that many proteins give rise to amyloid fibrils when heated, and indeed there is firm evidence for such a conclusion from the studies we have described here. There is, however, no evidence that such species have any link with the induction of disease, except in the case of prion disorders arising either from voluntary or involuntary cannibalism (e.g. Kuru and Bovine Spongiform Encephalopathy) or from contamination of pharmaceutical products or medical instruments, e.g. Creutzfeldt–Jakob disease from growth hormone usage or from surgical procedures (Prusiner 1997).

M. KARPLUS (*Laboratoire de Chimie Biophysique, ISIS, Université Louis Pasteur, Strasbourg, France*). If fibrous gels are the ultimate stable state of polypeptides at high concentrations, could you speculate on their role in evolution. Were there originally gels from which structured proteins evolved? What would have been the survival value of these gels?

C. M. DOBSON. It is highly tempting to speculate on the role that the amyloid state has had in the evolution of biological systems. We have thought most about the effects on protein sequences of the need to avoid conversion to this type of structure (Dobson 1999). It is interesting, however, that amyloid structures have been associated with some functional properties in bacteria and fungi in particular and we have wondered if it could have been more important in early life forms (Dobson *et al*. 2000). But such ideas are of course just speculation!

J. P. SIMONS (*Physical and Theoretical Chemistry Laboratory, University of Oxford, UK*). If the late transition state for protein folding is associated with the formation of a small set of properly disposed 'key contact' groups, does this imply that the role of the amino acid sequence is principally to provide the appropriate spacings? If so, does this imply (for the set of natural proteins) that the sequence of residues does not control their global minimum conformation? And is the role of the chaperone simply to facilitate access to the late transition state configuration, for example, by removing or lowering the intervening barriers?

C. M. DOBSON. The key residues that we have described have the role of promoting the folding process by creating a critical nucleus. The uniqueness of the folded state is achieved because the formation of the critical nucleus results in the establishment of the correct topology (Vendruscolo *et al*. 2001). The remainder of the sequence must be fully compatible with such a topology and therefore one can still say that the sequence controls the formation of the folded state.

It is now widely accepted that the folded state is fully encoded by the sequence of amino acids and that the role of chaperones is to help the proteins to escape from metastable and misfolded states that occur during the process of folding.

Additional references

Angell, C. A. 2005 Energy landscapes for cooperative processes: nearly ideal glass transitions, liquid–liquid transitions and folding transitions. *Phil. Trans. R. Soc. A* **363**, 000–000.

M. Vendruscolo and C. M. Dobson

Dobson, C. M., Ellison, G. B., Tuck, A. F. & Vaida, V. 2000 Atmospheric aerosols as prebiotic chemical reactors. *Proc. Natl Acad. Sci. USA* **97**, 11 864–11 868.

Legname, G., Baskakov, I. V., Nguyen, H.-O. B., Riesner, D., Cohen, F. E., DeArmond, S. J. & Prusiner, S. B. 2004 Synthetic mammalian prions. *Science* **305**, 673–676.

Minton, A. P. 2000 Implication of macromolecular crowding for protein assembly. *Curr. Opin. Struct. Biol.* **10**, 34–39.

Prusiner, S. B. 1997 Prion diseases and the BSE crisis. *Science* **278**, 245–251.