Available online at www.sciencedirect.com
SCIENCE DIRECT®
Methods 34 (2004) 4–14
METHODS
www.elsevier.com/locate/ymeth

# Experimental investigation of protein folding and misfolding

Christopher M. Dobson

*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EX, UK*

Accepted 5 March 2004
Available online 7 June 2004

**Abstract**

Newly synthesised proteins need to fold, often to intricate and close-packed structures, in order to function. The underlying mechanism by which this complex process takes place both in vitro and in vivo is now becoming understood, at least in general terms, as a result of the application of a wide range of biophysical and computational methods used in combination with the techniques of biochemistry and protein engineering. It is increasingly apparent, however, that folding is not only crucial for generating biological activity, but that it is also coupled to a wide range of processes within the cell, ranging from the trafficking of proteins to specific organelles to the regulation of cell growth and differentiation. Not surprisingly, therefore, the failure of proteins to fold appropriately, or to remain correctly folded, is associated with a large number of cellular malfunctions that give rise to disease. Misfolding, and its consequences such as aggregation, can be investigated by extending the types of techniques used to study the normal folding process. Application of these techniques is enabling the development of a unified description of the interconversion and regulation of the different conformational states available to proteins in living systems. Such a description proves a generic basis for understanding the fundamental links between protein misfolding and its associated clinical disorders, such as Alzheimer's disease and Type II diabetes, and for exploring novel therapeutic strategies directed at their prevention and treatment on a rational basis.
© 2004 Elsevier Inc. All rights reserved.

*Keywords:* Biological self-assembly; Molecular chaperones; Molecular evolution; Amyloid disease; Neurodegenerative disorders

## 1. Introduction

One of the most remarkable characteristics of a living system is the ability of even the most intricate of its molecular components to self-assemble into their functional states. Developing an understanding of the ways in which such processes occur will reveal not just the mechanisms of some extraordinarily complex structural transitions, but also the manner in which biological molecules have developed their present level of sophistication through the process of evolutionary selection. The most fundamental example of biological self-assembly is protein folding, the process through which disordered polypeptide chains convert into the tightly packed protein structures through which they exert their biological functions [1]. These functions include the control and regulation of essentially every chemical process on which our lives depend, as well as providing key components of virtually all the structural frameworks within our bodies [2]. Given the multitude of their roles, it is not surprising that proteins are the most abundant molecules in biology other than water, and that even the simplest organism contains about 1000 different types of proteins. The number of different proteins within ourselves is about 100 times greater, but this is still only a minute fraction of the total possible number of different sequences for a polypeptide chain of the average length (about 300 residues) of those found in higher organisms. Indeed, as there are 20 different types of amino acid, the number of possible sequences comfortably exceeds estimates of the total number of atoms in the universe. Natural proteins are therefore, a very select group of molecules, and their properties have a number of very special characteristics when compared to those of random sequences of amino acids. The ability of given sequences to fold to unique structures, and hence to generate a vast range of functions and an astonishing degree of selectivity, is a particularly important example of such a characteristic [3]. In this article,

*E-mail address:* cmd44@cam.ac.uk.

1046-2023/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2004.03.002

we shall examine how the application of a wide range of physical and chemical techniques is leading to a comprehensive understanding of the manner in which natural proteins fold. In addition, we explore how extension of such these approaches reveals that even such exquisitely designed molecules can revert to behaviour that is more typical of polypeptide chains that have not been so carefully selected.

## 2. The protein folding problem

Since the work of Christian Anfinsen in the 1960s and 1970s, it has been clear that the essential information that encodes the structure of proteins is contained within the amino acid sequence [4]. The way in which such information is encoded, however, has only recently begun to emerge, as a result of the combined application of a wide range of experimental and theoretical approaches. Of crucial importance, as is so often the case when major scientific problems catch the collective imagination, has been the development of a whole battery of new and increasingly sophisticated techniques, and of the conceptual frameworks that are needed to analyse and interpret the results of their application [5]. The reviews that follow this present article describe many of the most powerful of the methods that have been applied to study both normal and aberrant protein folding. In this article, I shall try to provide an introduction to these reviews by giving a general overview of the field and of its importance in biology and medicine.

In addition to the question of how the information for the unique native state is encoded in the amino acid sequence has been a second question, namely how a protein could find such a state in a finite time [6]. Even conservative estimates of the time required to search systematically the vast number of possible conformations that are in principle accessible to a polypeptide chain exceed by many orders of magnitude experimental measurements of the times required for typical protein molecules to fold. One of the key advances in the folding field has been the development of experimental and theoretical approaches that have resulted in a solution to this apparent paradox. In particular, a "new view" has developed in which folding is described as a stochastic search of conformational space rather than as a series of mandatory structural transitions [5,7,8]. The conceptual basis of such a mechanism is shown in Fig. 1 [3], and incorporates the generally accepted assumption that the native state of a protein is that with the lowest free energy. Fluctuations in the conformation of a polypeptide enable contacts to be made between even those residues that are very different from each other in the amino acid sequence. Because native-like interactions are on average more stable than non-native ones, a search mechanism of this type is in principle able to find the native state [9]. For natural proteins, it appears that such a mechanism can be highly efficient as the shape of the landscape (that is encoded in the sequence) is such that only a very small fraction of all possible conformations needs to be sampled during the folding process.

To begin, to understand the way in which the folding process occurs in detail, an important approach has been to investigate the most elementary steps in the conversion of random interactions into organised elements of structure. Despite the complexity of the folding process, the fundamental steps associated with it can be very fast. Major advances in investigating such steps have resulted from the development and application of a wide range of fast reaction methods [11,12]. Techniques involving both the rapid mixing of solutions (e.g., by transferring a denatured protein into a refolding buffer or a native protein into denaturing conditions), and the rapid perturbation of solution conditions (e.g., of temperature or pressure), have been refined to enable both folding and unfolding to be studied (Table 1). Remarkably, some of these experiments now have the capability to probe the behaviour of individual molecules [13]. Processes occurring on a timescale of milliseconds can now be followed routinely, and events on a submicrosecond timescale are becoming accessible using state-of-the-art equipment [12]. Using such approaches, individual α-helices have been observed to fold in as little as 100 ns, and β-turns to form in ca. 1 μs [14,15]. Indeed, the complete folding of some proteins such as small helical bundles has been observed to occur in less than 50 μs [16,17], and it appears that the "speed limit" for folding may be such as to allow folding of the simplest proteins in less than a tenth of this time. Other proteins, particularly involving extensive β-structure, may, however, take many orders of magnitude longer to fold, as we discuss below.

## 3. The folding of small proteins

A crucial next step in understanding the way in which proteins fold is to explore the manner in which the sequence defines the energy landscape. Many studies directed towards this end have focussed on proteins with less than about 100 residues, as such proteins are able to fold without the significant population of partially folded intermediates [18]. Of particular importance in this regard are studies that involve the use of protein engineering techniques to probe the role of individual residues in the folding process through site-directed mutagenesis [19,20]. Such studies enable the transition state region of the energy landscape to be probed in detail; by measurement of the relative effects of mutation on folding and unfolding rates, the relative contributions of specific residues to the stability of the transition state ensemble can be defined. Studies of a range of

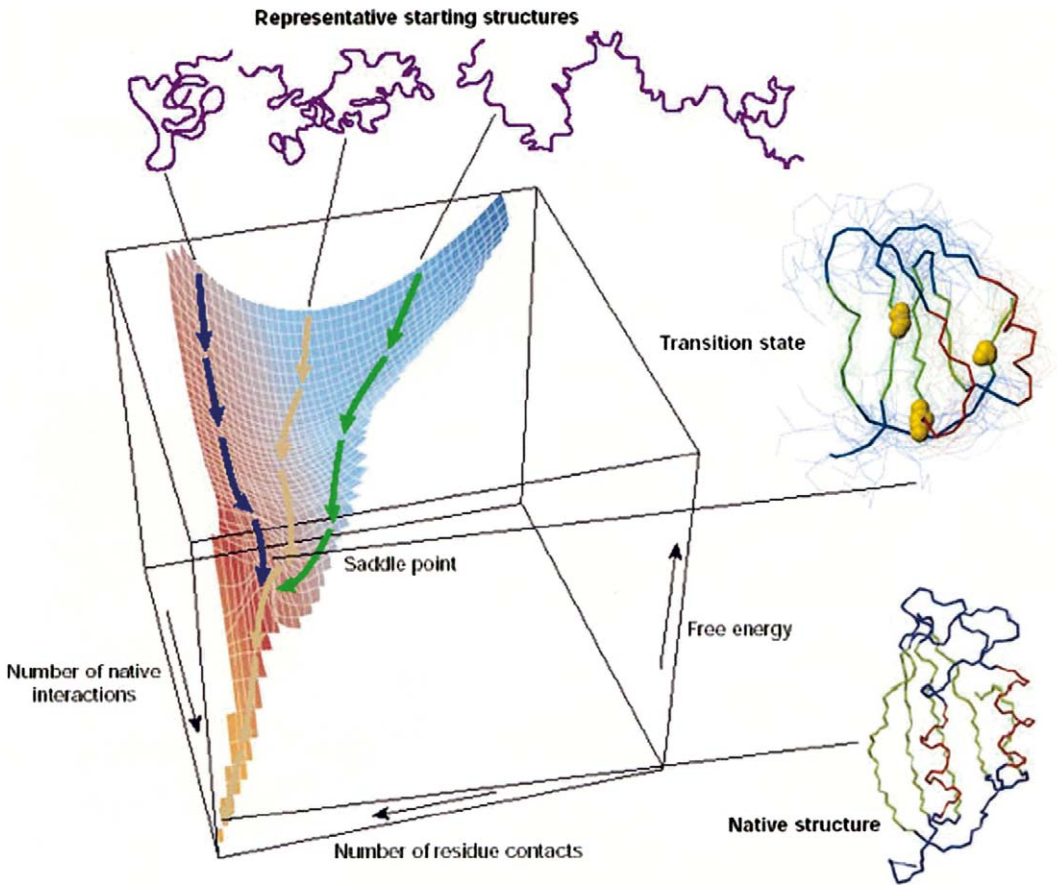


Fig. 1. A schematic energy landscape for protein folding. The surface shown here is derived from a computer simulation of the folding of a highly simplified model of a small protein [10]. Such a surface serves to "funnel" the multitude of denatured conformations to the unique native structure. The critical region on a simple surface, such as this one is the saddle point corresponding to the transition state, the barrier that all molecules must cross to be able to fold to the native state. An ensemble of structures corresponding to the experimental transition state for the folding of a small protein is indicated in the figure; this ensemble was calculated by using computer simulations constrained by experimental data from mutational studies of acylphosphatase [22]. The yellow spheres represent the three "key residues" in the structure; when these residues have formed their native-like contacts, the overall topology of the native fold is established. The structure of the native state is shown at the bottom of the surface, while at the top are indicated schematically some contributors to the distribution of unfolded states that represent the starting point for folding. Also indicated on the surface are highly simplified average trajectories for the folding of individual molecules. From [3].

small proteins have suggested that the fundamental mechanism of folding involves the formation of a folding nucleus, involving a relatively small number of residues in the protein, about which the remainder of the structure condenses rapidly [21]. Recently, this approach has been extended by combining such data with computer simulation techniques to generate 3D-structures of transition state ensembles, Fig. 1 [22]. The results suggest that a key feature of the transition state region of the energy landscape is that the overall fold or topology of the native state is already formed despite the fact that the structures are highly disordered. Once the correct topology has been defined in the transition state, the native structure will almost invariably be generated during the final stages of the folding process.

Given the emerging importance of the role of the overall topology of the protein structure [23,24], the question arises as to how it is encoded by the sequence. In addition to the experimental techniques discussed above, major advances have come from the use of computer simulation methods to explore from first principles the manner in which structure develops or is lost within the polypeptide chain during the folding and unfolding processes [25]. Such methods are particularly important as the cooperative nature of the folding of these small proteins severely limits the ability of even the most sensitive of biophysical methods to provide such information. Comparison of such simulations with the results of protein engineering experiments indicates that such simulations are now able to reproduce remarkably well the major features of these measurements [26,27]. One additional and particularly interesting example of the synergy between theory and experiment derives from the remarkable ability of AFM techniques to unfold proteins through the application of the physical force [28]. In such experiments, a protein molecule is literally

Table 1
Methods used to investigate protein folding and aggregation[a]

| Property | Technique | Measurement |
|---|---|---|
| Chain packing | Intrinsic fluorescence | The orientation and environment of (predominantly) tryptophan side chains |
| | Ultraviolet absorbance | The orientation and environment of aromatic side chains |
| | Extrinsic (ANS) fluorescence | Formation and disruption of organized hydrophobic patches and clefts |
| | Fluorescence quenching | Isolation of tryptophan side chains from hydrophilic fluorescence quenchers |
| | Cysteinyl quenching | Protection of cysteine side chains from hydrophilic reactants |
| Molecular dimensions | Fluorescence anisotropy | Tryptophan side chain mobility and overall molecular dimensions |
| | Fluorescence energy transfer | Scalar distance between tryptophan and a covalently attached fluorophore (or between two attached fluorophores) |
| | Small angle X-ray scattering/ Quasi-elastic light scattering | The average radius of gyration |
| | NMR diffusion measurements | The effective hydrodynamic radius |
| Secondary structure and persistent hydrogen bonds | Far-UV circular dichroism | Backbone conformation averaged over sequence and population |
| | Fourier transfer infra-red | Backbone conformation, hydrogen bond properties |
| | Pulse labelling NMR | Sequence specific formation of stable amide and tryptophan hydrogen bonds |
| | Pulse labelling mass spectrometry | The formation and cooperativity of persistent hydrogen bonds in discrete intermediates |
| Tertiary contacts and native structure | Biological activity | The formation of native tertiary structure at the active site |
| | Interrupted folding | The unfolding rate of discrete intermediates as a probe of their stability |
| | Near-UV circular dichroism | Formation of stable aromatic and disulphide bond tertiary contacts |
| | Real-time NMR | Formation of specific side chain tertiary contacts |
| | Protein engineering | The energetic contributions of side chains to discrete intermediates |
| | AFM/laser tweezers | Force required to unfold protein or region of protein |
| Aggregate structure | Congo red or thioflavin fluorescence | Existence of regular β-sheet structure |
| | Birefringence with Congo red | Well-defined amyloid core structure |
| | X-ray fibre diffraction | Spacings of regular elements of structure e.g. β-sheets |
| | AFM/EM | Dimensions and morphology of discrete aggregates |
| | Solid state NMR | Molecular conformation and intermolecular packing |

[a] Adapted from [11], but see also specific reference in text and the later reviews in this volume.

pulled apart by steadily increasing the applied force on the cantilever of the microscope. Because the "reaction coordinate" for such a process is relatively well-defined, simulations can be carried out that enable the experiments to be replicated in very great detail, and hence allows us to understand specific aspects of the forces through which proteins are stabilised [29].

Although the way in which the fold is encoded in the sequence is still not known in detail, approaches of the type described above suggest that the patterns of hydrophobic and hydrophilic residues, rather than the highly specific characters of the individual residues involved, play an important role [30]. It is clear, for example, that the same fold can often be generated by very different sequences of amino acid residues, although the distribution of residues of distinctive character is generally evident by comparison of such sequences, a result that underlies the method of structure prediction usually known as threading [30,31]. The distribution of residue types along the sequence appears to allow the stabilisation of native-like environments in the vicinity of the key residues in the absence of the remaining residue contacts. Such general properties of the fold also appear to be more important than characteristics such as the secondary structure or stability, although the latter undoubtedly influences aspects of the folding behaviour, as revealed most dramatically by the remarkable correlation between the folding rates of a series of small proteins and their "contact order" [32]. The latter is the average separation in sequence between residues that are in contact with each other in the native structure, and the smaller the contact order the faster the rate of folding. Such a correlation can be explained by the argument that a stochastic search process will be more efficient if the residues that form the nucleus are closer together in the sequence. The existence of this correlation also supports the concept that there could be relatively simple principles underlying the folding of proteins [33]. Further understanding of such principles will undoubtedly reveal in more detail just how the

sequence is able to encode the fold, and should increase significantly our ability both to predict the structures of proteins from their sequences and to design new sequences that encode novel folds [33,34].

## 4. The folding of larger proteins

For proteins that consist of more than about 100 residues, the folding process generally involves the population of one or more partially folded intermediates prior to the formation of the completely folded native state. There has been considerable debate over the significance of such species, whether they serve to help the protein find its correct fold efficiently or whether they are unavoidable traps that slow down folding [35,36]. Regardless of their specific role in the folding process, their existence provides a valuable source of information about the development of different aspects of the overall fold. Progress in this direction has resulted primarily from the use of the wide range of biophysical techniques that have been developed to probe fast events occurring during folding. An important characteristic of such techniques is that different methods are able to probe the development of different aspects of the native structures, as summarised in Table 1 [11]. Thus, far-UV CD spectroscopy monitors the formation of secondary structure while near-UV CD detects primarily the presence of tertiary interactions involving aromatic residues. Of particular significance is the application of NMR spectroscopy, as this technique has the potential to provide residue-specific information about the structure present at different stages of a folding reaction [37]. The ways in which such information can be extracted is now a very active area of research, and promises to revolutionise our knowledge of the folding process. It is already well established as a means to probe the persistence of individual elements of secondary structure through hydrogen exchange, an approach that is particularly valuable when combined with the complementary technique of mass spectrometry [38]. More recently, it has proved possible to probe other characteristics of the fold, including the behaviour of the side chains and the relative stabilities of different regions of globular structure. The potential for translating such data into 3D-structures is also being explored, using restrained simulation techniques analogous to those used to define the structures of the transition states of small proteins described above [39,40].

Using such techniques, it is becoming possible to begin to aspire to defining, at least in outline, experimental energy landscapes for the folding of these larger proteins [40]. Such studies have been assisted strongly by the fact that partially folded states of larger proteins can often be stabilised under equilibrium conditions, enabling time-consuming techniques such as NMR to be applied in greater detail than is possible in experiments carried out during the folding reaction itself. Simulations of the species involved are particularly valuable to establish the origins of the residual structure present at different stages of folding. The results of the various types of studies of the intermediates associated with larger proteins suggest that such proteins are likely to fold in modules, such that structure develops largely independently in different segments or domains of the structure [40–42]. It appears likely that interactions associated with key residues generate the native-like fold within such regions, and then interactions of other key residues ensure that these independent regions of structure assemble to generate the correct overall fold. Such a scenario is particularly attractive as it suggests, in accord with the principles underlying the new view of folding, that there is a common underlying mechanism for the folding of all proteins [5,15]. In addition, it suggests that highly complex structures are assembled in manageable pieces, and can readily be applied to other macromolecules such as nucleic acids, and even to the assembly of large complexes such as the ribosome.

## 5. Protein folding in the cell

Given that such progress is being made in understanding how proteins fold in vitro, one might ask, perhaps somewhat anxiously, how relevant such studies are to the events that occur in living systems. The answer that is emerging from a wide range of studies is that the underlying mechanism of folding is undoubtedly the same in vivo as in vitro—it would indeed have been remarkable if proteins had evolved so as to fold both in the laboratory and in the cell, but by different mechanisms. The cellular environment in which proteins fold following synthesis on ribosomes is of course extremely complex compared to that associated with most experiments conducted in vitro. For example, it is so densely packed with all the molecular components that are needed for its survival and replication that the macromolecular concentration can exceed 350 mg/ml [43]. While the molecules within a cell have evolved, remarkably, to be able to function correctly and often independently within such an environment, this degree of molecular crowding means that incompletely or improperly folded molecules will undoubtedly aggregate with each other or associate improperly with other cellular components. To avoid such problems, a series of auxiliary proteins have evolved to assist proteins to fold efficiently and without such complications [44,45]. These species include folding catalysts that enhance slow steps in protein folding, such as the formation of disulphide bonds and the isomerisation of peptide bonds involving proline residues, and "molecular chaperones" that act to avoid the consequences of protein misfolding and aggregation.

The key difference between protein folding in vitro and in vivo is in fact that biology has evolved mechanisms to control and regulate the entire process within living systems. In doing so, however, it is clear that biology has exploited many states of proteins other than the native one [46]. For example, unfolded and partially folded states are known to be important in events such as translocation across membranes, the trafficking of proteins to particular locations within the cell, and in targetting for destruction those proteins that have served their function and are no longer needed. In addition, some proteins are either partly or completely "natively unfolded," that is they do not adopt unique globular structures even under physiological conditions, at least in the absence of specific partners such as cofactors, other proteins or nucleic acids [47]. One can consider that biology has done to the folding process, using species such as molecular chaperones and folding catalyst, what it has done to chemical processes by means of enzymes and the regulation of gene expression, i.e., to generate specificity and control within the biological environment. In addition, it is clear from studies using a wide variety of biochemical techniques that living systems have not relied solely on such species to regulate protein folding. They have also developed remarkably sophisticated mechanisms of quality control to check, whether proteins are correctly folded, and to target for destruction any molecules that do not come up to scratch. The best-understood quality control mechanisms are in the endoplasmic reticulum, the major folding compartment of eukaryotic cells [48], and the best-characterised degradation mechanisms are part of the "unfolded protein response" that involves ubiquitination of proteins destined for disposal, followed by their destruction in the cytosol by the proteosome [49].

## 6. Protein misfolding and disease

Because of the key role played by protein folding and unfolding within the cell, it is inevitable that mistakes in folding will give rise to the malfunctioning of biological processes and hence to disease. A large number of diseases are already associated with misfolding and more are being added each year [50,51]. In some disorders, such as cystic fibrosis, the ability of a protein to fold correctly is reduced by familial mutations (in this case, in the gene encoding a membrane protein involved in chloride ion transport) resulting in a reduction in the level, or in some cases the complete absence of a key functional species. In other cases, such as the amyloid diseases, failure to fold or to remain correctly folded results in the aggregation and deposition of proteins in one or more types of tissue [52,53]. Here, pathological symptoms can result from the sheer quantity of protein aggregates (sometimes kilograms) found in organs such as the liver or spleen, in systemic diseases associated, for example, with mutations in the lysozyme gene. Alternatively, symptoms can result from a "gain of toxic function" associated with the aggregates; this latter situation is thought to be the primary origin of several neurodegenerative disorders, notably Alzheimer's and Parkinson's diseases [54,55]. In the amyloid diseases, the deposits contain intractable aggregates, often fibrillar in nature, which have a highly characteristic structure based on the stacking of regular arrays of β-strands into β-sheets, as we discuss below. Each disease, and some 20 are now identified, is associated primarily with one protein or fragment of a protein that forms the core structure of the fibrillar deposits [56].

It was once assumed that the formation of this type of pathological aggregate resulted from the existence of aberrant sequence motifs in a protein sequence that explicitly coded for the amyloid core structure. More recently, however, it has been found that many proteins without any connection with disease, including common proteins such as myoglobin and homopolymers such as polythreonine, can give rise to fibrillar structures with all the characteristics of those found associated with the clinical amyloidoses [57–59]. These findings have led to the suggestion that the ability to form amyloid fibrils is a generic property of polypeptide chains [60]. The existence of such a property can be rationalised on the grounds that the stability of the amyloid core structure results primarily from the hydrogen bonds that link the β-strands together involve the amide and carbonyl groups of the polypeptide main chain. As the main chain is common to all polypeptides (except for those containing the amino acid proline that does not possess an amide hydrogen atom and therefore, cannot hydrogen bond to carbonyl groups), the generic ability of peptides and proteins to form such a structure can readily be rationalised [60,61]. In essence, in amyloid fibrils the main chain dominates the structure and the side chains are incorporated in the most favourable manner consistent with this requirement. By contrast, in the evolved globular structures the close-packing of the side chains determines the fold, and the main chain is incorporated in the most favourable manner. The latter generally involves the formation of regions of helices and sheets as these secondary structure elements enable hydrogen bonds to form and stabilise the structure [2]. Differences do exist in the details of amyloid structures formed by different sequences, such as the fraction of the polypeptide chain incorporated into the core structure, the lengths of the β-strands, and the way that different protofilaments are assembled into the complete fibril. Such relatively minor variations in structure are likely to depend on the sequence and the way specific side chains interact with each other, rather than on any fundamental difference in the underlying core structure [62].

## 7. The formation and characterisation of amyloid structures

A crucial aspect of the investigation of the conversion of normally soluble proteins into amyloid deposits is to define the structures of the aggregates and the manner in which they are formed. Amyloid deposits all show characteristic optical behaviour, such as green birefringence, on binding certain dye molecules such as Congo red, an observation that is attributed to the existence of regularly spaced arrays of β-sheets [63]. Despite, their regularity the aggregates do not form 3D crystals and their structural investigation has involved the application of a wide range of biophysical techniques, each of which can provide information of different aspects of their structures, Table 1. The existence of a high degree of β-structure is clearly evident through studies using CD and FT-IR spectroscopy [64]. EM analysis shows that the fibrillar structures typical of many of the aggregates have a similar appearance, being long (often microns in length) and unbranched [56]. The fibrils are commonly about 10 nm in diameter, and often show evidence that they consist of between 2 and 6 protofibrils that are twisted around each other. X-ray fibre diffraction techniques indicate that the organised core has a "cross-β" structure in which sheets are assembled from β-strands that run perpendicular to the fibrillar axis [56]. Regardless of their differences in amino acid sequences, and of their structures in their biologically active states, different peptides and proteins form amyloid fibrils that are remarkably similar in these essential features.

Despite the fact that no structure of an amyloid fibril has been determined at atomic resolution, increasingly detailed models based on data from techniques such as X-ray fibre diffraction [56], cryo-electron microscopy [65,66], and solid-state NMR spectroscopy [67,68] are emerging. These structures support the conclusion that the fibrils are likely to be stabilised primarily by interactions involving the common polypeptide main chain. Biophysical studies have also provided insight into the manner in which the fibrils are formed from their soluble precursors, and again the general conclusions from studies of very different systems appear to be remarkably similar [54,69,70]. The first phase in amyloid formation seems to involve the formation of soluble oligomeric species whose structures are likely to be rather disordered. The first species visible by electron or atomic force microscopy generally resemble small bead-like structures, sometimes linked together and often described as amorphous or micellar structures. These early prefibrillar species then transform into species with more distinctive morphologies, often called protofilaments or protofibrils. These structures are commonly short, thin, and often curly, fibrillar species that are thought to assemble into mature fibrils, perhaps sometimes by simple lateral association accompanied by a degree of structural reorganisation. By this stage in the assembly process, the structures are highly organised and are often relatively inert and highly resistant to proteases [71]. By contrast, the early aggregates appear to expose to the solution a variety of regions of the polypeptide chain that are normally buried in a globular state. In some cases, however, these early species adopt quite distinctive structures, including well-defined annular species, whose role in determining their interactions with cell membranes could be particularly significant [54,72].

Even though it appears that the ability to convert into amyloid fibrils is in principle universal for polypeptides (except for polyproline-based sequences), the propensities to convert into this structure from a disordered state vary very substantially. Some sequences are, for example, inherently more soluble than others, and some have a greater tendency than others to form β-structure. Recent studies, particularly based on protein engineering techniques similar to those used to study protein folding, suggest that the propensities to form amyloid structures can be rationalised to a remarkably high degree by a consideration of the physicochemical properties of a polypeptide chain—such as charge, hydrophobicity, and secondary structure propensity—along with a consideration of the distribution of hydrophobic and polar residues [73–76]. The latter influences the tendency to form structures such as β-sheets in which the interactions between side chains on adjacent β-strands influence their stability [77]. Indeed, the fact that patterns of residue types that favour the formation of amyloid-like structures are less common in natural proteins than expected provides evidence that evolutionary selection has avoided sequences that tend to promote their formation [78]. Such selection, along with the ability of many sequences to fold under physiological conditions to globular structures and hence to bury the aggregation prone main chain along with hydrophobic residues within the close-packed structure, has undoubtedly been a major factor in enabling biology to avoid the conversion of its functional polypeptide chains into intractable fibrils under normal circumstances [55,79]. Despite such safeguards, there is an inherent tendency for proteins to revert into the generic, or "primordial," amyloid structure if the normal homeostasis of an organism is disrupted. Indeed, it is possible to bring together ideas as to the fundamental origin of the various amyloid disorders from this underlying principle.

## 8. A unified view of folding and aggregation

The ideas encapsulated in this article can be summarised in the schematic representation shown in Fig. 2

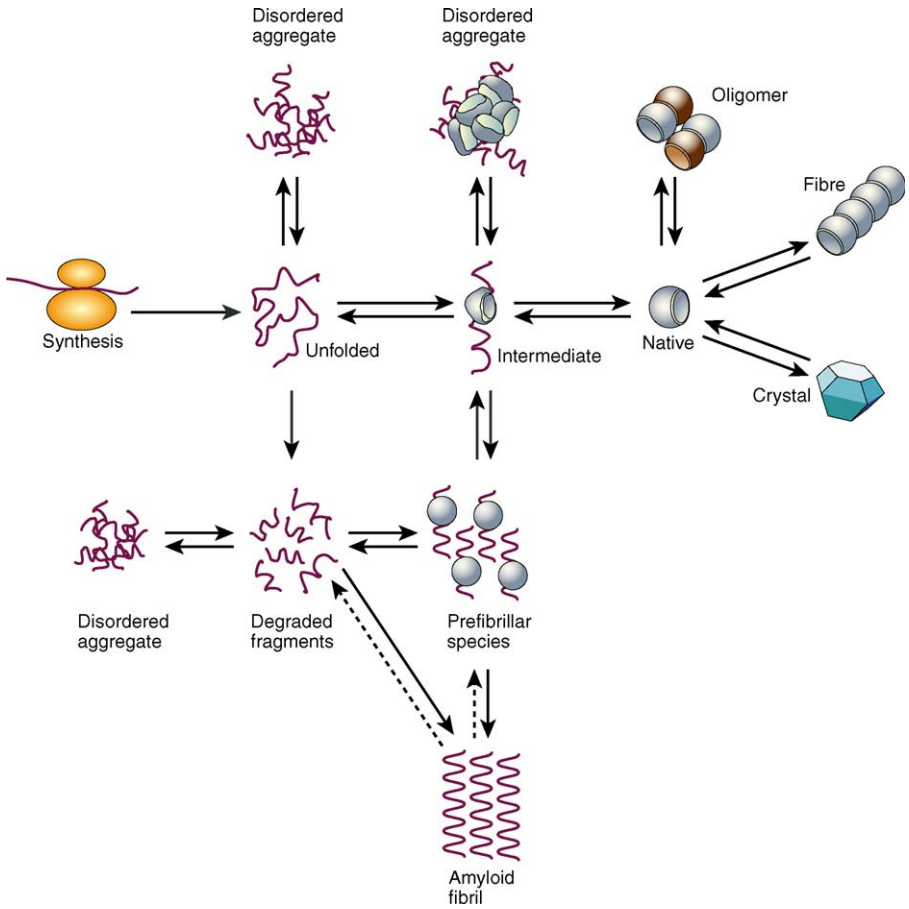


Fig. 2. A unified view of some of the types of structure that can be formed by polypeptide chains. An unstructured chain, for example, newly synthesized on a ribosome, may fold to a native structure, perhaps via one or more partially folded intermediates. It can, however, experience other fates such as degradation or aggregation. An amyloid fibril is just one form of aggregate, but it is unique in having a highly organized "misfolded" structure. Other assemblies, including functional multi-protein complexes and natural protein fibres, contain natively folded molecules, as do the protein crystals produced in vitro for X-ray diffraction studies of their structures. The populations and interconversions of the various states are determined by their relative thermodynamic and kinetic stabilities under any given conditions. In living systems, however, transitions between the different states are highly regulated by control of the environment, and by the presence of molecular chaperones, proteolytic enzymes, and other factors. Failure of such regulatory mechanisms is likely to be a major factor in the onset of misfolding diseases. From [3].

[3]. We conclude that a variety of distinct states are in principle accessible to a protein following its synthesis. The extension of the techniques of molecular biology, biophysics, and computation used to study folding to investigate misfolding and aggregation is enabling the development of a unified and comprehensive view of the behaviour of proteins in living systems. The particular state that is adopted under specific conditions by a given protein depends on the relative thermodynamic stabilities of the various accessible conformations and on the kinetics of their interconversion. Amyloid fibrils are just one of the types of aggregate that can be formed by a protein, although this species is of particular interest because it appears to have unique kinetic stability because of its highly organised hydrogen-bonded structure. It is therefore likely that, once it forms, it can persist for long periods of time, allowing a progressive build-up of proteinaceous deposits in tissue. The presence of such aggregates can also enhance the rate of conversion of the normally soluble form of the same protein into fibrils through a seeding mechanism [80]. It is therefore not surprising that living systems have normally avoided the deliberate formation of such material even as a structural scaffold for which at first sight it appears to be ideally suited. Nevertheless, there is increasing evidence that the unique properties of amyloid structures have been exploited for specialised (and carefully controlled) purposes by a wide range of organisms including bacteria, fungi, and even mammals [81–83].

It was noted above that there is evidence that evolutionary selection has tended to avoid sequences that strongly favour amyloid formation. But it is clear that the conformational states of biological systems are also highly regulated through the action of molecular chaperones and degradation mechanisms. Such regulation is analogous to that exerted on chemical reactions by enzymes [1]. The ideas encapsulated in Fig. 2, therefore,

act as a framework for understanding the fundamental events that underlie misfolding diseases. For example, many amyloid diseases are familial (e.g., lysozyme amyloidosis) as the mutations can increase the population of unfolded or partially unfolded species (usually by destabilising the native state or increasing the propensities of such species to aggregate) [84,85]. Other amyloid disorders (e.g., Alzheimer's disease) are associated with the accumulation of deposits whose major components are fragments of proteins produced by proteolytic cleavage or aberrant processing; such fragments are unable to fold into cooperative aggregation-resistant structures. Pathogenic mutations generating familial forms of such diseases can result from the increased tendency for such species to form or, when formed, to aggregate [85]. Other diseases (e.g., the transmissible spongiform encephalopathies) are likely to result at least in part from the increase in the rate of aggregation through seeding generated by ingestion of pre-formed aggregates, for example, through cannibalistic practices or contaminated surgical instruments or pharmaceuticals [80,86]. But perhaps, the most significant forms of amyloid diseases are those generally termed sporadic, many of which are associated with old age. It appears likely that the greatly increased prevalence of these diseases in the modern era results from the fact that more of us live longer as a result of improved hygiene and medical procedures. Evolutionary pressure has done a tremendous job in preventing such events during the time required to pass on our genes and to provide initial protection to our offspring, but there is little or no pressure to enhance the ability of our proteins or their "housekeeping" systems to operate in old age [3,61]. In such circumstances, it seems likely that we are simply seeing in the prevalence of amyloid disease the results of the inherent tendency of proteins to revert to their "primordial" structure when not prevented from so doing.

## 9. Looking to the future

The fundamental origin of amyloid disorders appears, therefore, to arise from the increased tendency of proteins to aggregate under circumstances such as old age. It is therefore, apparent that an increased understanding of protein folding and misfolding is crucial for the rational development of therapeutic strategies directed against these diseases. Already, a range of approaches is being developed with such ideas in mind, for example, to stabilise native states of amyloidogenic proteins, to lower the levels of aggregation prone species, and to inhibit selectively the aggregation process that results in the formation of the amyloid structure [87,88]. Of particular importance in such enterprises is an understanding of the specific links between the aggregation process and pathological behaviour. As we have mentioned above, in some amyloid disorders, such as systemic lysozyme amyloidosis, very large quantities of protein can be deposited in vital organs such as the kidney, liver, and spleen. Such deposits are likely to cause rupture or other malfunctions simply from the presence of such quantities of proteinaceous aggregate. In other amyloid disorders, notably the neurodegenerative diseases, it appears that the primary symptoms are the result of toxicity associated with the aggregates. Recent evidence suggests that the species formed during the early stages of amyloid formation, the precursors of the mature fibrils (see above), are the most dangerous species, generating cellular damage or cell death [54,55]. Such toxicity could arise simply from the accessibility of hydrophobic patches that might interact inappropriately with cellular species such as membranes, or it could arise from the specific properties of the annular species discussed to disrupt the ion balance across membranes [79].

Recent experiments suggest that the toxicity of these early aggregates is not limited to those proteins associated with clinical manifestations of disease, but could also be a property of such species formed from proteins that are not associated with specific known diseases. [89]. In support of this idea, antibodies raised against the early aggregates of one type of protein can cross-react with similar species from other proteins, suggesting that their characteristic properties are generically similar [90]. It is becoming clear, therefore, that the cellular housekeeping mechanisms (such as molecular chaperones and targetted degradation mechanisms) are essential for the ability of all cellular systems to function effectively, to neutralise the toxic effects of any protein aggregates that chance to form during the normal functioning of the cell [79,89]. Such ideas, still in their infancy, are particularly exciting not just because they provide an opportunity to understand the underlying molecular mechanism of the whole family of amyloid diseases, but also, because they raise the possibility that there are generic solutions for a generic form of disease. Perhaps, for example, it is even possible that the natural housekeeping mechanisms can be enhanced in a way that protects against all these disorders. If so, one of the principal goals of the early alchemists, to produce an elixir of life, might, at least in a small way, be realisable in the forseeable future! More prosaically, an enhanced understanding of protein folding, and the prevention of misfolding, will undoubtedly bring great intellectual satisfaction and novel insight into the nature and evolution of biological molecules, and also generate new ideas for the biotechnology and pharmaceutical industries, and for medical science. The key to the development of such an understanding is undoubtedly the further development and application of the type of methodologies described in this volume.

## Acknowledgments

The ideas in this article, which is based in part on an earlier review [3], have emerged from extensive discussions with outstanding students, research fellows, and colleagues over many years. I am most grateful to all of them; they are too numerous to mention here but the names of many appear in the list of references. The research of CMD is supported by Programme Grants from the Wellcome Trust and the Leverhulme Trust, as well as by the BBSRC, EPSRC, and MRC.

## References

[1] R.H. Pain (Ed.) Protein Folding, second ed., Oxford University Press, Oxford, 2000.
[2] C. Branden, J. Tooze, Introduction to Protein Structure, second ed., Garland Publishing, New York, 1999.
[3] C.M. Dobson, Nature 426 (2003) 884–890.
[4] C.B. Anfinsen, Science 181 (1973) 223–230.
[5] C.M. Dobson, A. Sali, M. Karplus, Angew. Chem. Int. Ed. Engl. 37 (1998) 868–893.
[6] M. Karplus, Fold. Des. 2 (1997) 569–576.
[7] P.G. Wolynes, J.N. Onuchic, D. Thirumalai, Science 267 (1995) 1619–1623.
[8] K.A. Dill, H.S. Chan, Nat. Struct. Biol. 4 (1997) 10–19.
[9] R.L. Baldwin, Nature 369 (1994) 183–184.
[10] A.R. Dinner, A. Sali, L.J. Smith, C.M. Dobson, M. Karplus, Trends Biochem. Sci. 25 (2000) 331–339.
[11] K. Plaxco, C.M. Dobson, Curr. Opin. Struct. Biol. 6 (1996) 630–636.
[12] R.H. Callendar, R.B. Dyer, R. Gilmanshin, W.H. Woodruff, Annu. Rev. Phys. Chem. 49 (1998) 173–202.
[13] B. Schuler, E.A. Lipman, W.A. Eaton, Nature 419 (2002) 743–747.
[14] W.A. Eaton, V. Munoz, P.A. Thompson, E.R. Henry, J. Hofrichter, Acc. Chem. Res. 31 (1998) 745–753.
[15] C.D. Snow, H. Nguyen, V.S. Pande, M. Gruebele, Nature 420 (2002) 102–106.
[16] W.Y. Yang, M. Gruebele, Nature 423 (2003) 193–197.
[17] U. Mayor, N.R. Guydoch, C.M. Johnson, S. Sato, G.S. Jas, S.M.V. Freund, J.G. Grossman, D.O.V. Alonso, V. Daggett, A.R. Fersht, Nature 421 (2003) 863–867.
[18] S.E. Jackson, Fold. Des. 3 (1998) R81–R91.
[19] A. Matouschek, J.T. Kellis Jr, L. Serrano, A.R. Fersht, Nature 340 (1989) 122–126.
[20] A.R. Fersht, Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding, W.H. Freeman, NewYork, 1999.
[21] A.R. Fersht, Proc. Natl. Acad. Sci. USA 97 (2000) 1525–1529.
[22] M. Vendruscolo, E. Paci, C.M. Dobson, M. Karplus, Nature 409 (2001) 641–646.
[23] D.A. Debe, M.J. Carlson, W.A. Goddard III, Proc. Natl. Acad. Sci. USA 96 (1999) 2596–2601.
[24] D.E. Makarov, K.W. Plaxco, Prot. Sci. 12 (2003) 17–26.
[25] J.E. Shea, C.L. Brooks, Annu. Rev. Phys. Chem. 52 (2001) 499–535.
[26] C.M. Dobson, M. Karplus, Curr. Opin. Struct. Biol. 9 (1999) 92–101.
[27] A.R. Fersht, V. Daggett, Cell 108 (2002) 573–582.
[28] M. Rief, M. Gautel, F. Oesterhelt, J.M. Fernandez, H.E. Gaub, Science 276 (1997) 1109–1112.
[29] S.B. Fowler, T.J. Rutherford, A. Steward, E. Paci, M. Karplus, J.E. Clarke, J. Mol. Biol. 322 (2002) 841–849.
[30] J.U. Bowie, R. Luthy, D. Eisenberg, Science 253 (1991) 164–170.
[31] A.V. Finkelstein, B.A. Reva, Nature 351 (1991) 497–499.
[32] K.W. Plaxco, K.T. Simons, D. Baker, J. Mol. Biol. 277 (1998) 985–994.
[33] D. Baker, Nature 405 (2000) 39–42.
[34] B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, D. Baker, Science 302 (2003) 1364–1368.
[35] A.P. Capaldi, C. Kleanthous, S.E. Radford, Nat. Struct. Biol. 9 (2002) 209–216.
[36] H. Roder, W. Colon, Curr. Opin. Struct. Biol. 7 (1997) 15–28.
[37] N.A.J. van Nuland, V. Forge, J. Balbach, C.M. Dobson, Acc. Chem. Res. 31 (1998) 733–780.
[38] A. Miranker, C.V. Robinson, S.E. Radford, R.T. Aplin, C.M. Dobson, Science 262 (2003) 896–900.
[39] M. Vendruscolo, E. Paci, M. Karplus, C.M. Dobson, Proc. Natl. Acad. Sci., USA 100 (2003) 14817–14821.
[40] M. Vendruscolo, E. Paci, C.M. Dobson, M. Karplus, J. Am. Chem. Soc. 125 (2003) 15686–15687.
[41] A.R. Panchenko, Z. Luthey-Schulten, P.G. Wolynes, Proc. Natl. Acad. Sci. USA 93 (1996) 2008–2013.
[42] F. Khan, J.I. Chuang, S. Gianni, A.R. Fersht, J. Mol. Biol. 333 (2003) 169–186.
[43] R.J. Ellis, Curr. Opin. Struct. Biol. 11 (2001) 114–119.
[44] M.-J. Gething, J. Sambrook, Nature 355 (1992) 33–45.
[45] F.U. Hartl, M. Hayer-Hartl, Science 295 (2002) 1852–1858.
[46] S.E. Radford, C.M. Dobson, Cell 97 (1999) 291–298.
[47] A.K. Dunkar, Z. Obradovic, Nat. Biotechnol. 19 (2001) 805–806.
[48] R. Sitia, I. Braakman, Nature 426 (2003) 891–894.
[49] A.L. Goldberg, Nature 426 (2003) 895–899.
[50] M.B. Pepys, in: D.J. Weatherall, J.G. Ledingham, D.A. Warrel (Eds.), The Oxford Textbook of Medicine, third ed., Oxford University Press, Oxford, 1995, pp. 1512–1524.
[51] C.M. Dobson, Phil. Trans. R. Soc. Lond. B 356 (2001) 133–145.
[52] D.J. Selkoe, Nature 426 (2003) 900–904.
[53] E.H. Koo, P.T. Lansbury Jr., J.W. Kelly, Proc. Natl. Acad. Sci. USA 96 (1999) 9989–9990.
[54] B. Caughey, P.T. Lansbury Jr., Annu. Rev. Neurosci. 26 (2003) 267–298.
[55] D.M. Walsh, I. Klyubin, J.V. Fadeeva, W.K. Cullen, R. Anwyl, M.S. Wolfe, M.J. Rowan, D.J. Selkoe, Nature 416 (2002) 535–539.
[56] M. Sunde, C.C.F. Blake, Adv. Protein Chem. 50 (1997) 123–159.
[57] F. Chiti, P. Webster, N. Taddei, A. Clark, M. Stefani, G. Ramponi, C.M. Dobson, Proc. Natl. Acad. Sci. USA 96 (1999) 3590–3594.
[58] M. Fandrich, M.A. Fletcher, C.M. Dobson, Nature 410 (2001) 165–166.
[59] M. Fändrich, C.M. Dobson, EMBO J. 21 (2002) 5682–5690.
[60] C.M. Dobson, Trends Biochem. Sci. 24 (1999) 329–332.
[61] C.M. Dobson, Nature 418 (2002) 729–730.
[62] A. Chamberlain, C.E. MacPhee, J. Zurdo, L.A. Morozova-Roche, H.A.O. Hill, C.M. Dobson, J. Davis, Biophys. J. 79 (2000) 3282–3293.
[63] W.E. Klunk, J.W. Pettigrew, D.J. Abraham, J. Histochem. Cytochem. 37 (1989) 1273–1282.
[64] M. Bouchard, J. Zurdo, E.J. Nettleton, C.M. Dobson, C.V. Robinson, Protein Sci. 9 (2000) 1960–1967.
[65] J.L. Jimenez, J.I. Guijarro, E. Orlova, J. Zurdo, C.M. Dobson, M. Sunde, H.R. Saibil, EMBO J. 18 (1999) 815–821.
[66] J.L. Jimenez, E.J. Nettleton, M. Bouchard, C.V. Robinson, C.M. Dobson, H.R. Saibil, Proc. Natl. Acad. Sci. USA 99 (2002) 9196–9201.
[67] A.T. Petkova, Y. Ishii, J.J. Balbach, O.N. Antzutkin, R.D. Leapman, F. Delaglio, R. Tycko, Proc. Natl. Acad. Sci. USA 99 (2002) 16742–16747.

[68] C. Jaroniec, C.E. MacPhee, V.S. Bajaj, M.T. McMahon, C.M. Dobson, R.G. Griffin, Proc. Natl. Acad. Sci. USA 101 (2004) 711–716.
[69] D.K. Wilkins, C.M. Dobson, M. Groβ, Eur. J. Biochem. 267 (2000) 2609–2616.
[70] G. Bitan, M.D. Kirkitadze, A. Lomakin, S.S. Vollers, G.B. Benedek, D.B. Teplow, Proc. Natl. Acad. Sci. USA 100 (2003) 330–335.
[71] P. Polverino de Laureto, N. Taddei, E. Frare, C. Capanni, S. Constantini, J. Zurdo, F. Chiti, C.M. Dobson, A. Fontana, J. Mol. Biol. 334 (2003) 129–141.
[72] H.A. Lashuel, D. Hartley, B.M. Petre, T. Walz, P.T. Lansbury Jr., Nature 418 (2002) 291.
[73] V. Villegas, J. Zurdo, V.V. Filamonov, F.X. Aviles, C.M. Dobson, L. Serrano, Protein Sci. 9 (2000) 1700–1708.
[74] M. Lopez de la Paz, K. Goldie, J. Zurdo, E. Lacrois, C.M. Dobson, A. Hoenger, L. Serrano, Proc. Natl. Acad. Sci. USA 99 (2002) 16052–16057.
[75] F. Chiti, M. Stefani, N. Taddei, G. Ramponi, C.M. Dobson, Nature 424 (2003) 805–808.
[76] K.F. DuBay, F. Chiti, J. Zurdo, C.M. Dobson, M. Vendruscolo, J. Mol. Biol., in press.
[77] M.W. West, W. Wang, J. Patterson, J.D. Mancias, J.R. Beasley, M.H. Hecht, Proc. Natl. Acad. Sci. USA 96 (1999) 11211–11216.
[78] B.M. Broome, M.H. Hecht, J. Mol. Biol. 296 (2000) 961–968.
[79] M. Stefani, C.M. Dobson, J. Mol. Med. 81 (2003) 678–699.
[80] J.D. Harper, P.T. Lansbury Jr., Annu. Rev. Biochem. 66 (1997) 385–407.
[81] H.L. True, S.L. Lindquist, Nature 407 (2000) 477–483.
[82] M.R. Chapman, L.S. Robinson, J.S. Pinkner, R. Roth, J. Heuser, M. Hammar, S. Normark, S.J. Hultgren, Science 295 (2002) 851–855.
[83] J.W. Kelly, W.E. Balch, J. Cell Biol. 161 (2003) 461–462.
[84] J.W. Kelly, Curr. Opin. Struct. Biol. 8 (1998) 101–106.
[85] D.R. Booth, M. Sunde, V. Bellotti, C.V. Robinson, W.L. Hutchinson, P.E. Fraser, P.N. Hawkins, C.M. Dobson, S.E. Radford, C.C.F. Blake, M.B. Pepys, Nature 385 (1997) 787–793.
[86] S. Prusiner, Science 278 (1997) 245–251.
[87] C.M. Dobson, Nat. Rev. Drug Disc. 2 (2003) 154–160.
[88] F.E. Cohen, J.W. Kelly, Nature 426 (2003) 905–909.
[89] M. Bucciantini, E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C.M. Dobson, M. Stefani, Nature 416 (2002) 507–511.
[90] R. Kayed, E. Head, J.L. Thompson, T.M. McIntire, S.C. Milton, C.W. Cotman, C.G. Glabe, Science 300 (2003) 486–489.