Quantitative approaches to defining normal and aberrant protein homeostasis

Michele Vendruscolo* and Christopher M. Dobson*

Received 23rd March 2009, Accepted 17th April 2009
First published as an Advance Article on the web 28th July 2009
DOI: 10.1039/b905825g

Protein homeostasis refers to the ability of cells to generate and regulate the levels of their constituent proteins in terms of conformations, interactions, concentrations and cellular localisation. We discuss here an approach in which physico-chemical properties of proteins and their environments are used to understand the underlying principles governing this process, which is crucial in all living systems. By adopting the strategy of characterising the origins of specific diseases to inform us about normal biology, we are bringing together methods and concepts from chemistry, physics, engineering, genetics and medicine. In particular, we are using a combination of *in vitro*, *in silico* and *in vivo* approaches to study protein homeostasis through the analysis of the effects that result from its perturbation in a select group of specific proteins, from either amino acid mutations, or changes in concentration and solubility, or interactions with other molecules. By developing a coherent and quantitative description of such phenomena, we are finding that it is possible to shed new light on how the physical and chemical properties of the cellular components can provide an understanding of the normal and aberrant behaviour of living systems. Through such an approach it is possible to provide new insights into the origin and consequences of the failure to maintain homeostasis that is associated with neurodegenerative diseases, in particular, and the phenomenon of ageing, in general, and hence provide a framework for the rational design of therapeutic approaches.

Introduction

One of the essential characteristics of living systems is the ability of their molecular components to self-assemble into functional structures.[1,2] Equally important, however, is the way in which the processes leading to this organisation are balanced within the cellular environment through the mechanism of homeostasis.[3–7] Of central importance in the study of this mechanism is to focus specifically on proteins, since these are the molecules that enable, regulate and control essentially all chemical processes on which life depends. In order to function, the large majority of our proteins need to fold into a specific three-dimensional structure.[4,8–10] Indeed, the wide variety of highly specific structures that results from protein folding, and which serve to bring key functional groups into close proximity, has enabled living systems to develop astonishing diversity and selectivity in their underlying chemical processes by using a common set of just twenty building blocks—the amino acids.[11] Much research has addressed the fundamental mechanism of protein folding through a combination of *in vitro* and *in silico* studies, and we now have considerable understanding at a molecular level of the

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, UK CB2 1EW. E-mail: mv245@cam.ac.uk; cmd44@cam.ac.uk

fundamental principles underlying this complex process.[8,12–14] The next challenge is to relate this information to events occurring in living systems.

As well as simply generating biological activity, however, we now know that protein folding in living systems takes place in a complex environment, and as a polypeptide chain emerges from the ribosome on which it has been synthesised, it interacts with a wide range of ancillary molecules including molecular chaperones.[4,15,16] Much less is known about such events at a molecular level and a primary objective in protein science is to extend the studies of folding from the test tube to the cell, and to understand how this process takes place in the cellular environment. Moreover, it is clear that protein folding and unfolding are closely coupled to many other biological processes ranging from the trafficking of molecules to specific cellular locations to the regulation of the growth and differentiation of cells.[7,10] In addition, only correctly folded proteins have the ability to remain soluble in crowded biological environments and to interact selectively with their natural partners.[10,15] The manner in which proteins are able to maintain such homeostasis is a subject of central interest in molecular biology.

Given the tremendous importance of protein folding, it is not surprising that the failure of proteins to fold correctly, or to remain correctly folded, is the origin of a wide variety of pathological conditions, including cystic fibrosis, α1-antitrypsin deficiency and Alzheimer's disease.[10,17–20] In many of these diseases proteins self-assemble in an aberrant manner into large molecular aggregates, including amyloid fibrils. Considerable attention has been devoted to exploring the nature and origin of such disorders from a structural viewpoint and to understanding the manner in which the balance between normal and aberrant conformational transitions can be perturbed.[20] Several studies have involved *in vitro* studies coupled with computer simulations,[20–23] and many others have been concerned with the goal of relating processes studied in atomic detail in the test tube to their quantitative effects in living systems.[24–26] Moreover, recent findings suggest that further developments in this area could have much more general relevance to understanding the way in which well-established physical and chemical principles can provide new insights into the apparent complexity of biology.[27]

The discovery of the common existence of amyloid and amyloid-like states is of unique importance in understanding the nature of biological systems because it reveals that there is an alternative stable and highly ordered state, accessible essentially to all proteins, in addition to the native one;[10,28,29] this observation has profound implications in diverse fields ranging from medicine to materials science. Because the structural interactions within the amyloid state and the native state are similar—although the latter are largely intramolecular whilst the former also include strong intermolecular contributions—the stability of the native and amyloid states can be comparable.[30] There is thus a competition between the two states that results in normal or aberrant biological behaviour depending on whether the native or the aggregated state is populated.[10,28,29] More generally, the maintenance of the correct balance in the populations of different states of proteins, one facet of protein homeostasis, is of great significance, as even marginal alterations in such populations can result in disease in the long term.[7,27] Indeed, it has been recently realised that the limit to the safe concentration of proteins in living systems is likely to be reached when the amyloid state becomes more stable than the native state.[27]

It is therefore of great importance to complement the well-established characterisation of the structure, folding and stability of native states with a similar analysis of the structure, assembly and stability of other states—ranging from unfolded and partially folded species, including natively unfolded states, to aggregated species such as amyloid fibrils. This is one of the main thrusts of our own work, together with the exploration of the effects of the balance between the normal and aberrant states of proteins in living organisms such as the *Drosophila* model system, which we believe will inform us on the origins of amyloid-related disease and hence more generally on the mechanism of protein homeostasis.

A conceptual framework for understanding protein homeostasis

It is becoming clear that the interplay between the various states of different protein molecules creates a highly complex system, whose behaviour determines whether a living organism functions in a normal or aberrant manner, and yet, as with other complex systems,[31,32] may be determined by the combination of relatively simple underlying processes.[3,6,33] This complexity[6] underlies the phenomena now often referred to as protein homeostasis or "proteostasis"[7] perhaps in a similar manner that, for example, individual organisms interact[34] in an ecosystem. The investigation of this particular class of biological molecules could therefore potentially shed a great deal of light on more general questions of the design and evolution of biological molecules and the environments in which they function. Such information lies at the heart both of understanding the molecular aspects of the phenomenon of life and of rational approaches to molecular medicine.

In this paper we present a strategy for describing and understanding in a coherent manner the behaviour of proteins in living systems, including their folding, misfolding and assembly processes. Our approach is primarily based on five technical and conceptual developments that have recently been made in protein science:

(1) The ability to describe quantitatively, by a combination of experimental and computational approaches, the often disordered and dynamic structures of the multiple states of proteins on which their biological behaviour depends.[35–38]

(2) New ideas about protein aggregation,[10,28] including the finding that the ability to assemble into stable and highly organised structures (*e.g.* amyloid fibrils) is not an unusual feature exhibited by a small group of peptides and proteins with special sequence or structural properties, but rather a property shared by most, if not all, proteins;

(3) The discovery that specific aspects of protein behaviour, including their aggregation propensities[21,23,39,40] and the cellular toxicity associated with the aggregation process,[24,41] can be predicted with a remarkable degree of accuracy from the knowledge of their amino acid sequences;

(4) The realisation that a wide variety of techniques originally devised for applications in nanotechnology can be used to probe the nature of protein aggregation and assembly and of the structures that emerge;[30,42–44] and

(5) The development of powerful approaches using model organisms for probing the origins and progression of misfolding diseases by linking concepts and principles emerging from *in vitro* studies to *in vivo* phenomena such as neurodegeneration.[24]

An analysis of these results, which span across a wide range of subjects from neuroscience to nanoscience, reveals that the ability to keep proteins in their soluble form is absolutely central for the maintenance of cell homeostasis.

Protein solubility and biological complexity

Considerable advances have been made in recent years in the search for the chemical and physical principles that underlie the complexity of biological phenomena. We believe that it is possible to describe, for example, processes such as folding and aggregation, at least in outline, in generic terms and link them to well-established and quantifiable concepts of chemistry and physics. In this context, it should then be possible to study in depth a relatively small number of carefully chosen proteins, and yet extract general principles from such studies. In particular, we believe that there are many common features underlying the diseases associated with amyloid formation. Thus, much of our research is focused on the development of this theme using amyloid-related diseases as a paradigm of the way the ideas of chemistry and physics can provide fundamental insight into both normal and aberrant behaviour and suggest novel therapeutic strategies.

Biological systems have evolved to be efficient by achieving astonishing levels of molecular crowding within cells and extracellular space, which is of the order of

300 g l^{-1}.[10] Information can then be transmitted rapidly, largely by molecular diffusion, between different components enabling them to function efficiently.[3] Molecules such as proteins remain soluble and able to avoid interaction with all but a relatively small and specific selection of other molecules, yet are composed of chemical species that are often extremely hydrophobic and prone to self-assemble. We are beginning to think that the ability to maintain the solubility of its component molecules is of much more general significance in biology than previously imagined.[27] Thus the observation that the sequence determines the solubility of proteins[45] could be just as important as the fact that it determines their structure and the ability to fold. As we understand in increasing detail how sequences define solubility, it is becoming possible to predict aspects of biology in ways that were previously unsuspected.

Multiple forms of protein structure

Much is understood about the nature of globular protein structures and about the principles by which isolated denaturated polypeptide chains are able to achieve such states.[4,8–10] A more complete knowledge of the behaviour of proteins in the cell has, however, been limited by the challenges involved in defining the structures of native states in complex environments and of the highly dynamic structural ensembles that describe most of the additional forms of proteins that are now known to be of biological importance, including natively unfolded states and partially folded states involved in folding and in aggregation.

A detailed structural description of native and non-native states lies at the heart of our ability to describe in a quantitative manner the complex behaviour of proteins within a cell. Powerful techniques are being developed to complement more established methods to overcome the challenges posed by the task of providing such a description.[30,42–44,46] Our own approach is based primarily on methods that directly combine experimental and computational techniques.[5,6,35] These procedures involve the use of experimental data, largely derived from NMR spectroscopy, as restraints in computer simulations.[36,37] We have already used in this way a range of different types of experimental data, for example distance measurements from paramagnetic probes introduced by mutagenesis,[47,48] and structural and dynamical information from hydrogen–deuterium exchange experiments.[49–51] But a major breakthrough has come recently through the discovery[37,52–55] of ways in which chemical shift data can be used in this approach to generate structures of native states to an accuracy comparable to that of conventional methods (Fig. 1). The use of chemical shifts requires only a resolved and assigned spectrum and thus renders unnecessary the measurement of large numbers of additional parameters such as interatomic distances derived from nuclear Overhauser effects (NOEs). The latter measurements are very challenging (and in practice virtually impossible) in highly dynamical systems and for the conformationally heterogeneous states populated, for example, along the folding and misfolding pathways.

These chemical shift techniques, particularly in conjunction with the measurement of residual dipolar couplings (RDCs), can be used to generate structural ensembles of non-native states for which these parameters are often the only ones that can be readily measured.[56,57] These methods can be used to study a wider range of protein states, including highly dynamical ones such as low-populated conformations involved in the folding and misfolding processes. Indeed we have already provided a proof of principle that the use of chemical shift restraints, here derived from relaxation dispersion techniques, can be used to characterise transient species.[58] We believe that these computational approaches, as well as advances in experimental techniques that enable the systematic measurements of chemical shifts in non-native states of proteins,[56] will enable us to increase very considerably the resolution to which this type of non-native structure can be determined.

This approach should enable the characterisation of proteins whose structures have proved elusive by conventional means, but which are crucial to understanding

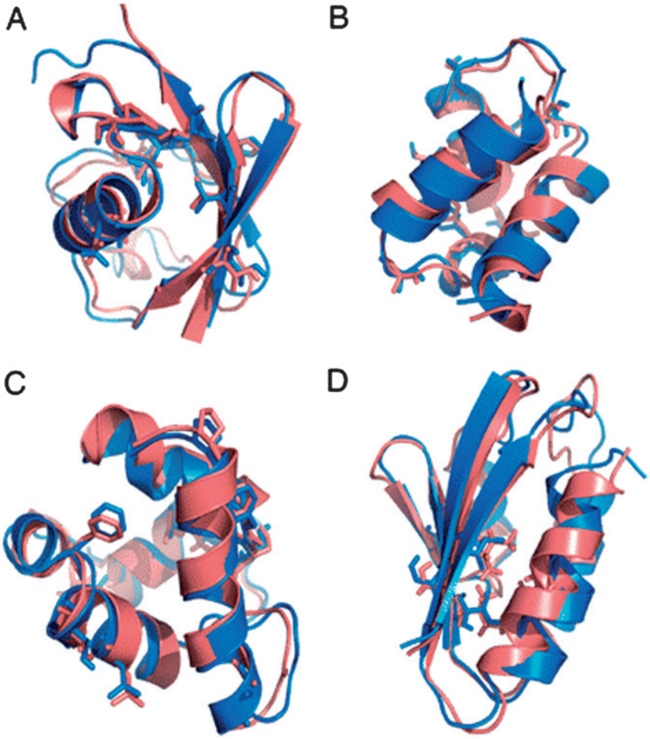

Fig. 1 Comparison between protein structures determined by X-ray crystallography (pink) and by the technique that we have recently introduced that uses only NMR chemical shift information (blue).[37] Despite being only at the initial stages of development of the method, we are already able to generate structures of globular proteins of up to 120 residues in length that agree with the those determined by conventional methods with RMSD values of 1.2–1.8 Å for the backbone atoms and 2.1–2.6 Å for the side-chain atoms.

the process of protein folding. Two recent studies[59,60] have already demonstrated this approach by using as test cases calmodulin,[59] a protein that plays an essential role in signal transduction pathways, and ubiquitin,[60] a protein that is a key component in degradation pathways, that the use of molecular dynamics simulations with NMR restraints[36] enables the changes in structure and dynamics upon binding to be described at nearly atomic level resolution, thus enabling analysis of the molecular mechanisms responsible for binding to be carried out. These studies have provided strong support for the "equilibrium shift" model,[61,62] according to which the conformations that proteins adopt in the bound state are already present, although with low statistical weights, in the unbound states in solution. This mechanism enables proteins such as calmodulin and ubiquitin to interact with large numbers of other proteins in a selective and efficient manner.

Another crucial area in which the inclusion of experimental measurements in molecular dynamics simulations is having a major impact is in the investigation of the structures of protein complexes, where even weak interactions can be detected by NMR methods.[63] In a first study,[53] we have already established, using the case of the structure of the cytotoxic endonuclease domain from bacterial toxin colicin (E9) in complex with its cognate immunity protein (Im9),[64] that chemical shifts enable the determination of protein complexes even when the complexes themselves exhibit significant dynamics and the component proteins undergo conformational rearrangements upon binding. It is also possible to determine the structure of protein–protein complexes using chemical shift information when the chemical shift changes upon binding are relatively small and hence particularly difficult to compute accurately. This result is a consequence of the well-known fact in NMR spectroscopy that the availability of a large number of restraints—in this case derived from chemical shifts—can provide enough information for high-resolution structure determination, even if they are not accurately known individually.[65] We have developed a computer code called CamDock to enable the structures of protein–protein complexes to be determined by combining advanced docking methods[53] with the information provided by chemical shifts[53] and residual dipolar couplings.[66]

It is then of very great importance to be able to relate the principles that emerge from studies in the test tube to analogous events occurring in the cell. Interesting work has been done using NMR spectroscopy in environments designed to mimic the cellular milieu,[67] but our ambition is to go further and ultimately to explore processes taking place in the cellular environment itself.

One of the most fundamental, and yet so far elusive, aspects of protein folding concerns the way in which this process is initiated during or following biosynthesis on the ribosome, *i.e.* the manner in which folding occurs in the cellular environment.[4,9,15,16] We have recently shown the feasibility of applying advanced NMR techniques to obtain detailed structural insights into the conformations of nascent proteins during the process of their synthesis on the ribosome.[38] In collaboration with Dr John Christodoulou (UCL) we generated ribosome–nascent chain complexes (RNCs) by arresting RNA translation, and used this technique in the first instance to study a tandem immunoglobulin (Ig) domain repeat (Ig2) of an actin-binding protein.[38] Analysis of the spectra of these RNCs selectively ^{15}N/^{13}C-labelled in the nascent chains reveals that the first Ig domain of the translation-arrested nascent chain is able to fold to a native-like state that remains tethered to the ribosome by the second highly disordered Ig domain. This study is now being extended by studying nascent chains of different lengths to probe the progressive development of structure in a growing nascent chain. We believe that this approach can open the door to descriptions at an atomistic level of detail of the process of co-translational folding, so as to characterise in detail the process by which proteins fold as the nascent chain emerges from the ribosomal exit tunnel.

In the immediate future, there are also great opportunities provided by the inclusion of NMR observables, including chemical shifts,[37,52–54] residual dipolar couplings[68] and interatomic distances obtained by paramagnetic relaxation enhancement experiments,[47,48] with molecular dynamics simulations to probe in detail the crucial processes by which cellular components such as molecular chaperones, including Hsp70 and trigger factor, interact with nascent chains and help to promote correct folding and inhibiting misfolding and aggregation.[4,15] This work has the potential of opening up a vast range of new opportunities to explore the study of the fundamental mechanism of folding in environments directly relevant to living systems.

Molecular basis of protein aggregation

The structures, dynamics and interactions that stabilise protein aggregates are difficult to study, since these species are often insoluble and resist crystallisation, thus making it very challenging to apply standard solution NMR spectroscopy and X-ray crystallography techniques.[20,69] Interdisciplinary approaches appear to be particularly suitable to address the problem of describing the structures of a variety of protein assemblies.

Very considerable progress has been recently made to characterise quantitatively the physical properties of fully formed amyloid fibrils and of their partially ordered protofibrillar precursors by bringing together solid-state NMR spectroscopy, cryo-electron microscopy and techniques of nanoscience. Advances in solid-state NMR methods are making it possible to obtain interatomic distance information in states that are insoluble and non-crystalline such as amyloid fibrils.[69] In addition to information about the structure of the amyloid fibrils formed by an 11-residue fragment of human transthyretin, these approaches have provided great insight into the structures of amyloid fibrils formed by several other peptides and proteins, including Het-S and Aβ.[70–72]

Together with the strategy mentioned above, nanoscience techniques are also emerging as powerful tools that can provide insight into the factors that stabilise amyloid fibrils.[30,42,44] In an initial study, by using atomic force microscopy (AFM) imaging we described the changes in the distribution of inter- *versus* intra-molecular bonding interactions associated with the transition of proteins from their native globular structures into ordered supramolecular assemblies. This work reveals that hydrogen bonded arrays of polypeptide chains exhibit material properties intermediate between those of hard materials such as steel and carbon nanotubes and softer biological fibres such as tubulin and actin (Fig. 2).[30] AFM and other techniques also generate information about the dimensions of the cross-β structural core of amyloid fibrils and of the less structured regions flanking them. These measurements provide unique insight into the kinetic and thermodynamic factors responsible for the stability of amyloid fibrils.

Much of our current understanding of the process of aggregation has been obtained by light scattering and fluorescence measurements of the kinetics of their growth.[20,29] As the spectroscopic signals in these techniques do not exclusively arise from amyloid fibrils but also from other types of aggregates that may be present in solution, and because of the non-linear relationship between aggregate abundance and signal intensity, the results can be difficult to interpret in a highly quantitative manner. In order to overcome such problems we have shown that very accurate measurements of growth rates can be obtained by a strategy in which real-time monitoring of the increase in mass of the fibrils themselves is carried out by measuring the variation in the frequency of a quartz crystal oscillator.[42] The application of this quartz crystal microbalance technique for monitoring the kinetics of aggregation of a series of proteins and under a variety of conditions is enabling a systematic analysis of the factors that can influence protein aggregation, particularly the mechanism by which molecular chaperones can inhibit fibril growth.[42]

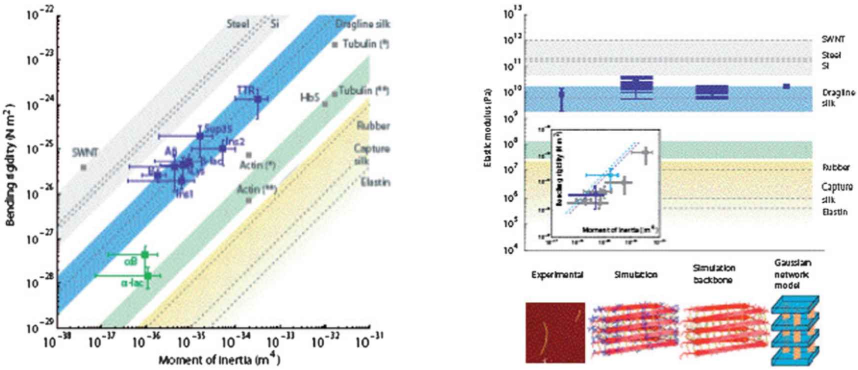

Fig. 2 A clear relationship between bending rigidity and moment of inertia has revealed the existence of universal mechanical properties of amyloid fibrils.[30] Left: illustration for a set of proteins of the values determined by atomic force microscopy (AFM) for the bending rigidities as a function of their cross-sectional moments of inertia (blue rectangles). For comparison, values for single-wall carbon nanotubes (SWNT), steel, actin, tubulin, rubber and elastin are also shown. The green shaded region shows the range of elastic moduli for materials held together by amphiphilic interactions. Right: comparison of the elastic moduli of fibrils from (blue squares, left to right) AFM measurements, full atom simulations, contribution of backbone alone, and results from the Gaussian network theory, in which the stability of a structure is estimated from the contributions of its hydrogen bonds.

Origins and consequences of aggregation in living systems

As we have discusses above, a wide range of human diseases is associated with protein misfolding and aggregation.[10,17–20] A powerful approach that has recently been proposed to study these phenomena is based on the combination of *in silico*, *in vitro* and *in vivo* studies to investigate the factors responsible for the abnormal assembly of proteins into insoluble aggregates and the effects that these conformational species have once released in the cellular environment.

Considerable progress has been made in characterising the major physicochemical factors that promote the aggregation of polypeptide chains, and increasingly sophisticated computational approaches have been developed[21,23,39,40] that enable predictions to be made about a variety of features of the process of aggregation of peptides and proteins. On this basis a method of predicting "aggregation propensity profiles" has been established that enables the identification of regions with a high intrinsic propensity for aggregation,[21,23,39,40] providing a platform for further development of this approach.

In the case of fully folded proteins, we have shown that it is possible to take into account the fact that the most amyloidogenic regions are protected from aggregation since they are located in the structural core of the native state and hence they are unable to form inter-molecular interactions without at least a degree of unfolding.[23] In essence, given the amino acid sequence of a protein, we have shown how it is possible to combine the predictions of the intrinsic aggregation propensity profiles with those for folding into stable structures to determine new aggregation propensity profiles of structured or partially structured proteins that account for the influence of the structural context. We have provided a initial demonstration of the potential of this approach through its application to the prediction of aggregation profiles for a range of peptides and proteins whose aggregation propensities have been characterized experimentally in particular detail, including the human prion protein, which is involved in sporadic, inherited and infectious forms of Creutzfeld–Jakob disease.

We anticipate that, in addition to its relevance for understanding misfolding diseases, the insight provided by these studies will in time represent a significant contribution to improving the biotechnological production of therapeutic peptides and proteins, in drug discovery initiatives and for antibody production. The ability to design rationally, and with increasing reliability, specific amino acid substitutions capable of altering significantly the aggregation propensities of peptides and proteins will enable us to investigate the physico-chemical factors responsible for the formation of amyloid fibrils and their oligomeric precursors.[20,21]

A range of strategies is being developed to combine *in vivo* approaches with *in vitro* and *in silico* methods to obtain a quantitative understanding of the molecular basis of neurodegenerative and other misfolding diseases. In an initial study we have demonstrated the potential of this approach in the case of the Aβ peptide, by showing that the relative toxicity in *Drosophila* of its mutational variants can be predicted with a remarkable 83% accuracy from their amino acid sequences (Fig. 3).[24] The advantage of using *Drosophila* models for such studies is that the brevity of their lifecycle, the power of the associated genetic tools, and the ease with which a range of toxicity-related phenotypes may be measured allows us to quantify the links between the *in vivo* toxicity of protein aggregates and the fundamental chemical properties of peptides and proteins.[24,41,73,74]

The results discussed above were obtained by developing a method for predicting the rate of formation of protofibrillar aggregates based on the physico-chemical properties of the amino acids comprising the sequences of the mutational variants of the Aβ42 peptide that we have investigated. It is also remarkable that, despite the fact that the intrinsic aggregation propensities of typical protein sequences vary by at least five orders of magnitude, we have been able to achieve profound alterations in the pathogenic effects of Aβ42 by increasing or decreasing its propensity to aggregate by less than 15%. This result suggests that proteins implicated in

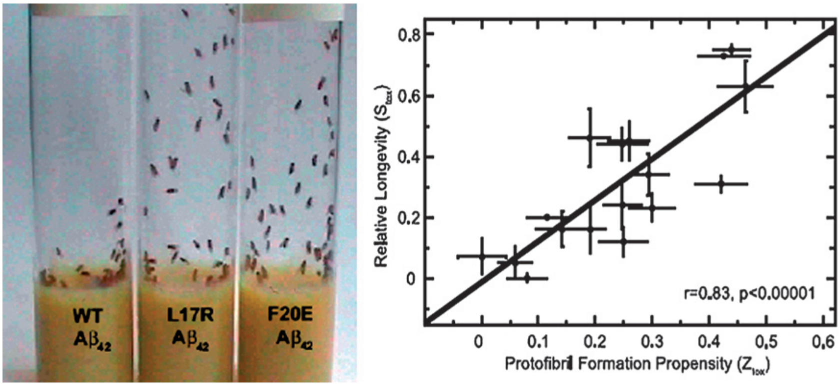

Fig. 3 Rational design of the toxic effects of Aβ42 mutants in a transgenic *Drosophila* model of Alzheimer's disease.[24] The relative longevity ($S_{\rm tox}$, y-axis) of flies expressing a range of Aβ42 variants is predicted accurately by a score ($Z_{\rm tox}$, x-axis) for the propensity to form protofibrillar aggregates ($r = 0.83$, $p < 0.00001$).

misfolding diseases are likely to be extremely close to the limit of their solubility under normal physiological conditions and consequently the small alterations in their concentration, environment or sequence, such as those that occur with genetic mutations or with increasing age, are likely to be the fundamental origin of these highly debilitating and increasingly common conditions.

The approach that we are developing is already enabling us to obtain accurate quantitative measurements of the relationships between the manifestations of neuronal dysfunction in a complex organism, such as locomotor defects and reduced lifespans, and the fundamental physico-chemical factors that determine the propensities of peptides and proteins to aggregate into oligomeric species and protofibrils. Our research is aimed at demonstrating that, despite the presence within the cell of multiple regulatory mechanisms such as molecular chaperones and degradation systems, it is the intrinsic, sequence-dependent propensity of the polypeptide chains to aggregate to form protofibrillar aggregates that is the primary determinant of its pathological behaviour in living systems.

Thus, by using quantitative *in vivo* and *in vitro* techniques we are exploring the links between various conformational states and pathological effects. Our strategy is to use the results of *in vitro* biophysical methods, including NMR spectroscopy, fibril formation assays and amyloid staining, to deduce the events occurring *in vivo*. We have pioneered this strategy in the case of the Aβ peptide to differentiate the effects of the propensities to form either fibrillar or protofibrillar aggregates, by rationally designing mutations that alter either the fibrillar or the protofibrillar propensities.[24]

Protein homeostasis in normal and aberrant biology

The fundamental connection between two aspects of proteins in the cell—their abundance and their solubility—is increasingly evident. A direct characterisation of this relationship is the very high correlation (97%) observed between the *in vitro* aggregation rates of a series of human proteins and the corresponding *in vivo* mRNA expression levels.[27] Thus, even relatively small alterations in protein abundance and *in vivo* solubility can be linked to human disease, as described below.

The existence of a close relationship between mRNA expression levels and protein aggregation rates (Fig. 4) provides a new perspective on the phenomenon of protein aggregation and of its connections with misfolding diseases.[27] In essence, these results suggest that the amino acid sequences of proteins determine not only their

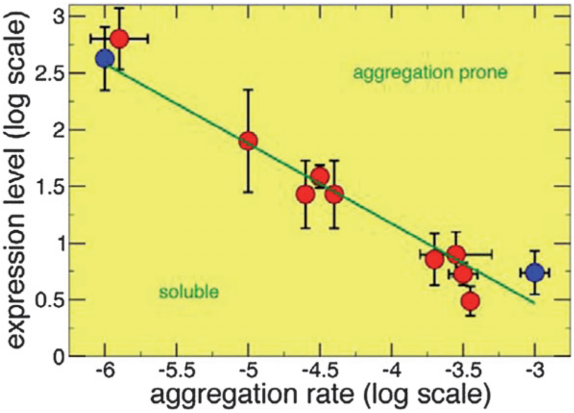

Fig. 4 Relationship between mRNA expression levels measured *in vivo* through microarray technologies and the aggregation rates of the corresponding proteins measured *in vitro*. We have considered all the 11 human proteins, either involved in disease (red circles) or not (blue circles), for which protein aggregation rates have been measured under near-physiological conditions.[27]

folding behaviour, but also their aggregation propensities and ultimately the susceptibility of an organism to contracting aggregation-related diseases.

The strong degree of anticorrelation between expression levels and aggregation rates suggests that the aggregation propensities of the proteins needed by the cell are precisely tuned to levels that enable them to be functional at the concentrations required for optimally efficient performance. It also indicates that protein molecules have co-evolved with their cellular environments to be sufficiently soluble for their biological roles, but no more so, and hence that aggregation can result from even minor changes in the chemistry and in the regulation of otherwise harmless proteins. Indeed, expression at higher levels than those found naturally is likely to result in enhanced clearance or in deposition of the proteins involved, both of which can generate disease. The intimate relationship between expression levels and aggregation rates is the net result of the opposing effects of an evolutionary pressure to remain soluble at the concentrations needed by the cell and random mutational processes that tend to increase their aggregation propensity.[27] Thus, over time, evolutionary selection generates proteins able to resist aggregation just at the levels required. Such a relationship provides dramatic evidence for the generic ideas about aggregation[10,28]—specifically that the propensity of proteins to revert to the amyloid state is the ultimate origin of amyloid-related diseases.

We are just beginning to explore in detail the hypothesis that proteins have evolved to have low enough aggregation propensities to enable an organism to function optimally, but with almost no scope for dealing with any situation where these levels increase further. We are now investigating the way in which variations in the concentration of proteins influence their behaviour in the cell. These studies are enabling us to understand the interplay of the physico-chemical properties of proteins and of the quality control mechanisms present in the cell to regulate the way in which they act. The approach that we are developing will enable us to obtain accurate measurements of the relationships between manifestations of neuronal dysfunction in a complex organism, such as locomotor defects and reduced lifespans, and the fundamental chemical factors that determine the propensities of peptides and proteins to aggregate into protofibrils.[24,75] We aim to understand how, despite the presence within the cell of multiple regulatory mechanisms such as molecular chaperones and degradation systems to avoid protein deposition, the intrinsic sequence-dependent propensity of polypeptide chains to aggregate remains a major determinant of the pathologies associated with misfolding and aggregation in living systems. In addition to demonstrating that rational mutagenesis can be used to alter systematically the toxicity of peptides and proteins, the transgenic *Drosophila* models that we are developing are enabling us to perform a quantitative analysis of the effects on the lifespan and mobility of *Drosophila* of other factors likely to be relevant to the *in vivo* aggregation process, in particular molecular chaperones, small therapeutic molecules and antibodies, or antibody-like species, by co-expression techniques.

Towards a quantitative biology based on physico-chemical principles

Complex regulatory networks, involving primarily nucleic acids and proteins, orchestrate the cellular functions required to maintain protein homeostasis.[3,33] These same cellular functions are also, however, dependent on the basic chemistry of the molecules taking part in them. Therefore, the “chemical” and the “cellular” views of cell biology are closely related, as is revealed, for example, by the high correlation between expression levels and aggregation propensities. By this statement, we do not mean that gene regulation itself is not important, as there is a huge amount of evidence that demonstrates its key role in protein homeostasis.[76] What we are saying is that very significant advances can be made by considering the “chemical view” of protein homeostasis. Studies are under way to explore the validity of the hypothesis that we have formulated according to which the necessity of avoiding aggregation plays a major evolutionary role in ensuring that proteins can remain soluble in the cell at the concentrations required for their function.

As an initial example, we discuss here the case of gene expression, which is the process through which the information contained in the DNA sequence of an organism is converted into functional proteins.[76] In response to the requirements of a cell, each step in the process of gene expression is regulated by complex cellular mechanisms, from the transcription of DNA into mRNA to the post-translational modification of proteins. The conversion of the information stored in DNA into proteins takes place through several phases that are highly regulated in response to the functional requirements of proteins by the cell. A detailed knowledge of the mechanisms of regulation can be used to rationalise and ultimately predict the

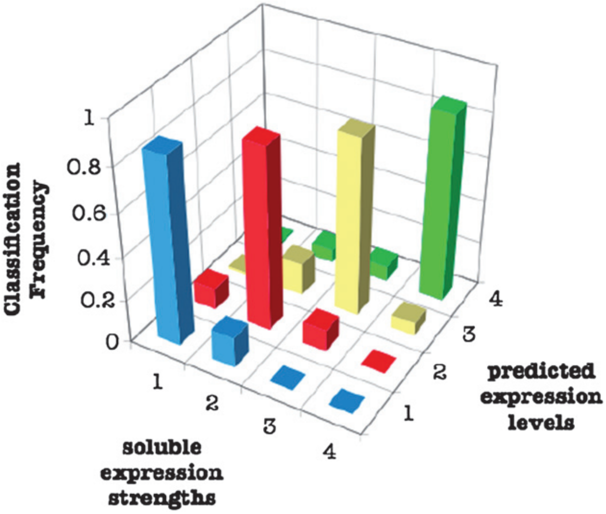

Fig. 5 Prediction of mRNA expression levels of recombinant human proteins in *E. coli*. When soluble expression strengths are compared with our predictions the overall accuracy is above 90%.[45]

outcome of the gene expression process. For example, the study of *cis*-regulatory motifs encoded in DNA sequences has reported an accuracy of about 70% for the prediction of expression patterns, while the correlation between the frequency of translational codons and gene expression levels is estimated to be around 60%.[77]

As an alternative to the strategy of exploiting the knowledge of the cellular regulatory processes to predict gene expression, we have proposed an approach, which has been prompted by the observation that proteins, once expressed, must remain soluble and avoid misfolding and aggregation in order to function optimally and to avoid cellular damage.[27] Since, as discussed above, we have established quantitative methods for predicting aggregation rates of proteins from the knowledge of the chemical properties of their sequences, we are in a position to investigate the relationship between these properties and the levels of expression of the genes. Our results indicate that it is possible to predict mRNA expression levels in *E. coli* with an accuracy of 90% or better from the knowledge of the sequences of the corresponding proteins (Fig. 5).[45]

Conclusions

Quantitative methods in molecular biology are providing unprecedented insights into the molecular mechanisms by which protein homeostasis is maintained in the cell. By drawing primarily on results from our own research we have discussed a variety of strategies based on the physico-chemical properties of proteins that appear to be particularly promising for increasing our ability to describe their behaviour *in vivo* and to suggest rational approaches to modulate it.

Acknowledgements

The results that we have described from our own research have been generated through a series of interdisciplinary collaborations. We would like in particular to acknowledge the contributions from Dr Andrea Cavalli, Dr Damian Crowther, Dr Anthony Fitzpatrick, Dr Bob Griffin, Dr Thomas Jahn, Dr Tuomas Knowles, Dr David Lomas, Dr Leila Luheshi, Dr Rinaldo Montalvao, Dr Gian Gaetano Tartaglia, Dr Mark Welland and Dr Helen Saibil. Our research is supported by the Royal Society, the Wellcome Trust, the Leverhulme Trust, the Alzheimer's Research Trust, the European Community, the European Molecular Biology Organisation, the Biotechnology and Biological Sciences Research Council and the Medical Research Council.

References

1 P. Aloy, B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A. C. Gavin, P. Bork, G. Superti-Furga, L. Serrano and R. B. Russell, Structure-based assembly of protein complexes in yeast, *Science*, 2004, **303**, 2026–2029.
2 C. V. Robinson, A. Sali and W. Baumeister, The molecular sociology of the cell, *Nature*, 2007, **450**, 973–982.
3 L. H. Hartwell, J. J. Hopfield, S. Leibler and A. W. Murray, From molecular to modular cell biology, *Nature*, 1999, **402**, C47–C52.
4 F. U. Hartl and M. Hayer-Hartl, Protein folding—Molecular chaperones in the cytosol: from nascent chain to folded protein, *Science*, 2002, **295**, 1852–1858.
5 M. Vendruscolo, J. Zurdo, C. E. MacPhee and C. M. Dobson, Protein folding and misfolding: a paradigm of self-assembly and regulation in complex biological systems, *Philos. Trans. R. Soc. London, Ser. A*, 2003, **361**, 1205–1222.
6 M. Vendruscolo and C. M. Dobson, Towards complete descriptions of the free-energy landscapes of proteins, *Philos. Trans. R. Soc. London, Ser. A*, 2005, **363**, 433–450.
7 W. E. Balch, R. I. Morimoto, A. Dillin and J. W. Kelly, Adapting proteostasis for disease intervention, *Science*, 2008, **319**, 916–919.
8 A. R. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, W.H. Freeman, New York, 1999.

9 M. J. Gething and J. Sambrook, Protein folding in the cell, *Nature*, 1992, **355**, 33–45.

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

11 C. M. Dobson, Chemical space and biology, *Nature*, 2004, **432**, 824–828.

12 K. A. Dill and H. S. Chan, From Levinthal to pathways to funnels, *Nat. Struct. Biol.*, 1997, **4**, 10–19.

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

14 J. N. Onuchic and P. G. Wolynes, Theory of protein folding, *Curr. Opin. Struct. Biol.*, 2004, **14**, 70–75.

15 J. Frydman, Folding of newly translated proteins *in vivo*: The role of molecular chaperones, *Annu. Rev. Biochem.*, 2001, **70**, 603–647.

16 B. Bukau, J. Weissman and A. Horwich, Molecular chaperones and protein quality control, *Cell*, 2006, **125**, 443–451.

17 R. W. Carrell and D. A. Lomas, Conformational disease, *Lancet*, 1997, **350**, 134–138.

18 B. Caughey and P. T. Lansbury, Protofibrils, pores, fibrils, and neurodegeneration: Separating the responsible protein aggregates from the innocent bystanders, *Annu. Rev. Neurosci.*, 2003, **26**, 267–298.

19 C. Haass and D. J. Selkoe, Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid beta-peptide, *Nat. Rev. Mol. Cell Biol.*, 2007, **8**, 101–112.

20 F. Chiti and C. M. Dobson, Protein misfolding, functional amyloid, and human disease, *Annu. Rev. Biochem.*, 2006, **75**, 333–366.

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

22 A. M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz and L. Serrano, Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins, *Nat. Biotechnol.*, 2004, **22**, 1302–1306.

23 G. G. Tartaglia, A. Pawar, S. Campioni, F. Chiti and M. Vendruscolo, Prediction of aggregation-prone regions of structured proteins, *J. Mol. Biol.*, 2008, **380**, 425–436.

24 L. M. Luheshi, G. G. Tartaglia, A.-C. Brorsson, A. P. Pawar, I. E. Watson, F. Chiti, M. Vendruscolo, D. A. Lomas, C. M. Dobson and D. C. Crowther, Systematic *in vivo* analysis of the intrinsic determinants of amyloid beta pathogenicity, *PLoS Biol.*, 2007, **5**, e290.

25 R. L. Wiseman, E. T. Powers, J. N. Buxbaum, J. W. Kelly and W. E. Balch, An adaptable standard for protein export from the endoplasmic reticulum, *Cell*, 2007, **131**, 809–821.

26 T. W. Mu, D. S. T. Ong, Y. J. Wang, W. E. Balch, J. R. Yates, L. Segatori and J. W. Kelly, Chemical and biological approaches synergize to ameliorate protein-folding diseases, *Cell*, 2008, **134**, 769–781.

27 G. G. Tartaglia, S. Pechmann, C. M. Dobson and M. Vendruscolo, Life on the edge: A link between gene expression levels and aggregation rates of human proteins, *Trends Biochem. Sci.*, 2007, **32**, 204–206.

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

29 T. R. Jahn and S. E. Radford, Folding *versus* aggregation: Polypeptide conformations on competing pathways, *Arch. Biochem. Biophys.*, 2008, **469**, 100–117.

30 T. P. J. Knowles, A. W. Fitzpatrick, H. R. Mott, S. Meehan, M. Vendruscolo, C. M. Dobson and M. E. Welland, Mechanical properties reveal the dominance of backbone interactions in stabilising amyloid fibrils, *Science*, 2007, **318**, 1900–1903.

31 A. L. Barabási and R. Albert, Emergence of scaling in random networks, *Science*, 1999, **286**, 509–512.

32 H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A. L. Barabasi, The large-scale organization of metabolic networks, *Nature*, 2000, **407**, 651–654.

33 A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell and G. Superti-Furga, Proteome survey reveals modularity of the yeast cell machinery, *Nature*, 2006, **440**, 631–636.

34 I. Volkov, J. R. Banavar, S. P. Hubbell and A. Maritan, Patterns of relative species abundance in rainforests and coral reefs, *Nature*, 2007, **450**, 45–49.

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

36 K. Lindorff-Larsen, R. B. Best, M. A. DePristo, C. M. Dobson and M. Vendruscolo, Simultaneous determination of protein structure and dynamics, *Nature*, 2005, **433**, 128–132.

37 A. Cavalli, X. Salvatella, C. M. Dobson and M. Vendruscolo, Protein structure determination from NMR chemical shifts, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 9615–9620.

38 S. T. D. Hsu, P. Fucini, L. D. Cabrita, H. Launay, C. M. Dobson and J. Christodoulou, Structure and dynamics of a ribosome-bound nascent chain by NMR spectroscopy, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 16516–16521.

39 K. F. Dubay, A. P. Pawar, F. Chiti, J. Zurdo, C. M. Dobson and M. Vendruscolo, Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains, *J. Mol. Biol.*, 2004, **341**, 1317–1326.

40 A. P. Pawar, K. F. DuBay, J. Zurdo, F. Chiti, M. Vendruscolo and C. M. Dobson, Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases, *J. Mol. Biol.*, 2005, **350**, 379–392.

41 D. C. Crowther, R. Page, D. Chandraratna and D. A. Lomas, A *Drosophila* model of Alzheimer's disease, *Methods Enzymol.*, 2006, **412**, 234–255.

42 T. P. J. Knowles, W. M. Shu, G. L. Devin, S. Meehan, S. Auer, C. M. Dobson and M. E. Welland, Kinetics and thermodynamics of amyloid formation from direct measurements of fluctuations of fibril mass, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 10016–10021.

43 T. P. J. Knowles, J. F. Smith, A. Craig, C. M. Dobson and M. E. Welland, Spatial persistence of angular correlations in amyloid fibrils, *Phys. Rev. Lett.*, 2006, **96**, 238301.

44 J. F. Smith, T. P. J. Knowles, C. M. Dobson, C. E. MacPhee and M. E. Welland, Characterization of the nanoscale properties of individual amyloid fibrils, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 15806–15811.

45 G. G. Tartaglia, S. Pechmann, C. M. Dobson and M. Vendruscolo, A relationship between mRNA expression levels and protein solubility in *E. coli*, *J. Mol. Biol.*, 2009, **388**, 381–389.

46 H. J. Dyson and P. E. Wright, Intrinsically unstructured proteins and their functions, *Nat. Rev. Mol. Cell Biol.*, 2005, **6**, 197–208.

47 M. M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo and C. M. Dobson, Mapping long-range interactions in alpha-synuclein using spin- label NMR and ensemble molecular dynamics simulations, *J. Am. Chem. Soc.*, 2005, **127**, 476–477.

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

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

50 R. B. Best and M. Vendruscolo, Structural interpretation of hydrogen exchange protection factors in proteins: Characterization of the native state fluctuations of C12, *Structure*, 2006, **14**, 97–106.

51 J. Gsponer, H. Hopearuoho, S. B. M. Whittaker, G. R. Spence, G. R. Moore, E. Paci, S. E. Radford and M. Vendruscolo, Determination of an ensemble of structures representing the intermediate state of the bacterial immunity protein Im7, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 99–104.

52 Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. H. Liu, A. Eletsky, Y. B. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker and A. Bax, Consistent blind protein structure generation from NMR chemical shift data, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 4685–4690.

53 R. W. Montalvao, A. Cavalli, X. Salvatella, T. L. Blundell and M. Vendruscolo, Structure determination of protein–protein complexes using NMR chemical shifts: Case of an endonuclease colicin–immunity protein complex, *J. Am. Chem. Soc.*, 2008, **130**, 15990–15996.

54 P. Robustelli, A. Cavalli and M. Vendruscolo, Determination of protein structures from solid-state NMR chemical shifts, *Structure*, 2008, **16**, 1764–1769.

55 D. S. Wishart, D. Arndt, M. Berjanskii, P. Tang, J. Zhou and G. Lin, CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data, *Nucleic Acids Res.*, 2008, **36**, W496–W502.

56 P. Vallurupalli, D. F. Hansen and L. E. Kay, Structures of invisible, excited protein states by relaxation dispersion NMR spectroscopy, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 11766–11771.

57 A. De Simone, B. Richter, X. Salvatella and M. Vendruscolo, Toward an accurate determination of free energy landscapes in solution states of proteins, *J. Am. Chem. Soc.*, 2009, **131**, 3810–3811.

58 D. M. Korzhnev, X. Salvatella, M. Vendruscolo, A. A. Di Nardo, A. R. Davidson, C. M. Dobson and L. E. Kay, Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR, *Nature*, 2004, **430**, 586–590.

59 J. Gsponer, J. Christodoulou, A. Cavalli, J. M. Bui, B. Richter, C. M. Dobson and M. Vendruscolo, A coupled equilibrium shift mechanism in calmodulin-mediated signal transduction, *Structure*, 2008, **16**, 736–746.

60 O. F. Lange, N. A. Lakomek, C. Fares, G. F. Schroder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmuller, C. Griesinger and B. L. de Groot, Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution, *Science*, 2008, **320**, 1471–1475.

61 D. D. Boehr, D. McElheny, H. J. Dyson and P. E. Wright, The dynamic energy landscape of dihydrofolate reductase catalysis, *Science*, 2006, **313**, 1638–1642.

62 M. Vendruscolo and C. M. Dobson, Dynamic visions of enzymatic reactions, *Science*, 2006, **313**, 1586–1587.

63 E. R. P. Zuiderweg, Mapping protein–protein interactions in solution by NMR spectroscopy, *Biochemistry*, 2002, **41**, 1–7.

64 C. Kleanthous, U. C. Kuhlmann, A. J. Pommer, N. Ferguson, S. E. Radford, G. R. Moore, R. James and A. M. Hemmings, Structural and mechanistic basis of immunity toward endonuclease colicins, *Nat. Struct. Biol.*, 1999, **6**, 243–252.

65 K. Wuthrich, Protein structure determination in solution by nuclear magnetic resonance spectroscopy, *Science*, 1989, **243**, 45–50.

66 A. Grishaev, J. Wu, J. Trewhella and A. Bax, Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data, *J. Am. Chem. Soc.*, 2005, **127**, 16621–16628.

67 P. Selenko, Z. Serber, B. Gade, J. Ruderman and G. Wagner, Quantitative NMR analysis of the protein G B1 domain in *Xenopus laevis* egg extracts and intact oocytes, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 11904–11909.

68 G. M. Clore and C. D. Schwieters, How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation?, *J. Am. Chem. Soc.*, 2004, **126**, 2923–2938.

69 M. Baldus, ICMRBS founder's medal 2006: Biological solid-state NMR, methods and applications, *J. Biomol. NMR*, 2007, **39**, 73–86.

70 C. Wasmer, A. Lange, H. Van Melckebeke, A. B. Siemer, R. Riek and B. H. Meier, Amyloid fibrils of the HET-s(218–289) prion form a beta solenoid with a triangular hydrophobic core, *Science*, 2008, **319**, 1523–1526.

71 A. T. Petkova, W. M. Yau and R. Tycko, Experimental constraints on quaternary structure in Alzheimer's beta-amyloid fibrils, *Biochemistry*, 2006, **45**, 498–512.

72 N. Ferguson, J. Becker, H. Tidow, S. Tremmel, T. D. Sharpe, G. Krause, J. Flinders, M. Petrovich, J. Berriman, H. Oschkinat and A. R. Fersht, General structural motifs of amyloid protofilaments, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 16248–16253.

73 M. B. Feany and W. W. Bender, A *Drosophila* model of Parkinson's disease, *Nature*, 2000, **404**, 394–398.

74 J. Bilen and N. M. Bonini, *Drosophila* as a model for human neurodegenerative disease, *Annu. Rev. Genet.*, 2005, **39**, 153–171.

75 L. M. Luheshi, D. C. Crowther and C. M. Dobson, Protein misfolding and disease: from the test tube to the organism, *Curr. Opin. Chem. Biol.*, 2008, **12**, 25–31.

76 M. Levine and R. Tjian, Transcription regulation and animal diversity, *Nature*, 2003, **424**, 147–151.

77 H. J. Bussemaker, B. C. Foat and L. D. Ward, Predictive modeling of genome-wide mRNA expression: From modules to molecules, *Annu. Rev. Biophys. Biomol. Struct.*, 2007, **36**, 329–347.