# Towards quantitative predictions in cell biology using chemical properties of proteins

Michele Vendruscolo and Gian Gaetano Tartaglia*

DOI: 10.1039/b805710a

It has recently been suggested that the concentrations of proteins in the cell are tuned towards their critical values, and that the alteration of this balance often results in misfolding diseases. This concept is intriguing because the *in vivo* concentrations of proteins are closely regulated by complex cellular processes, while their critical concentrations are primarily determined by the chemical characters of their amino acid sequences. We discuss here how the presence of a link between the upper levels of *in vivo* concentrations and critical concentrations offers an opportunity to make quantitative predictions in cell biology based on the chemical properties of proteins.

## Introduction: protein chemistry and cell biology

Cellular functions are orchestrated by complex regulatory networks involving primarily nucleic acids and proteins.[1,2] These same cellular functions are also, however, dependent on the basic chemistry of the molecules that carry them out.[3–6] Therefore, the ''chemical'' and the ''cellular'' views of cell biology are complementary and closely related. A specific link between them emerges from the suggestion that proteins remain soluble in the cellular environment only up to the concentrations required for their functions (Fig. 1). This notion follows from the strong correlation that has been recently reported between expression levels of human genes *in vivo* and the aggregation rates of the corresponding proteins measured *in vitro*[7] (Fig. 2).

In this *highlight* we explore the opportunities provided by the relationships between the upper levels of *in vivo* concentrations and the critical concentrations of proteins (Fig. 1) and between mRNA expression levels and aggregation rates of the corresponding proteins (Fig. 2) to develop strategies to make quantitative predictions in cell biology through the analysis of the chemical properties of proteins. We first describe the advances that have been recently made for predicting the solubility of proteins[8–12] and their propensity for aggregation.[13–20] We then explain how the relationship between protein aggregation rates and mRNA expression levels (Fig. 2) provides a way to use the chemical properties of proteins to estimate the expression levels themselves (Fig. 3), and to assess the degree of neuronal dysfunction and degeneration in a *Drosophila* model of Alzheimer's disease (Fig. 4).[21]

*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK. E-mail: ggt23@cam.ac.uk*

## Solubility and chemical properties of proteins

The solubility of proteins can be defined in terms of their critical concentration, which is the concentration above which they form a condensed phase.[9] The solubility depends strongly on the thermodynamic conditions, including temperature, pH and ionic strength and on the composition of the solvent. Therefore the solubility under standard *in vitro* conditions might differ from the solubility in the highly complex cellular environment.[9] Moreover, as the most insoluble regions of protein sequences are secluded from the solvent through the folding process, the solubility of proteins depends also on the stability of their native states.[9] Despite these complications, Wilkinson and Harrison demonstrated that the solubility of proteins in *Escherichia coli* can be predicted from their chemical properties, including charge, propensity for forming

**Michele Vendruscolo**

*Michele Vendruscolo obtained a PhD in condensed matter physics at SISSA in Trieste. He is now a reader in theoretical chemical biology at the University of Cambridge and an EMBO Young Investigator. His research interests are primarily focused on the investigation of the structures and properties of proteins, and their relationship to biological evolution and misfolding diseases.*

**Gian Gaetano Tartaglia**

*Gian Gaetano Tartaglia obtained an MPhil in theoretical physics at La Sapienza in Rome. He then received his PhD in biochemistry at the University of Zurich, and he is currently working as a postdoctoral researcher in Cambridge in the field of protein misfolding and aggregation.*

 

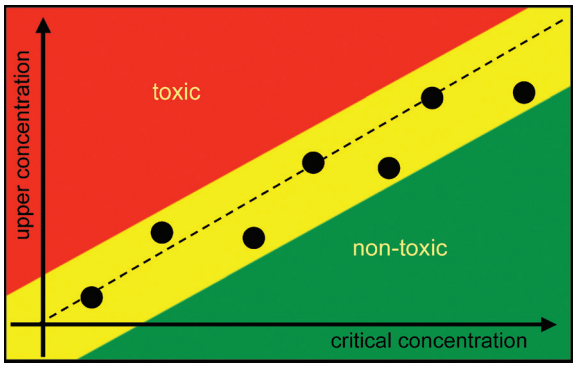


**Fig. 1** Schematic illustration of the expected relationship between the *in vivo* concentrations and the critical concentrations of proteins under physiological conditions. The amino acid sequences of proteins have evolved to remain soluble at the upper levels of abundance required by the cell during their cycles. As random mutations tend to destabilize their native folds[25–27] and increase their propensities to aggregate,[28] proteins slowly drift towards the toxic region of low solubility (shown in red). Evolutionary selection prevents them from entering this region and stops the drift at the boundary between the toxic and the non-toxic regions (the latter shown in green). This boundary is rather sharp (yellow band) and therefore causes the high correlation between aggregation rates and expression levels reported in Fig. 2.

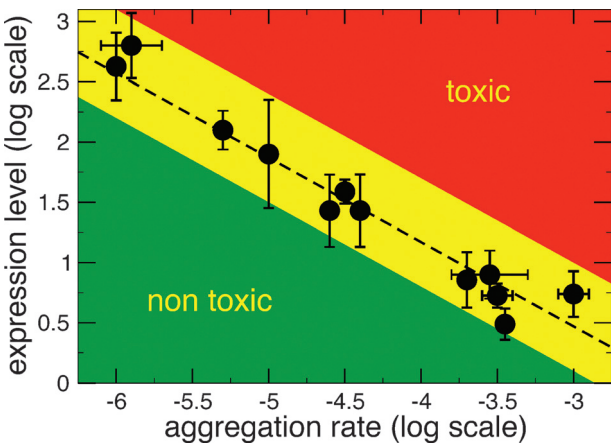


**Fig. 2** Relationship between mRNA expression levels and aggregation rates for 11 human proteins for which the latter have been measured *in vitro* under near-physiological conditions.[7] The coefficient of correlation between these two quantities is 0.97.

turns, hydrophilicity and length of the sequence.[8] The degree of solubility of a protein, *cv*, was estimated as:

$$cv = \alpha\left[\frac{N+G+P+S}{n}\right] + \beta\left[\frac{R+K-D-E}{n}+\gamma\right] \quad (1)$$

In this equation, $n$ is the number of amino acids in the protein, $\alpha$ and $\beta$ are two parameters describing, respectively, the electrostatic charge and the propensity to form turns, and $\gamma$ is a constant; furthermore $N$, $G$, $P$, and $S$ are the number of Asn, Gly, Pro, and Ser residues and $R$, $K$, $D$, and $E$ the number of Arg, Lys, Asp, and Glu residues, respectively. Using a dataset of 81 proteins, the simple mathematical model given by eqn (1) was proposed for the identification of soluble eukaryotic targets in *E. coli*.[22] After this initial work, several other studies have increased our understanding of the relationship between the chemical properties of amino acid sequences and their solubility.[10–12]

## Aggregation and chemical properties of proteins

It is well known that proteins have a strong tendency to aggregate into insoluble assemblies,[23] and that the ability to form aggregates is not restricted to the proteins whose deposition is associated with specific diseases, but is a generic property of polypeptide chains.[24]

In 2003, Chiti, Dobson and co-workers reported the existence of a significant correlation between the changes in the aggregation rates resulting from single amino acid substitutions and their effects on the chemical properties of polypeptide chains:[13]

$$\ln(\nu_{wt}/\nu_{mut}) = \alpha\Delta H + \beta(\Delta\Delta G_{coil\text{-}\alpha} + \Delta\Delta G_{\beta\text{-}coil}) + \gamma\Delta C \quad (2)$$

In this equation, $\ln(\nu_{mut}/\nu_{wt})$ is the logarithm in base 10 of the ratio between $\nu_{wt}$, the aggregation rate of the wild type protein, and $\nu_{mut}$, the aggregation rate of its mutational variant. $\Delta H$, $\Delta\Delta G_{coil\text{-}\alpha} + \Delta\Delta G_{\beta\text{-}coil}$, and $\Delta C$ represent the change in hydrophobicity, secondary structure propensity and charge upon mutation, respectively. Eqn (2) reproduces to a remarkable extent the changes in the aggregation rates observed experimentally *in vitro* for single amino acid substitutions for series of peptide and proteins, including those associated with disease.[13]

This initial observation was then developed into an approach in which environmental conditions are also taken into account in order to predict the overall aggregation propensities of peptides and proteins from the knowledge of their amino acid sequences and of the conditions used experimentally to monitor the aggregation process, including the pH, the ionic strength and the protein concentration of the solution in which aggregation occurs.[16] The parameters corresponding to all these terms were estimated by fitting the elongation rates of a dataset of proteins compiled through an extensive literature search.[16] The same approach was also shown to lead to accurate predictions of the regions of the sequence of a peptide or a protein that are most important in determining the aggregation process.[16,17,19,20]

## A link between protein chemistry and cell biology

By considering together the results described in the previous two sections, we observe that the same chemical parameters used for predicting the solubility

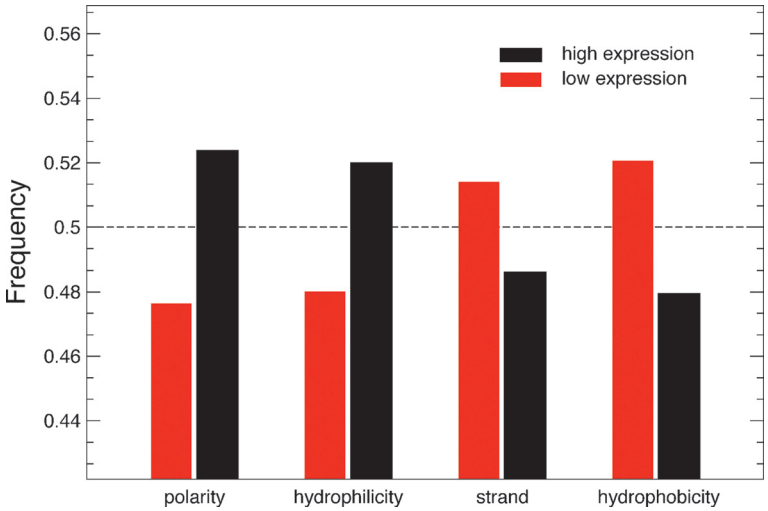


**Fig. 3** Relationship between gene expression levels and the chemical properties of the corresponding protein sequences. High expression levels are associated with protein sequences with large numbers of polar (D, E, N, R, Q, and H) and hydrophilic (T, K, C, G, A, and R) amino acids. Low expression levels are observed for protein sequences with large numbers of amino acids with high hydrophobicity (F, M, I, L, V, Y, and P), and high β-sheet propensity (V, I, F, W, L, Y, and T). These results were obtained by sorting the expression levels of about 2000 *E. coli* genes (http://redpoll.pharmacy.ualberta.ca/CCDB/) and by selecting the 50 top-ranked and 50 bottom-ranked genes. The horizontal line represents the average value for the two datasets taken together. Chemical propensities are defined following a consensus of experimental scales (http://expasy.org/tools/protscale.html).

of proteins in *E. coli*[8–12,22] are also employed for predicting aggregation propensities of proteins *in vitro*.[13–20] These results hint at the existence of a link between aggregation rates measured *in vitro* and concentrations detected *in vivo*.

By investigating the nature of this link, we have reported a strong relationship between *in vivo* mRNA expression levels and *in vitro* protein aggregation rates[7] (Fig. 2). We suggested that this correlation is the consequence of an evolutionary pressure acting to decrease the risk of aggregation in the cell.[7] This evolutionary pressure is exerted since failure of proteins to fold correctly can give rise to cellular malfunctions and diseases. Random amino acid mutations have a strong tendency to destabilize the native fold of a protein[25–27] and enhance the propensity to aggregate.[28] Since protein concentrations and mRNA expression levels are correlated in *E. coli*,[29] the relationship that we identified indicates a close correlation between concentrations and solubility of proteins (Fig. 1).

Since the aggregation propensities of proteins can be predicted from their chemical properties.[8–20] and the aggregation propensities themselves are correlated to the *in vivo* concentrations,[7] it should also be possible to make predictions about a

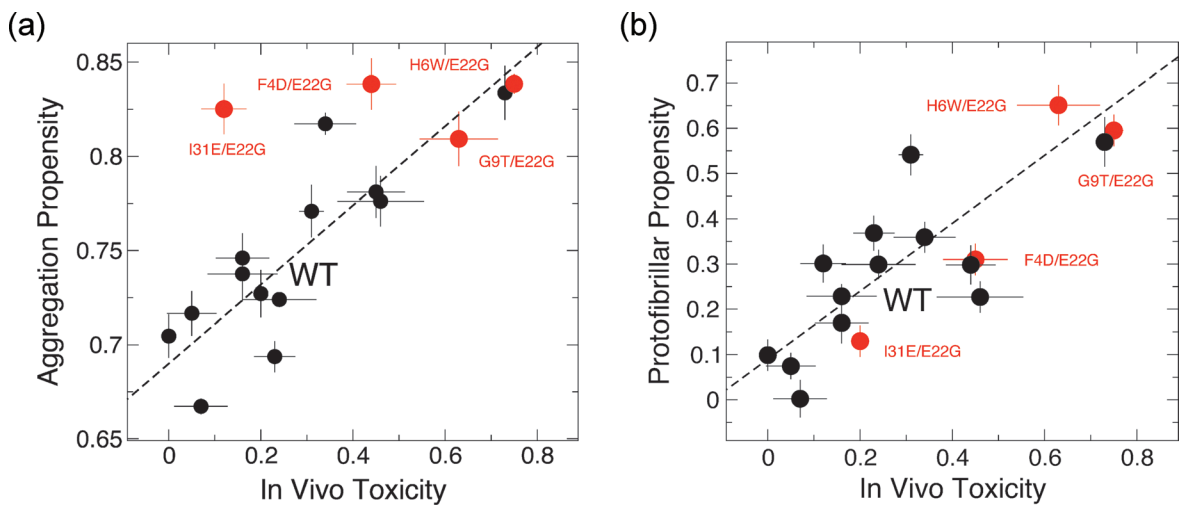


**Fig. 4** The toxicity of mutational variants of the Aβ peptide can be better predicted using the propensities to form protofibrillar assemblies rather than the propensities to form fibrillar aggregates.[21] (a) Correlation between aggregation propensity and neurotoxicity in *Drosophila* of Aβ mutational variants. The Zyggregator algorithm[19,20] (http://www-vendruscolo.ch.cam.ac.uk/zyggregator.php) was used to rationally design 17 variants of the Aβ peptide on the basis of their aggregation propensities. The toxicity of the variants, measured as deficiencies in the locomotor activity, shows a correlation of 75% with the predicted aggregation propensities.[21] Mutants of the E22G "Arctic" variant (red circles) diverge from the trend defined by the mutants of the wild-type Aβ (black circles). Notably, different fibril morphologies are associated with the E22G variants, which aggregate into low molecular weight protofibrils. Taken together, these results show that the algorithm used for predicting aggregation propensities, while capable of predicting the *in vitro* aggregation rates of all the mutants considered, including those on the E22G background,[21] is unable to fully capture their toxic effects. (b) Correlation between the predictions of the propensities to form protofibrillar aggregates and toxicities measured *in vivo* in transgenic *Drosophila*.[21] The toxicity of the variants, measured as deficiencies in the locomotor activity, shows a correlation of 83% with the predicted protofibrillar propensities, which is significantly higher than the one obtained using the aggregation propensities. In particular, mutations on the E22G background (red circles) do not diverge significantly from the trend given by the wild type mutants (black circles).

variety of cellular processes by using the chemical properties of the proteins involved in them. In the next two sections we provide two examples that show how this idea can be applied.

## Relationship between gene expression levels and amino acid composition of the corresponding proteins

We first illustrate the close connection between protein chemistry and cell biology by showing that gene expression levels and the amino acid compositions of the corresponding protein sequences are linked in *E. coli* (Fig. 3). Highly-expressed genes are associated with protein sequences with high charge and hydrophilicity, while genes with low levels of expression are associated with protein sequences with high hydrophobicity and propensity to form β-sheets (Fig. 3). Consistent with these results, proteins with many polar amino acids represent an optimal target for heterologous expression in *E. coli*.[30] In addition, membrane proteins, which are characterized by the presence of several extended hydrophobic regions, show very low levels of expression in *E. coli*.[31] Furthermore, natively unfolded proteins tend to have more charged and fewer hydrophobic amino acids[32] in order to increase their solubility in the absence of the protection against aggregation provided by a folded structure,[19] but despite these features their expression levels are in some cases lower than those of globular proteins,[33] thus indicating the key role played by the native structure in maintaining the solubility of polypeptide chains.[19]

## Protein concentrations and neurotoxicity of Aβ peptide aggregates in *Drosophila melanogaster*

The possibility of using protein chemistry to make quantitative predictions about cellular processes is further illustrated here by reviewing a study in which the aggregation propensities of a series of mutational variants of the Aβ peptide were related to the lifespan of transgenic *Drosophila melanogaster*.[21] In that study the aggregation propensities of all the possible 798 single-point mutations of the Aβ peptide were predicted. Thirteen non-naturally occurring variants were then selected according to such propensities for expression in the central nervous system of *Drosophila*. In addition, four mutational variants of the E22G ("Arctic") form of the Aβ peptide, which leads to early onset of Alzheimer's disease in humans,[34] were designed to modulate its aggregation propensity. It has been shown by *in vitro* experiments that the E22G mutant forms aggregates and fibrils at much lower concentrations ($<2.5$ μM) than the wild type (12.5 μM).[35] At the same concentration, the aggregation process occurs more rapidly for the E22G mutant compared with the wild type protein and amyloid fibrils appear about four days earlier after the onset of sample incubation without seeding.[36]

Using the *Drosophila* longevity to quantify the effect of Aβ variants, a correlation of 75% was found between the predicted propensities to aggregate and the severity of neurodegeneration (Fig. 4a). Although the transcription levels of the wild type Aβ and of the E22G variant were found to be practically the same, the lifespan of flies expressing the wild-type Aβ peptide was observed to be about 45 days, while the lifespan of the flies expressing the E22G variant was reduced to about 10 days.[21] We also found that flies expressing the F20E mutant lived 25% longer than flies expressing the wild-type Aβ peptide, and that the F20E mutant does not form *in vivo* deposits, even when its transcription levels are three times higher than those of the wild-type Aβ peptide. High correlation was also found between predicted aggregation propensities and locomotor deficits. Some deviations from this trend were, however, observed. The I31E mutation on the E22G background (I31E/E22G) exhibited neuronal effects that did not match the predicted aggregation propensity. Intriguingly, the I31E/E22G peptide was found to aggregate *in vitro* at a rate that matches the predictions but its deposits were not accompanied by cavities in brain tissue. Moreover, even when the transcription levels of the I31E/E22G peptide are twice as much as the levels of the E22G peptide, the protein was not observed to be toxic. Importantly, the case of the F20E and I31E/E22G variants suggests that non-toxic proteins can be expressed at higher levels without impairing cellular processes, which strongly supports our observation of a link between aggregation rates and expression levels.

In order to rationalize the deviations of some of the mutants that we observed, we considered the series of experimental studies that have shown for a range of peptides and proteins that amyloid fibril formation is preceded by the appearance of organized molecular assemblies usually termed protofibrils.[37,38] Interest in these low molecular weight oligomers has increased since these species have been detected in the brains of patients suffering from Alzheimer's disease.[39,40] In fact, at different stages of Aβ assembly, Aβ oligomers and fibrils are observed that differ in structure as well as toxic function. In particular, earlier oligomeric assemblies have been found to be toxic to cells in cell cultures and in a transgenic mouse model.[41,42] Hence, we investigated the relationship between neurodegenerative effects of the Aβ variants and their associated protofibrillar propensities.[21] In order to build a predictor for protofibrillar propensities, we used the aggregation rates of the mutants of the protein AcP[43] to parametrize an equation that has the same form as the one used for prediction of aggregation propensities. In the new predictions, we assumed that regions characterized by either positive or negative propensities are important to predict the protofibrillar propensity. We made this hypothesis because regions associated with negative propensities are usually charged or have a high propensity for being unstructured, which may play a key role at the early stages of the oligomerization process.[44–46] In the approach that we have described, the chemical properties that we use for predicting the aggregation rates into fibrillar or protofibrillar assemblies are similar. As we, however, determined the parameters through separate fitting procedures in which the aggregation rates into specific types of assemblies were measured, the relative weights of the chemical properties are different, thus resulting in specific predictions for each type of assembly process.[20]

The predicted protofibrillar propensities showed a correlation of 83% with the fly longevity, revealing a stronger

relationship with the *in vivo* toxicity than the aggregation propensity (Fig. 4b). These results show not only that predictions made on the basis of the chemical properties of Aβ peptide variants can account quantitatively for the observed changes in the lifespan of transgenic flies expressing such peptides in their central nervous system, but also that this type of prediction can provide information about the nature of the molecular species having the most serious neurotoxic effects.

## Conclusions

We have described close relationships between the chemical properties of proteins and their solubility, expression levels and aggregation propensities. These results suggest that strategies for predicting the outcome of cellular processes based on chemical properties of proteins can be used as an alternative to those based on the analysis of the cellular processes themselves.[47–51]

We have illustrated this approach by showing that neurotoxic effects of misfolded protein assemblies in a living organism can be predicted on the basis of the analysis of the chemical properties of proteins that determine their propensities to form protofibrillar aggregates. Thus, although cells have evolved complex regulatory mechanisms to supervise folding and combat misfolding, the intrinsic tendency of polypeptide chains to aggregate plays a key role in determining the pathological consequences of their aberrant assemblies.

## Acknowledgements

We thank Dr Sebastian Pechmann, Dr D. Crowther, Dr L. Luheshi, and Prof. C. M. Dobson for stimulating discussions. This work was supported by the Medical Research Council, the Wellcome Trust, and Leverhulme Trust.

## References

1 M. Levine and R. Tjian, Transcription regulation and animal diversity, *Nature*, 2003, **424**, 147–151.
2 E. H. Davidson, D. R. McClay and L. Hood, Regulatory gene networks and the properties of the developmental process, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 1475–1480.
3 R. L. Strausberg and S. L. Schreiber, From knowing to controlling: a path from genomics to drugs using small molecule probes, *Science*, 2003, **300**, 294–295.
4 B. R. Stockwell, Chemical genetics: ligand-based discovery of gene function, *Nat. Rev. Genet.*, 2000, **1**, 116–125.
5 C. M. Dobson, Chemical space and biology, *Nature*, 2004, **432**, 824–828.
6 S. Ding and P. G. Schultz, A role for chemistry in stem cell biology, *Nat. Biotechnol.*, 2004, **22**, 833–840.
7 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.
8 D. L. Wilkinson and R. G. Harrison, Predicting the solubility of recombinant proteins in *Escherichia coli*, *Bio/Technology*, 1991, **9**, 443–448.
9 C. H. Schein, Solubility as a function of protein structure and solvent components, *Bio/Technology*, 1990, **8**, 308–315.
10 C. S. Goh, N. Lan, S. M. Douglas, B. Wu, N. Echols, A. Smith, D. Milburn, G. T. Montelione, H. Zhao and M. Gersterin, Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis, *J. Mol. Biol.*, 2004, **336**, 115–130.
11 S. Idicula-Thomas, A. J. Kulkarni, B. D. Kulkarni, V. K. Jayaraman and P. V. Balaji, A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in *Escherichia coli*, *Bioinformatics*, 2006, **22**, 278–284.
12 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.
13 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.
14 A. Trovato, F. Chiti, A. Maritan and F. Seno, Insight into the structure of amyloid fibrils from the analysis of globular proteins, *PLoS Comput. Biol.*, 2007, **2**, 1608–1618.
15 G. G. Tartaglia, A. Pawar, S. Campioni, F. Chiti, C. M. Dobson and M. Vendruscolo, Prediction of aggregation-prone regions in structured proteins, *J. Mol. Biol.*, 2008, **380**, 425–436.
16 G. G. Tartaglia and M. Vendruscolo, The Zyggregator method for predicting protein aggregation propensities, *Chem. Soc. Rev.*, 2008, **37**, 1395–1401.
17 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.
18 G. D. Davis, C. Elisee, D. M. Newham and R. G. Harrison, New fusion protein systems designed to give soluble expression in *Escherichia coli*, *Biotechnol. Bioeng.*, 1999, **65**, 382–388.
19 F. Chiti and C. M. Dobson, Protein misfolding, functional amyloid, and human disease, *Annu. Rev. Biochem.*, 2006, **75**, 333–366.
20 C. M. Dobson, Protein misfolding, evolution and disease, *Trends Biochem. Sci.*, 1999, **24**, 329–332.
21 A. A. Pakula and R. T. Sauer, Genetic analysis of protein stability and function, *Annu. Rev. Genet.*, 1989, **23**, 289–310.
22 K. A. Bava, M. M. Gromiha, H. Uedaira, K. Kitajima and A. Sarai, ProTherm, version 4.0: thermodynamic database for proteins and mutants, *Nucleic Acids Res.*, 2004, **32**, D120–D121.
23 R. Godoy-Ruiz, R. Perez-Jiminez, B. Ibarra-Molero and J. M. Sanchez-Ruiz, Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations, *J. Mol. Biol.*, 2004, **336**, 313–318.
24 Z. Ignatova, Monitoring protein stability *in vivo*, *Microb. Cell Fact.*, 2005, **4**, 23.
25 P. Lu, C. Vogel, R. Wang, X. Yao and E. M. Marcotte, Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation, *Nat. Biotechnol.*, 2007, **25**, 117–124.
26 K. Büssow, C. Scheich, V. Sievert, U. Harttig, J. Schultz, B. Simon, P. Bork, H. Lehrach and U. Heinemann, Structural genomics of human proteins—target selection and generation of a public catalogue of expression clones, *Microb. Cell Fact.*, 2005, **4**, 21.
27 S. Sundararaj, A. Guo, B. Habibi-Nazhad, M. Rouani, P. Stothard, M. Ellison and D. S. Wishart, The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of *Escherichia coli*, *Nucleic Acids Res.*, 2004, **32**, D293–D295.
28 V. N. Uversky, Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go?, *Cell. Mol. Life Sci.*, 2003, **60**, 1852–1871.
29 A. Mohan, W. J. Sullivan, Jr, P. Radivojac, A. K. Dunker and V. N. Uversky, Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes, *Mol. BioSyst.*, 2008, **4**, 328–340.
30 K. Sennvik, C. Nilsberth, C. Stenh, L. Lannfelt and E. Bendedikz, The Arctic Alzheimer mutation enhances sensitivity to toxic stress in human neuroblastoma cells, *Neurosci. Lett.*, 2002, **326**, 51–55.
31 O. N. Antzutkin, Amyloidosis of Alzheimer's Abeta peptides: solid-state nuclear magnetic resonance, electron paramagnetic resonance, transmission electron microscopy, scanning transmission electron microscopy and atomic force microscopy studies, *Magn. Reson. Chem.*, 2004, **42**, 231–246.
32 J. Meinhardt, G. G. Tartaglia, A. Pawar, T. Christopeit, P. Hortschansky, V. Schroeckh, C. M. Dobson, M. Vendruscolo and M. Fändrich, Similarities in the thermodynamics and kinetics

 

of aggregation of disease-related Abeta(1-40) peptides, *Protein Sci.*, 2007, **16**, 1214–1222.
33 J. D. Harper, S. S. Wong, C. M. Lieber and P. T. Lansbury, Observation of metastable Abeta amyloid protofibrils by atomic force microscopy, *Chem. Biol.*, 1997, **4**, 119–125.
34 R. Bader, R. Bamford, J. Zurdo, B. F. Luisi and C. M. Dobson, Probing the mechanism of amyloidogenesis through a tandem repeat of the PI3-SH3 domain suggests a generic model for protein aggregation and fibril formation, *J. Mol. Biol.*, 2006, **356**, 189–208.
35 A. E. Roher, M. O. Chaney, Y. M. Kuo, S. D. Webster, W. B. Stine, L. J. Haverskamp, A. S. Woods, R. J. Cotter, J. M. Tuohy, G. A. KRafft, B. S. Bonnel and M. R. Emmerling, Morphology and toxicity of Abeta-(1-42) dimer derived from neuritic and vascular amyloid deposits of Alzheimer's disease, *J. Biol. Chem.*, 1996, **271**, 20631–20635.
36 L. F. Lue, Y. M. Kuo, A. E. Roher, L. Brachova, Y. Shen, L. Sue, T. Beach, J. H. Kurth, R. E. Rydel and J. Rogers, Soluble amyloid beta peptide concentration as a predictor of synaptic change in Alzheimer's disease, *Am. J. Pathol.*, 1999, **155**, 853–862.
37 M. D. Kirkitadze, G. Bitan and D. B. Teplow, Paradigm shifts in Alzheimer's disease and other neurodegenerative disorders: the emerging role of oligomeric assemblies, *J. Neurosci. Res.*, 2002, **69**, 567–577.
38 M. P. Lambert, A. K. Barlow, B. A. Chromy, C. Edwards, C. Freed, M. Liosatos, T. E. Morgan, I. Rozovsky, B. Trommer, K. L. Viola, P. Wals, C. Zhang, C. E. Finch, G. A. Krafft and W. L. Klein, Diffusible, nonfibrillar ligands derived from Abeta1-42 are potent central nervous system neurotoxins, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 6448–6453.
39 F. Chiti, M. Calamai, N. Taddei, M. Stefani, G. Ramponi and C. M. Dobson, Studies of the aggregation of mutant proteins *in vitro* provide insights into the genetics of amyloid diseases, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 16419–16426.
40 T. Lührs, C. Ritter, M. Adrian, D. Riek-Loher, B. Bohrmann, H. Döbeli, D. Schubert and R. Riek, 3D structure of Alzheimer's amyloid-beta(1-42) fibrils, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 17342–17347.
41 J. H. Brown, N. Volkmann, G. Jun, A. H. Henschen-Edman and C. Cohen, The crystal structure of modified bovine fibrinogen, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 85–90.
42 F. Massi, D. Kilmov, D. Thirumalai and J. E. Straub, Charge states rather than propensity for beta-structure determine enhanced fibrillogenesis in wild-type Alzheimer's beta-amyloid peptide compared to E22Q Dutch mutant, *Protein Sci.*, 2002, **11**, 1639–1647.
43 D. H. Lackner, T. H. Beilharz, S. Marguerat, J. Mata, S. Watt, F. Shubert, T. Preiss and J. Bäler, A network of multiple regulatory layers shapes gene expression in fission yeast, *Mol. Cell*, 2007, **26**, 145–155.
44 D. H. Nguyen and P. D'haeseleer, Deciphering principles of transcription regulation in eukaryotic genomes, *Mol. Syst. Biol.*, 2006, **2**, 1–9.
45 E. Segal, T. Ravek-Sadka, M. Schroeder, U. Unnerstall and U. Gaul, Predicting expression patterns from regulatory sequence in *Drosophila* segmentation, *Nature*, 2008, **451**, 535–540.
46 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.
47 M. A. Beer and S. Tavazoie, Predicting gene expression from sequence, *Cell*, 2004, **117**, 185–198.