EMBO reports

review

Prediction of amyloid aggregation *in vivo*

Mattia Belli, Matteo Ramazzotti & Fabrizio Chiti[+]

Department of Biochemical Sciences, University of Florence, Firenze, Italy

Many human diseases owe their pathology, to some degree, to the erroneous conversion of proteins from their soluble state into fibrillar, β-structured aggregates, often referred to as amyloid fibrils. Neurodegenerative diseases, such as Alzheimer and spongiform encephalopathies, as well as type 2 diabetes and both localized and systemic amyloidosis, are among the conditions that are associated with the formation of amyloid fibrils. Several mathematical tools can rationalize and even predict important parameters of amyloid fibril formation. It is not clear, however, whether such algorithms have predictive powers for *in vivo* systems, in which protein aggregation is affected by the presence of other biological factors. In this review, we briefly describe the existing algorithms and use them to predict the effects of mutations on the aggregation of specific proteins, for which *in vivo* experimental data are available. The comparison between the theoretical predictions and the experimental data obtained *in vivo* is shown for each algorithm and experimental data set, and statistically significant correlations are found in most cases.

Keywords: prediction; protein misfolding; software; TANGO; Zyggregator

EMBO *reports* (2011) **12,** 657–663. doi:10.1038/embor.2011.116

See Glossary for abbreviations used in this article.

Introduction

Biological fluids—be they bacterial or eukaryotic cytosols, or the plasma of superior mammals—are colloidal, aqueous solutions in which various proteins are present at high concentrations. One of the properties of a colloidal solution is the aggregation of its constituent particles into large assemblies that affect cell and tissue function. In the case of peptides and proteins, the resulting aggregates generally consist of well-organized fibrillar assemblies, often referred to as amyloid fibrils. In these structures the various polypeptide chains are held together by continuous β-sheet structures running along the fibrils, in which the β-strands are perpendicular to the fibril axes (Chiti & Dobson, 2006).

From a physicochemical perspective, amyloid fibril formation represents an essential feature of the behaviour of polypeptide chains. Understanding this behaviour is vital for a thorough characterization of the nature of proteins—for example to decipher their folding code—and is a worthwhile pursuit, even without reference to human disease (Jahn & Radford, 2008). In addition, aggregation is a problem in biotechnology, in which the large-scale expression of proteins that are of potential interest to the market often results in their self-assembly into amyloid-like structures in inclusion bodies (Ventura & Villaverde, 2006). From a reversed biotechnological perspective, amyloid fibril formation is an important source of innovation, as huge numbers of new materials can be designed by exploiting the amyloid motif and the inherent diversity of protein sequences (Cherny & Gazit, 2008).

In biology, the aggregation of peptides and proteins into both structured amyloid fibrils and more amorphous aggregates is a constant challenge for living organisms. To prevent, dedicated systems have evolved that range from inherent sequence and structural adaptations in natural proteins that directly inhibit their own aggregation, to more complex cellular machineries such as the heat-shock response, the unfolded-protein response, endoplasmic-reticulum-associated degradation and autophagy, among others (Monsellier & Chiti, 2007; Voellmy & Boellmann, 2007; Kapoor & Sanyal, 2009; Hoseki *et al*, 2009; Bejarano & Cuervo, 2010). The failure of proteins and peptides to remain soluble can result in pathology and many human diseases are associated with the formation of fibrillar aggregates: neurodegenerative conditions (such as Alzheimer disease, Parkinson disease and spongiform encephalopathies), systemic amyloidoses (such as light chain amyloidosis and dialysis-related amyloidosis), and localized amyloidosis (such as type 2 diabetes and atrial amyloidosis; Chiti & Dobson, 2006).

Given the importance of amyloid fibril formation to protein chemistry, biotechnology, biology and medicine, it is important to elucidate the mechanisms by which proteins are converted from their soluble states into amyloid fibrils. In this context, improved knowledge of the structure of amyloid fibrils and the forces that promote and stabilize their formation has given rise to several mathematical tools that allow the fundamental aspects of the aggregation process to be predicted. These include identifying the regions of the sequence of a protein that promote amyloid fibril formation and the effect of mutations.

The ability to predict amyloid fibril formation is important for several reasons in different scenarios. For biotechnology, accurately predicting aggregation allows the design of new proteins—or modified versions of existing ones—with properly modulated aggregation tendencies to avoid or enhance aggregation, depending on the intended application (Ventura & Villaverde, 2006; Cherny & Gazit 2008). In medicine, many of the genetic diseases associated with fibril formation result from mutations in the sequence of the aggregating protein. Therefore, the ability to predict the effect of mutations on the

Department of Biochemical Sciences, University of Florence, Viale Morgagni 50, 50134 Firenze, Italy
[+]Corresponding author. Tel: +39 055 4598319; Fax: +39 055 4598905;
E-mail: fabrizio.chiti@unifi.it

Received 10 December 2010; accepted 26 May 2011; published online 17 June 2011

Table 1 | A survey of available methods and programs for predicting amyloid aggregation

Algorithm	Class	Prediction provided	Usability*	Reference
Chiti & Dobson	Empirical	Relative aggregation rate after mutation	Equation	Chiti *et al*, 2003
Tartaglia *et al*	Empirical	Relative aggregation rate after mutation	Equation	Tartaglia *et al*, 2004
DuBay *et al*	Empirical	Absolute aggregation rate	Equation	DuBay *et al*, 2004
Net-CSSP	Structure-based	Aggregation-prone regions	Server	Yoon & Welsh, 2004
TANGO	Empirical	Aggregation-prone regions	Software, server	Fernandez-Escamilla *et al*, 2004
Pawar *et al*	Empirical	Aggregation-prone regions	Equation	Pawar *et al*, 2005
3D profile	Structure-based	Aggregation-prone regions	On demand	Thompson *et al*, 2006
PASTA	Structure-based	Aggregation-prone regions, orientation of the β-strands	Software, server	Trovato *et al*, 2006
FoldAmyloid	Structure-based	Aggregation-prone regions	Server	Galzitskaya *et al*, 2006
AGGRESCAN	Empirical	Aggregation-prone regions	Server	Conchillo-Solé *et al*, 2007
SALSA	Empirical	Aggregation-prone regions	Equation	Zibaee *et al*, 2007
Zyggregator	Empirical	Aggregation-prone regions	Server, upon request	Tartaglia & Vendruscolo, 2008
BETASCAN	Structure-based	Aggregation-prone regions	Server	Bryan *et al*, 2009
Waltz	Structure-based	Aggregation-prone regions	Server	Maurer-Stroh *et al*, 2010

*Equation, only the mathematical description is available; software, dedicated software was developed and is freely downloadable; server, dedicated software was developed and is hosted by a server, not downloadable. See supplementary Table 1 for greater detail of the *modus operandi* and development of each method.

Glossary

$A\beta_{42}$	42-residue-long amyloid-β peptide
AcP	human muscle acylphosphatase
GFP	green fluorescent protein
HypF-N	amino-terminal domain of the HypF protein from *E. coli*
Net-CSSP	neural networks for calculating contact-dependent secondary structure propensity
PASTA	prediction of amyloid-structure aggregation
SALSA	simple algorithm for sliding averages
SOD1	superoxide dismutase type 1
TFE	2,2,2-trifluoroethanol

propensity for amyloid formation also allows the elucidation of whether such mutations are pathogenic because they increase the inherent propensity of the associated protein to aggregate, or for other reasons (Sekijima *et al*, 2005). From a protein-chemistry perspective, there is an interesting bonus arising from the development of increasingly accurate predictive tools: the factors on which the accurate tools are based must be relevant to amyloid formation and can therefore be flagged up for research in the field. In addition, the availability of computational tools should allow for the rapid and systematic analysis of full proteomes, the comparison of categories of proteins and of proteins from different organisms, as well as the statistical analysis of strategies that have arisen to counteract aggregation throughout evolution (Rousseau *et al*, 2006; Monsellier *et al*, 2008).

The birth of predictive algorithms

The first mathematical tool shown to yield predictions consistent with *in vitro* experimental data was published in 2003 (Chiti *et al*, 2003). In this study, the aggregation rates of several single-point mutants of the model protein human muscle acylphosphatase (AcP) were compared with those of the wild-type protein under the same conditions. The process for both wild-type and mutant AcP was the conversion of TFE-denatured AcP into β-sheet-containing, thioflavine-T-binding and Congo-red-binding protofibrils. The change in the rate of aggregation of the mutants was found to correlate with changes in their hydrophobicity, propensity to form α-helical structure, propensity to form β-sheet structure at the site of the mutation, and overall charge. These intrinsic factors were combined to produce an empirical formula to predict the effect of a given mutation on the aggregation rate, as a function of the changes in these parameters after mutation. When tested on a range of mutants of different peptides and proteins (all unstructured or intrinsically disordered), the formula was found to yield accurate predictions. The algorithm was empirical and simple, but only valid for unstructured peptides and proteins and only able to predict the effect of a given mutation on amyloid fibril formation or β-sheet aggregation generally, not on other observable factors. Despite its limitations, the formula paved the way for work based on the idea that amyloid fibril formation follows generic rules that can be rationalized.

Today, approximately 14 computational tools have been published that aim to extend this predictive power to more difficult subjects (Table 1). These include predicting the aggregation behaviour of initially folded proteins, as well as other observable characteristics of protein aggregation such as the absolute rate of amyloid fibril elongation or the aggregation-promoting regions within a given sequence.

The predictive models can be divided into two main categories: empirical and structure-based. Empirical tools try to explain experimental results and make predictions by identifying and considering the appropriate factors, either individually or in combination, mainly of the amino acid properties (for example, hydrophobicity, β-propensity and solubility). By contrast, models based on structure try to identify the determinants of amyloid aggregation by observing the existing three-dimensional (3D) structures of peptides that adopt a known fibrillar structure or native proteins that belong to distinct structural classes. Table 1 lists the algorithms and references the

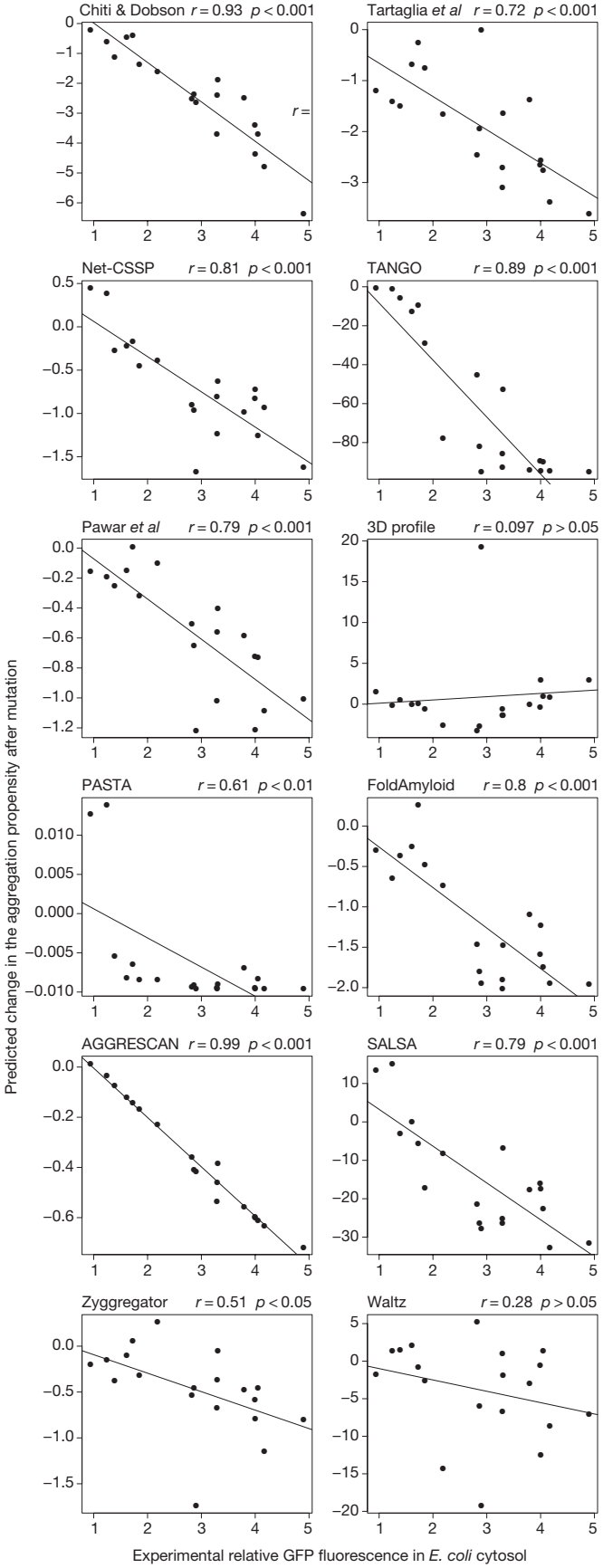


Fig 1 | Predicted change in the aggregation propensity after mutation compared with *in vivo* experimental solubility in *Escherichia coli* cytosol for $A\beta_{42}$ variants. Each graph reports the predicted change in the aggregation propensity after mutation (calculated according to the algorithm indicated in each graph and following the procedure described in the supplementary information online) compared with experimental relative fluorescence of GFP fused to $A\beta_{42}$ mutants in *E. coli* cytosol, as described previously (de Groot *et al*, 2006). A value of 0 on the *y*-axis corresponds to a value of 1 on the *x*-axis (no change, relative to wild type). Scales on the *y*-axis have been adjusted to show, for each plot, the full data set. The lines represent the best fits of the data to linear functions. For each plot, the name of the algorithm used is reported, as well as the absolute value of the Pearson linear correlation coefficient (*r*) and the statistical significance of the slope (*p*). $A\beta_{42}$, 42-residue-long amyloid-β peptide; GFP, green fluorescent protein. ►

reports in which the algorithms were first described. A full description of the algorithms and of the observations they predict is available in supplementary Table 1 online. The algorithms that have been proposed so far have each been tested on several polypeptide sequences known to aggregate into amyloid-like fibrils *in vitro*. In each publication, the predictions provided by the algorithm in question fit well with the experimental data. Some of the algorithms, such as Zyggregator, are refinements of older versions; others have combined and developed pre-existing algorithms, such as Waltz.

Measurement of amyloid formation *in vivo*

All algorithms published so far have been inspired by and tested against experimental data obtained mainly *in vitro*, where the protein or peptide being studied aggregates in a buffered solution in the absence of other proteins or biological factors. However, amyloid formation phenomena that are relevant to biology, medicine and even biotechnology occur in complex biological environments. In such biological contexts, the presence of several agents has been shown to affect—often dramatically—protein aggregation. These agents include lipids, glycosaminoglycans, collagen fibres, molecular chaperones, proteases, apolipoprotein E, the serum amyloid P component, metal ions and probably many other as-yet-unidentified agents. The question thus arises whether the predictive power embodied in the existing algorithms can be extended to amyloid formation in a biological context, or whether it is limited to simplified *in vitro* systems.

The answer to this question is not trivial, particularly because the quantitative measurement of the aggregation rates or propensities of proteins *in vivo* is technically challenging. Nevertheless, several papers have been published with data describing the aggregation propensities of wild-type and mutated proteins in *Escherichia coli* (Wurth *et al*, 2002; Kim & Hecht, 2006; de Groot *et al*, 2006; Winkelmann *et al*, 2010). These four reports use two techniques to detect aggregates in the *E. coli* cytosol.

The first technique is based on the observation that GFP can be used as a reporter for protein aggregation, as its structure is destabilized—and its associated fluorescence intensity is reduced—if the protein fused to its amino-terminus is aggregated (Wurth *et al*, 2002). By directly measuring the total fluorescence of the bacterial culture, the authors were able to determine the levels of aggregation of several variants of the $A\beta_{42}$ peptide in a quantitative, rapid and inexpensive manner. In the second approach, single-point mutants of the HypF-N protein—all truncated at the carboxy-terminus to

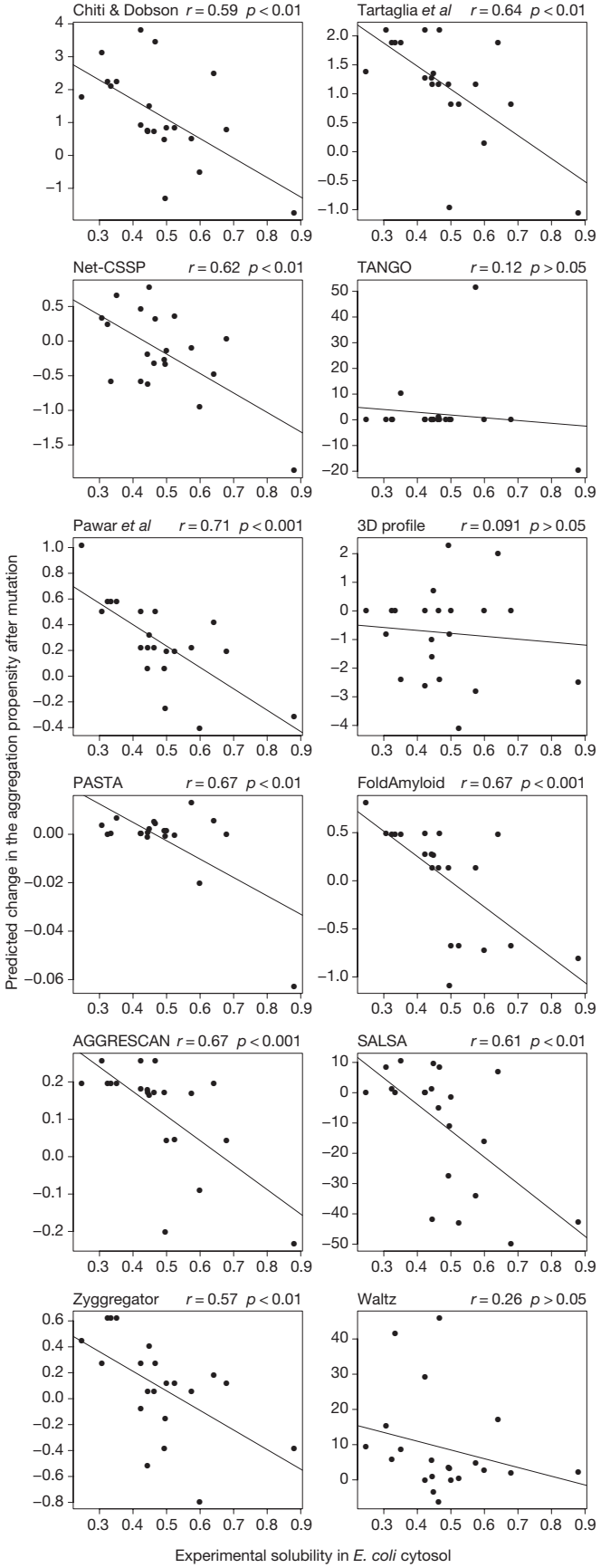


Fig 2 | Predicted change in the aggregation propensity after mutation compared with *in vivo* experimental solubility in *Escherichia coli* cytosol for HypF-N variants. Each graph reports the predicted change in the aggregation propensity after mutation (calculated according to the algorithm indicated in each graph and following the procedure described in the supplementary information online) compared with experimental solubility in *E. coli* cytosol (I_{SN}/I_P) of the mutants of HypF-N, as described previously (Winkelmann *et al*, 2010). A value of 0 on the *y*-axis corresponds to a value of 0.5 on the *x*-axis (no change, relative to wild type). Scales on the *y*-axis have been adjusted to show, for each plot, the full data set. The lines represent the best fits of the data to linear functions. For each plot, the name of the algorithm used is reported, as well as the absolute value of the Pearson linear correlation coefficient (*r*) and the statistical significance of the slope (*p*).

prevent folding—were expressed under strict, low-level expression conditions (Winkelmann *et al*, 2010). The authors were able to produce at the desired time very low levels of mutated HypF-N (only detectable by western blotting) in order to preserve the original cellular machinery dedicated to the maintenance of protein homeostasis. By using this system, the authors were able to quantify the ratio of protein in the supernatant and pellet fractions (I_{SN}/I_P) for each expressed variant—a quantitative inverse correlate of the fraction of protein aggregating *in vivo*.

Both of these methods rely on the accumulating evidence that bacterial inclusion bodies consist of amyloid-like material as a result of the misfolding and aggregation of cytoplasmic proteins (Carrió *et al*, 2005; Wang, L. *et al*, 2008). Both methods are based on the quantitative measurement of mutant protein levels in inclusion bodies relative to wild type. They cannot determine absolute aggregation rates of expressed proteins or aggregation-promoting regions, but can quantitatively determine the effect of a given mutation on the propensity to form amyloid-like fibrils for a given protein expressed *in vivo*.

Predictions correlate with aggregation *in vivo*

This ability to monitor amyloid formation in *E. coli* allows the assessment of the rules and algorithms determined or generated from *in vitro* studies. This is achieved by applying the predictive algorithms to the data sets generated from the expression of mutants *in vivo* and comparing the predictions with the corresponding propensities to form inclusion bodies determined experimentally.

For the same group of $A\beta_{42}$ variants (those of de Groot *et al*, 2006), Fig 1 shows plots of the change in the aggregation propensity predicted to result from mutation, against the measured fluorescence intensity of the fused GFP protein, where high fluorescence indicates low aggregation of $A\beta_{42}$. All 12 plots in Fig 1 use the same set of experimental data on the *x*-axis, but different sets of predicted values on the *y*-axis. Each plot refers to a well-defined algorithm listed in Table 1. excluding DuBay *et al* (2004), which is redundant with respect to Pawar *et al* (2005), and BETASCAN, which could not provide quantitative values suitable for the analysis. For those algorithms for which the method used to predict the effect of a given mutation on aggregation was not explained in the original report, the procedure we used to determine the predicted change in aggregation propensity of $A\beta_{42}$ following mutation, as well as the rationale behind the comparison with experimental parameters, are described in the supplementary information online. For the HypF-N mutants—those of Winkelmann *et al* (2010)—Fig 2 show plots of the predicted change in the aggregation propensity after mutation, against the solubility of

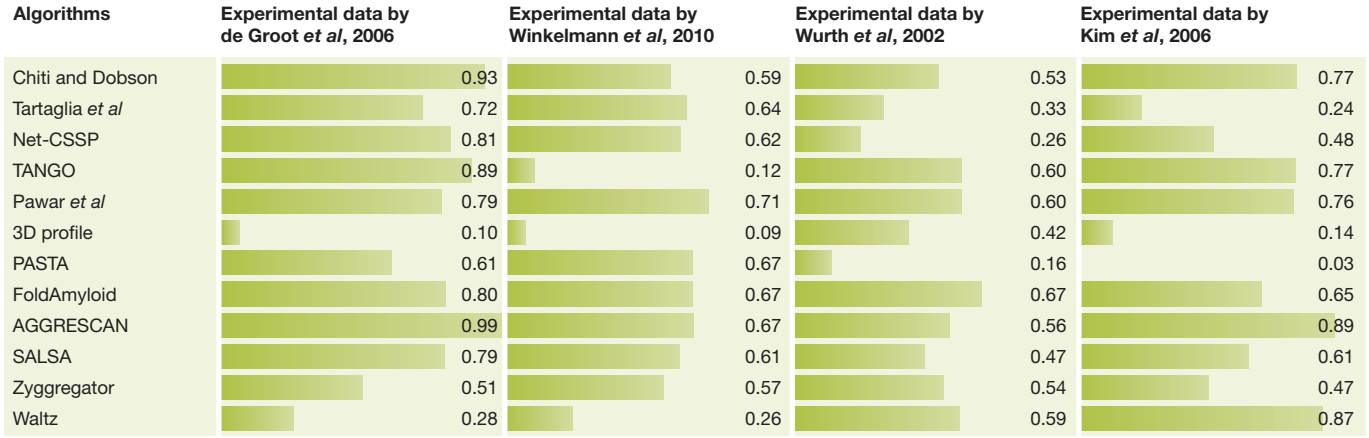


Fig 3 | Correlation between the change in aggregation propensity observed *in vivo* and the change in aggregation propensity predicted *in silico* by algorithms. For each algorithm and each data set, the absolute value of the Pearson linear correlation coefficient (*r*) is reported as both a numeric value and a horizontal bar.

the mutated proteins in the *E. coli* cytosol (I_{SN}/I_P), which is an inverse correlate of aggregation. The same analyses performed on the two remaining experimental data sets—that is, Wurth *et al* (2002) and Kim & Hecht (2006)—are reported in supplementary Figs S1 and S2.

For each of the four experimental data sets and for each of the 12 algorithms analysed, a plot of predicted versus experimental values was produced and analysed by linear regression, obtaining the linear correlation coefficient (*r*) and the *p* value associated with the slope of the regression line. The resulting *p* values indicate that in most cases the correlations are significant ($p < 0.05$). This indicates that most of the algorithms are able to predict the effect of a mutation on aggregation propensity *in vivo*.

Comparison between algorithms

Among the algorithms considered in this review, those initially designed to generate aggregation-propensity profiles and identify aggregation-prone regions (Table 1; supplementary Table 1) were not optimized to determine the effect of mutations, so a direct comparison with algorithms that aim to predict the effect of mutations on aggregation rates or propensity does not favour them. In addition, only four experimental data sets containing quantitative determination of amyloid-like aggregation *in vivo* are available so far. All are in *E. coli* and three involve the same peptide. For these reasons, it is not appropriate to rank the accuracy and reliability of the algorithms on the basis of a comparison of the *r* and *p* values as determined when they are applied to the same set of experimental results (Fig 3). Nevertheless, we believe that the general comparative strategy is useful for this review, in order to provide an overview of the predictive power of *in vivo* amyloid aggregation.

The 3D-profile method is the only algorithm that produces predictions that correlate poorly with all four experimental data sets. If we therefore exclude the 3D-profile method from the comparison, both empirical and structure-based algorithms perform similarly in predicting amyloid aggregation *in vivo*, confirming the validity of both approaches (mean *r* values of 0.631 ± 0.107 and 0.527 ± 0.136, respectively; median values of 0.622 and 0.522, respectively).

It is worth noting that algorithms that mark a neat distinction between aggregation-promoting regions and non-influential regions are not more accurate than those that do not consider the position of the mutated residue in the sequence. Indeed, the experimental data reviewed here show that mutations have an effect on aggregation propensity *in vivo* at many positions in the sequence. Taken together, this implies that the position of the mutation is not an important factor in amyloid formation *in vivo*. This might reflect the ability of chaperones to interact with and suppress the aggregation potential of hot spots, forcing other regions to emerge in promoting the process. It might also reflect the ability of the many biological macromolecules described so far to affect and nucleate amyloid fibril formation by promoting heterogeneous mechanisms of amyloid formation, on the basis of interactions that change with the biological factors involved and do not necessarily involve hot spots. Further investigation is needed to clarify this point.

Algorithms can predict disease manifestation

So far we have described the use of algorithms to predict amyloid formation *in vitro* and in the cytosol of *E. coli*—a simple organism in which the solubility or aggregation of an expressed peptide or protein can be measured quantitatively. Now, the question is whether this reliability and accuracy can be extended to predicting amyloid fibril formation in superior organisms. In the context of medicine, can the algorithms be used to gain an understanding of the pathogenesis of protein deposition diseases in humans, for example to predict disease severity from the mutations associated with familial forms of a disease?

The Zyggregator algorithm—one of the 14 algorithms listed in Table 1 and applied to the *E. coli* data in Figs 1, 2, S1 and S2—has been used to rationalize phenotypic effects of the aggregation of $A\beta_{42}$ in the central nervous system of flies (Luheshi *et al*, 2007). In a group of *Drosophila melanogaster* strains overexpressing variants of $A\beta_{42}$, a strong negative correlation was found between the longevity and locomotor ability of the flies and the predicted propensity of the variants to form amyloid fibrils, as determined by using Zyggregator (Fig 4A,B). This result is remarkable for two reasons: first, it demonstrates the power of a computational method based on physicochemical factors to predict amyloid formation in an animal model; second, the prediction extends to the phenotypic effect of amyloid formation, thereby extending the predictive power beyond protein aggregation *per se*.

In the light of increasing understanding that protein oligomers or protofibrils (as opposed to fibrils) are the pathogenic species

Sidebar A | In need of answers

(i) Can we dissect an algorithm into various forms that predict the different steps of aggregation (including rate of oligomerization, length of the lag phase for fibril formation, rate of fibril elongation, hot spots in oligomers and hot spots in fibrils)?
(ii) Can we improve the existing algorithms with additional physicochemical and biological factors?
(iii) How can we improve our ability to quantitatively measure protein aggregation and its various phases *in vivo*?
(iv) Will we be able to discriminate between the factors affecting protein aggregation and its various phases *in vitro* or *in vivo*? Could we weight their contributions appropriately in both cases?

in neurodegenerative diseases associated with amyloid deposition, the authors of the same study used a modified version of the Zyggregator algorithm to predict the effect of mutations on protofibril, rather than fibril, formation. The modified algorithm predicted the propensity of $A\beta_{42}$ mutants to form protofibrils that better correlated with the relative longevity and locomotor ability of the flies expressing the same mutants (Fig 4C,D), thus validating the power of algorithms adjusted to predict oligomer formation in medicine.

In another study, disease duration in patients with familial amyotrophic lateral sclerosis (fALS), which is associated with mutations of superoxide dismutase 1 (SOD1), was found to negatively correlate with the increase in the rate of formation of β-sheet aggregates by SOD1 after mutation, as determined by using the Chiti and Dobson equation (Fig 4E; Wang, Q. *et al*, 2008). Dissimilarly to $A\beta_{42}$ and truncated HypF-N, native SOD1 is a dimeric, folded protein containing secondary, tertiary and even quaternary structure. In a globular protein of this type, amyloid formation is known to require partial unfolding of the native structure and the amyloidogenic effect of a mutation is known to depend on the destabilization of the native structure caused by the mutation (Chiti & Dobson, 2006). Having observed that the disease duration of fALS also negatively correlated with the instability of the SOD1 variants, the authors introduced a combined function to their algorithm that considered both the intrinsic propensity to form β-sheet aggregates of unfolded SOD1 (as determined using the equation from Chiti & Dobson) and the instability of native SOD1 (Wang, Q. *et al*, 2008). Disease duration was found to better correlate with such a combined function (Fig 4F).

We have run all the algorithms on the $A\beta_{42}$ variants in the *Drosophila melanogaster* study (Luheshi *et al*, 2007) and on all SOD1 variants studied in the fALS study (Wang, Q. *et al* 2008). We found that some algorithms provide good predictions, whereas others fail to provide predictions of relative longevity and disease duration. As explained, the lower predictive power of the algorithms, in these cases relative to the *E. coli* studies, depends on the fact that phenotypic effects and clinical observations depend not only on amyloid formation, but also on other factors. The better performance of the modified algorithms used in the *D. melanogaster* and fALS studies to consider the effect of mutations on protofibril formation—rather than fibril formation—and protein instability, respectively, provide examples that this is the case.

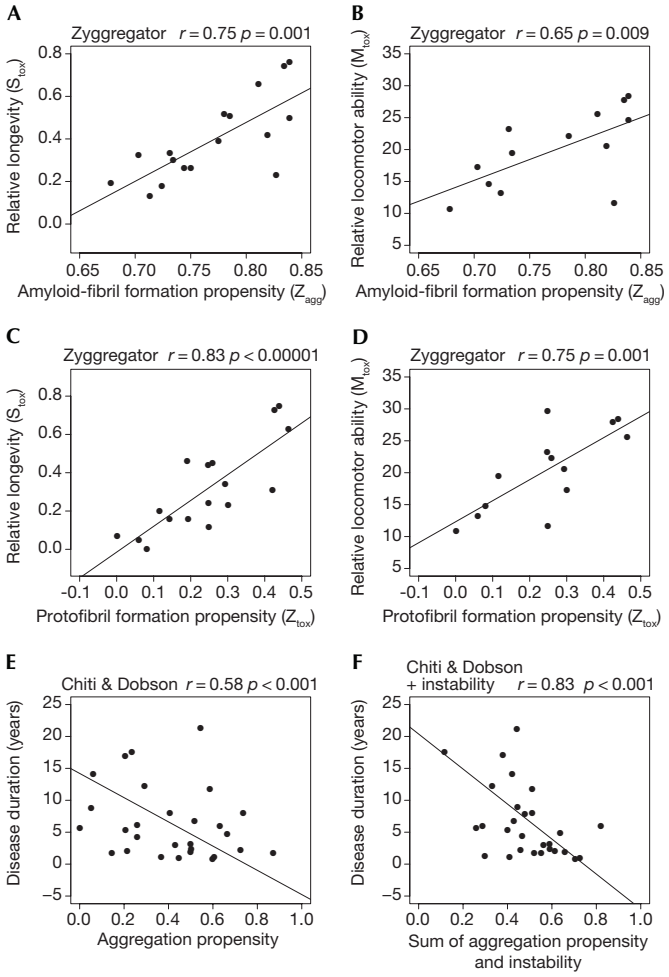


Fig 4 | Correlation between computational predictions and disease manifestations in higher organisms. (**A–D**) Correlation between relative longevity (S_{tox}) and relative locomotor ability (M_{tox}) of *Drosophila melanogaster* strains overexpressing variants of $A\beta_{42}$, and propensity of the overexpressed variants to form amyloid fibrils (**A**,**B**) or protofibrils (**C**,**D**). Following the original definition, high values of S_{tox} and M_{tox} correspond to low values of longevity and locomotor ability, respectively. Adapted with permission from Luheshi *et al*, 2007. (**E**,**F**) Correlation between disease duration of patients with fALS and aggregation propensity (**E**) or the sum of aggregation propensity and instability (**F**) of the variants of SOD1. Adapted with permission from Wang, Q. *et al*, 2008. The algorithms used for the calculation are indicated in each graph. The *r* and *p* values reported for each graph are those indicated by the original authors. SOD1, superoxide dismutase 1.

A database to collect amyloid-formation data

To facilitate the development and testing of new algorithms (or improved versions of existing ones), it seems wise to create a database to collect experimentally determined kinetic data about oligomer or amyloid fibril formation. This database could then be used to analyse and compare data from different labs, involving different proteins, mutants and solution conditions. To this end, we have prepared a supervised, interactive web server for the deposition of kinetic data of oligomer or amyloid fibril formation and related experimental conditions, which we offer as a resource to the community. We have already deposited the data used in this review, and we encourage our peers to enrich the database with their own results at www.unifi.it/scibio/bioinfo/AmyloBase.html.

Conclusions and future perspectives

The positive message from this review is that algorithms that have been derived from and used to predict amyloid fibril formation in the test tube—in which a protein is left to aggregate in the absence of biological factors—also offer a considerable degree of accuracy for predicting amyloid-aggregation propensity *in vivo*. Much remains to be done to improve our ability to rationalize and predict amyloid formation in biological contexts, as well as to extend this prediction to disease manifestation and pathology (Sidebar A). In this regard, it is important to increase the accuracy of existing algorithms, for example by identifying additional physicochemical factors that have a role in amyloid fibril formation and modifying the algorithms appropriately, and implementing them with additional biology-related factors. It will also be important to overcome limitations in the experimental measurement of amyloid-formation parameters in living organisms (for example, hot spots, rates of formation and stability of fibrils, oligomers and so on). However, the results reviewed and analysed here suggest that we have the instruments to elucidate and predict amyloid fibril formation *in vivo* and the phenotypic and clinical effects associated with it.

Supplementary information is available at EMBO *reports* online (http://www.emboreports.org).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REEFERENCES

Bejarano E, Cuervo AM (2010) Chaperone-mediated autophagy. *Proc Am Thorac Soc* **7:** 29–39

Bryan AW, Menke M, Cowen LJ, Lindquist SL, Berger B (2009) BETASCAN: probable β-amyloids identified by pairwise probabilistic analysis. *PLoS Comput Biol* **5:** e1000333

Carrió M, González-Montalbán N, Vera A, Villaverde A, Ventura S (2005) Amyloid-like properties of bacterial inclusion bodies. *J Mol Biol* **347:** 1025–1037

Cherny I, Gazit E (2008) Amyloids: not only pathological agents but also ordered nanomaterials. *Angew Chem Int Ed Engl* **47:** 4062–4069

Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem* **75:** 333–366

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

Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* **8:** 65

de Groot NS, Aviles FX, Vendrell J, Ventura S (2006) Mutagenesis of the central hydrophobic cluster in Aβ42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. *FEBS J* **273:** 658–668

DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M (2004) Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J Mol Biol* **341:** 1317–1326

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

Galzitskaya OV, Garbuzynskiy SO, Lobanov MY (2006) Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput Biol* **2**: e177

Hoseki J, Ushioda R, Nagata K (2009) Mechanism and components of endoplasmic reticulum-associated degradation. *J Biochem* **147:** 19–25

Jahn TR, Radford SE (2008) Folding versus aggregation: polypeptide conformations on competing pathways. *Arch Biochem Biophys* **469:** 100–117

Kapoor A, Sanyal AJ (2009) Endoplasmic reticulum stress and the unfolded protein response. *Clin Liver Dis* **13:** 581–590

Kim W, Hecht MH (2006) Generic hydrophobic residues are sufficient to promote aggregation of the Alzheimer's Aβ42 peptide. *Proc Natl Acad Sci USA* **103:** 15824–15829

Luheshi LM *et al* (2007) Systematic *in vivo* analysis of the intrinsic determinants of amyloid β pathogenicity. *PLoS Biol* **5**: e290

Maurer-Stroh S *et al* (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat Methods* **7:** 237–242

Monsellier E, Chiti F (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep* **8:** 737–742

Monsellier E, Ramazzotti M, Taddei N, Chiti F (2008) Aggregation propensity of the human proteome. *PLoS Comput Biol* **4:** e1000199

Pawar AP, Dubay KF, Zurdo J, Chiti F, Vendruscolo M, Dobson CM (2005) Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *J Mol Biol* **350:** 379–392

Rousseau F, Serrano L, Schymkowitz JW (2006) How evolutionary pressure against protein aggregation shaped chaperone specificity. *J Mol Biol* **355:** 1037–1047

Sekijima Y, Wiseman RL, Matteson J, Hammarström P, Miller SR, Sawkar AR, Balch WE, Kelly JW (2005) The biological and chemical basis for tissue-selective amyloid disease. *Cell* **121:** 73–85

Tartaglia GG, Vendruscolo M (2008) The Zyggregator method for predicting protein aggregation propensities. *Chem Soc Rev* **37:** 1395–1401

Tartaglia GG, Cavalli A, Pellarin R, Caflisch A (2004) The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci* **13:** 1939–1941

Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, Eisenberg D (2006) The 3D profile method for identifying fibril-forming segments of proteins. *Proc Natl Acad Sci USA* **103:** 4074–4078

Trovato A, Chiti F, Maritan A, Seno F (2006) Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLoS Comput Biol* **2:** e170

Ventura S, Villaverde A (2006) Protein quality in bacterial inclusion bodies. *Trends Biotechnol* **24:** 179–185

Voellmy R, Boellmann F (2007) Chaperone regulation of the heat shock protein response. *Adv Exp Med Biol* **594:** 89–99

Wang L, Maji SK, Sawaya MR, Eisenberg D, Riek R (2008) Bacterial inclusion bodies contain amyloid-like structure. *PLoS Biol* **6:** e195

Wang Q, Johnson JL, Agar NY, Agar JN (2008) Protein aggregation and protein instability govern familial amyotrophic lateral sclerosis patient survival. *PLoS Biol* **6:** e170

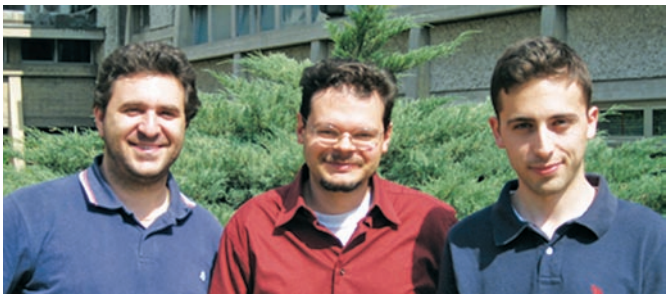
Winkelmann J, Calloni G, Campioni S, Mannini B, Taddei N, Chiti F (2010) Low-level expression of a folding-incompetent protein in *Escherichia coli*: search for the molecular determinants of protein aggregation *in vivo*. *J Mol Biol* **398:** 600–613

Wurth C, Guimard NK, Hecht MH (2002) Mutations that reduce aggregation of the Alzheimer's Aβ42 peptide: an unbiased search for the sequence determinants of Aβ amyloidogenesis. *J Mol Biol* **319:** 1279–1290

Yoon S, Welsh WJ (2004) Detecting hidden sequence propensity for amyloid fibril formation. *Protein Sci* **13:** 2149–2160

Yoon S, Welsh WJ, Jung H, Yoo YD (2007) CSSP2: an improved method for predicting contact-dependent secondary structure propensity. *Comput Biol Chem* **31:** 373–377

Zibaee S, Makin OS, Goedert M, Serpell LC (2007) A simple algorithm locates β-strands in the amyloid fibril core of α-synuclein, Aβ, and tau using the amino acid sequence alone. *Protein Sci* **16:** 906–918



Matteo Ramazzotti (left), Fabrizio Chiti & Mattia Belli (right)