

# Multiple Evolutionary Mechanisms Reduce Protein Aggregation

Joke Reumers<sup>1,2</sup>, Frederic Rousseau<sup>\*,1,2</sup> and Joost Schymkowitz<sup>\*,1,2</sup>

<sup>1</sup>VIB Switch Laboratory, Brussels, Belgium

<sup>2</sup>Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

**Abstract:** The folding of polypeptides into stable globular protein structures requires protein sequences with a relatively high hydrophobicity and secondary structure propensity. These biophysical properties, however, also favor protein aggregation via the formation of intermolecular beta-sheets and, as a result, globular structure and aggregation are inextricable properties of protein polypeptides. Aggregates that are enriched in beta-sheet structures have been found in diseased tissues in association with at least twenty different human disorders and the effect of aggregation on protein function include simple loss-of-function but also often a gain of toxicity. Given both the ubiquity and the potentially lethal consequences of protein aggregation, negative selective pressure strongly minimizes aggregation. Various evolutionary strategies keep aggregation in check, including (1) the optimisation of the thermodynamic stability of the protein, which precludes aggregation by burial of the aggregation prone regions in solvent inaccessible regions of the structure, (2) segregation between folding nuclei and aggregation nuclei within a protein sequence, (3) the placement of so-called gatekeeper residues at the flanks of aggregating segments, that reduce the aggregation rate of (partially) unfolded proteins, and (4) molecular chaperones that target aggregation nucleating sequences directly, thereby further suppressing aggregation in a cellular environment. In this review we describe the intrinsic features built into protein sequence and structure that protect against aggregation.

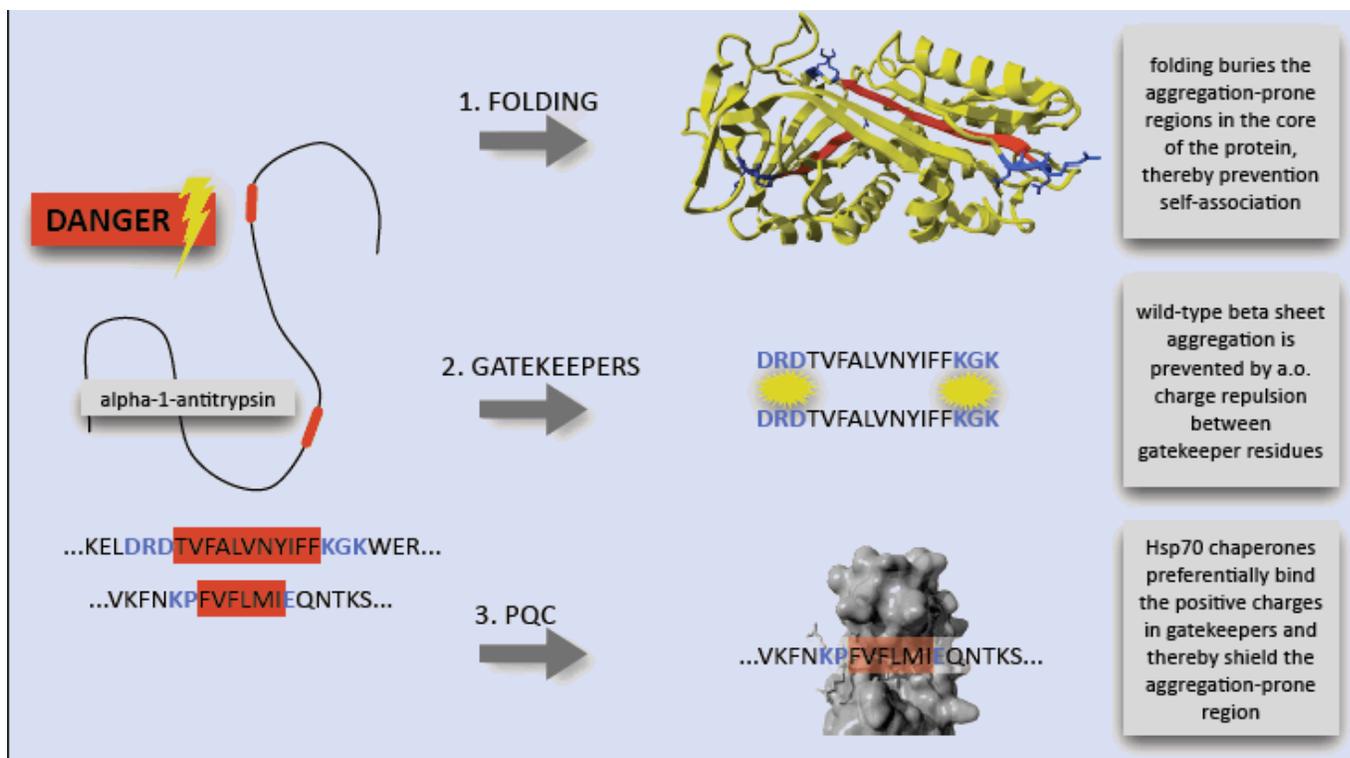
**Keywords:** Amyloid, protein aggregation, protein aggregation and evolution, protein evolution, protein folding and aggregation, evolutionary pressure.

## INTRODUCTION

Misfolding and the associated aggregation of proteins have been the object of intensive study in the last decade, as they appear to be the molecular basis of neurodegenerative disorders such as Alzheimer's and Parkinson's disease, and other diseases such as type 2 diabetes [1]. To date, circa 40 disorders have been linked to protein aggregation [2]. Aggregation is unavoidable in globular proteins, because non-native conformations can be adopted during or immediately after synthesis, under stress conditions or as a consequence of mutations or proteolysis. Although it seems that almost all proteins are able to form aggregates when expressed at high concentrations *in vitro*, they differ substantially in their intrinsic propensity to do so under physiological conditions [3]. Most importantly, aggregation is nucleated by short sequence segments with specific physical properties and the amino acid residues involved in aggregation are usually segregated in the primary structure from the residues that are critical for proper folding [4]. The major contributors to aggregation propensity have been identified as hydrophobicity, net charge and propensity to form secondary structure, i.e. a predisposition for beta-sheet formation and an aversion for alpha-helical structures [4]. The identification of these determinants of aggregation facilitated the development of prediction algorithms that assess the effect of mutations on aggregation, identify the regions in the protein sequence that promote aggregation, and quantify the aggrega-

tion rates of unfolded proteins [5-9]. These computational methods have enabled the large-scale analyses of the aggregation behavior of full proteomes [10-12], which have confirmed the ubiquity of aggregation propensity in proteomes of all kingdoms of life. Protein aggregation represents an enormous burden for cellular organisms: not only the loss-of function of the individual aggregating proteins imposes stress on the cell, but also the energy consumed by the ATP-dependent protection mechanisms of the protein quality control machinery. Hence, proteomes are subject to strong evolutionary pressure to minimize aggregation [13]. The different mechanisms hindering effective protein aggregation in the cell are illustrated in Fig. (1) with the example of alpha-1-antitrypsin (A1AT). The deficiency of antitrypsin has been associated with aggregation of this enzyme and results in liver dysfunction [14]. The enzyme has two predicted aggregation-prone regions, which are buried in the correctly folded form of the protein (i). These two regions are flanked by so-called gatekeeper residues, which in the unfolded state of the protein will prevent self-association through charge repulsion or steric hindrance (ii). In addition to these two mechanisms embedded in the protein's sequence and structure, the cell has developed a highly advanced protein quality control system (iii) [15]. A large variety of chaperones in the cell hinder the formation of aggregates, not only by shielding the aggregation-nucleating regions in the nascent chain, but also by sequestering unfolded proteins from other identical proteins, and by untangling partial aggregates [16]. In this review we will focus on the intrinsic protein characteristics that counteract aggregation.

\*Address correspondence to this author at the Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium; Tel: +321 6346227; Fax: 321 6347181; E-mail: frederic.rousseau@switch.vib-vub.be or joost.schymkowitz@switch.vib-vub.be



**Fig. (1).** The different strategies used to oppose the formation of protein aggregates. The structure and sequence shown is alpha-1-antitrypsin (AAT), of which the deficiency, caused by aggregation, is associated with liver disease [14]. 1) Folding buries the sticky regions in the core of the protein. 2) Well-placed gatekeeper residues prevent self-association and therefore inhibit the aggregation process. 3) The protein quality control system has evolved to oppose and invert aggregation. For example members of the Hsp70- family recognize the positive charged residues at the flanks of aggregation-nucleating regions. In the case of secreted AAT, quality control is performed in the ER by the Hsp70 family member BiP [14].

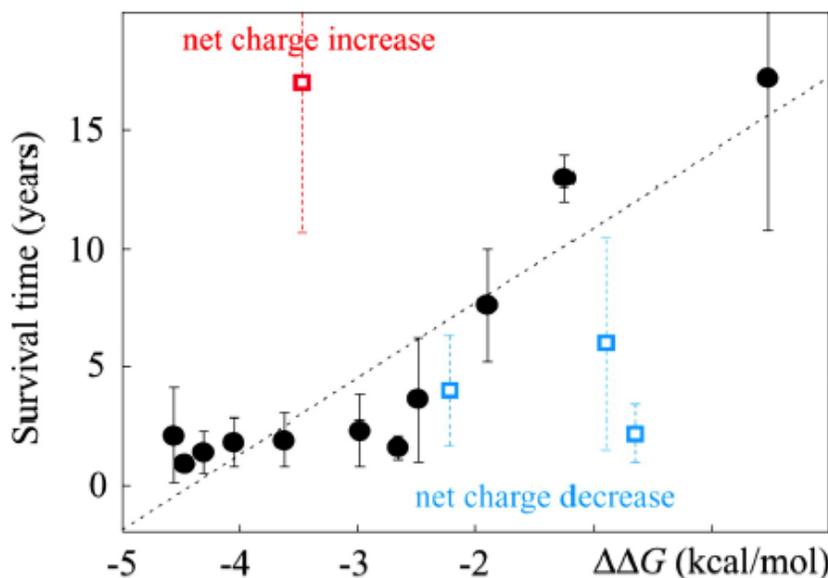
### AGGREGATION-PRONE SEQUENCES ARE BURIED INSIDE PROTEIN STRUCTURES

Protein folding and aggregation are competing conformational reactions. As a result, the first defense mechanism against aggregation is the stability of the native protein conformation itself: in a folded protein the backbone is locked in the tertiary structure of the protein and therefore not accessible to form the inter-chain hydrogen bonds that are a determining factor in cross-beta aggregated structures [17]. Cooperativity in folding is related with resistance against aggregation [18], and studies of the folding of a computationally designed protein suggest that the smooth folding pathways of small polypeptides are the result of negative selection against aggregation, and not a general property of proteins that fold into a unique stable structure [19]. Although it has been discovered that globular native structures can also form aggregates through intermolecular  $\beta$ -strand interactions at edges of individual  $\beta$ -sheets [20] or three dimensional domain swapping [21-23], it is still universally accepted that unfolded or partially unfolded proteins generally have a higher propensity to aggregate than the fully native states [24].

In a large-scale study using experimentally determined stability measurements of 2351 mutations in globular proteins, Serrano and colleagues showed that stability is the main evolutionary pressure in the absence of other factors

such as binding and catalysis [25]. Their analysis revealed that misfolding is avoided primarily by selection for stability, and also that avoiding misfolding-prone sequences compromises stability, emphasizing the inextricable tie between protein structure and aggregation. However, the maintenance of an aggregating segment within a sequence does not have negative consequences if the aggregation load is not too high. There exists a "permissive" window for aggregation: highly aggregating sequences are prevented but moderately aggregating ones are tolerated [26]. This is confirmed in proteome-wide studies of aggregation propensities where the majority of the proteins have low predicted aggregation scores, and only a small portion have very high tendency to aggregate [10-12].

The relation between the tolerance for aggregation-prone regions and the burial of these regions in folded proteins is underlined by the differences in aggregation propensities between globular and intrinsically disordered proteins (IDPs). Proteome-wide studies of aggregation propensities showed that globular proteins from all-alpha, all-beta and mixed alpha/beta SCOP classes showed similar levels of aggregation propensity, while natively unstructured proteins show much lower average aggregation loads [10, 27]. In a Monte Carlo simulation of small hydrophobic peptides with and without disordered flanks, Abeln & Frenkel showed that disordered flanks next to aggregating regions even prevent aggregation [28]. Small hydrophobic peptides without disor-



**Fig. (2). Interplay between stability and net charge determines the age of onset in familial ALS.** The average survival time after diagnosis is plotted in function of the protein stability changes ( $\Delta\Delta G$ ) in SOD1.  $\Delta\Delta G$  for ALS-associated mutations that do not alter the net charge of SOD shows a high correlation with survival time ( $R = 0.91$ ). Increasing the net charge of the protein causes a shift toward longer survival time, whereas decreasing the charge has the opposite effect. Reprinted with permission from [32].

dered flanks aggregated, while the peptides with unstructured flanks were stable as monomers or small micelle-like clusters. The disordered flanks have no effect on the native function of the motif, i.e. binding energy is not affected.

#### POINT MUTATIONS CAN MODULATE PROTEIN STABILITY AND AGGREGATION TENDENCY

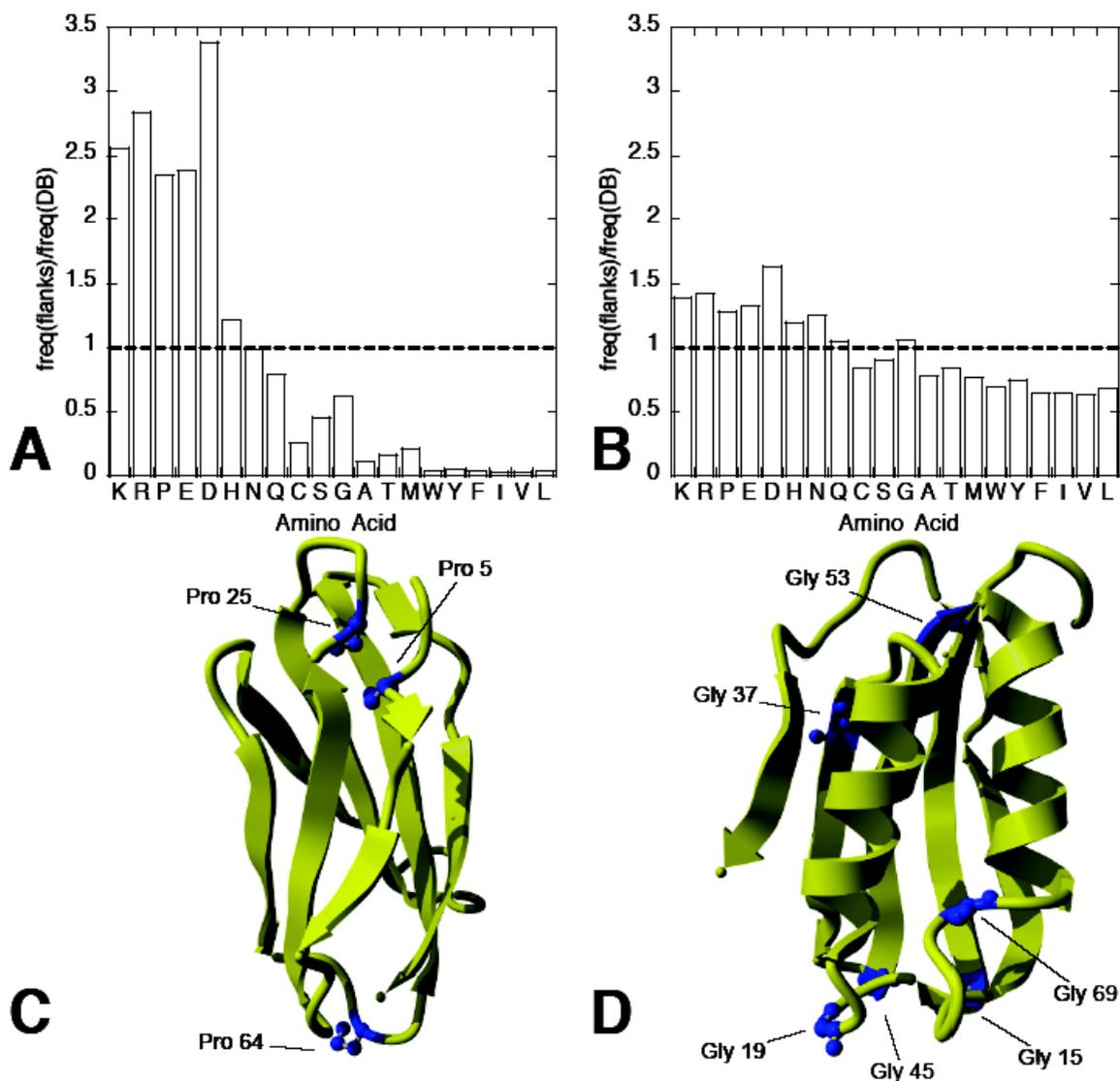
The amyloidogenicity of a protein can be reduced by stabilization of the native structure (reviewed in [29]); conversely, many mutations associated with increased aggregation have been shown to destabilize the native structure. This has been shown experimentally for several (disease related) proteins, such as transthyretin in amyloidosis [30] and Cu/Zn-superoxide-dismutase (SOD1) for amyotrophic lateral sclerosis (ALS) [31]. In the latter case, Oliveberg and co-workers could link experimentally determined stability differences of apo-SOD1 to survival time of ALS patients [32]. The stability change for ALS-associated mutations that do not alter the net charge of SOD shows a high correlation with survival time (Fig. 2). An additional large scale study showed that the combination of increased aggregation propensity and decreased protein stability can account for 69% of the variability in familial ALS patient survival times [31]. The link between increased destabilization and more aggressive disease development can also be found among transthyretin mutations in amyloidosis [30], where the rate of tetramer dissociation needed for amyloid formation influences both disease penetrance and age of onset.

Chiti and coworkers examined the interplay between decreased stability and increased aggregation *in vivo*: the solubility of mutations in the N-terminal domain of *Escherichia coli* HypF protein in the cell was compared with their effect on stability of the protein. HypF-N has been shown to convert to amyloid fibrils *in vitro* that are morphologically similar to those found in amyloid disease

[33]. HypF-N variants carrying destabilizing mutations aggregate after expression, whereas mutants with stability similar to the wild type protein remain soluble in the *E. coli* cytosol [34]. Although these studies show that destabilisation is the major factor contributing to misfolding, it is certainly not the only factor. Destabilisation does not always imply misfolding and vice versa, as demonstrated by mutations that affect aggregation independent of stability [35, 36].

#### GATEKEEPER RESIDUES DISRUPT STRETCHES OF HYDROPHOBIC RESIDUES TO MINIMIZE AGGREGATION PROPENSITY

The term *structural gatekeeper* was first introduced in the context of the two-state folding pathway of protein S6 [37], as residues that steer the folding process by blocking certain paths. It was later introduced in the context of the A $\beta$  amyloid peptide aggregation by the same researchers as "charged side chains that prevent aggregation by interrupting contiguous stretches of hydrophobic residues in the primary sequence" [38]. A computational analysis of the aggregation properties of 26 proteomes by Rousseau and co-workers [12] with the TANGO algorithm [6] revealed a strong enrichment of charged residues (arginine, lysine, aspartate and glutamate) and proline at the flanks of aggregation-prone regions. Their study showed that 90% of aggregation-prone regions are capped with at least one gatekeeper residue, with a bias for positively charged residues at regions with the highest aggregation propensities. A similar result was obtained by Chiti and co-workers in the analysis of the human proteome [10] with a different computational method [7]. In accordance with the aforementioned study by Rousseau *et al.*, they found that Arg, Lys and Pro had higher frequencies at the flanks of regions with high aggregation propensity. In a follow-up study of the human proteome, Rousseau *et al.* investigated the composition of the three



**Fig. (3). Disruption of aggregation motifs by polar residues and structure breakers.** A+B. Enrichment of gatekeeper residues at the flanks of aggregating regions. The ratio of amino acid frequency in the flanks versus the frequency of amino acids in the full data set is shown for each gatekeeper type, considering 1 position (A) or three positions (B) before and after each aggregation-nucleating region. All gatekeepers (P, R, K, D, E) are enriched in the flanks (ratio >1). The pattern is very distinct at the first position before and after the regions, but the broader flanks also show this enrichment. When taking into account three positions (B) we also see an enrichment of Histidine (H) and Asparagine (N), and to a lesser extent glycine (G) and glutamine (Q). Adapted from [11]. C+D. Opposition of aggregation by conserved structure breakers. C. Aggregation of fibronectin type III domains is limited by conserved proline residues. Adapted from [40]. D. Conserved glycines in human muscle acylphosphatase slow down the formation of aggregates. Adapted from [41].

amino acid positions before and after aggregation prone regions [11]. Due to the long-range effects of electrostatic interactions, the boundaries of aggregation nucleating zones may not be strictly defined. The elevated usage of the 5 previously identified gatekeepers (P, R, K, D, E) on the direct flanks of aggregation-prone regions (Fig. 3A) was confirmed in the three C-terminal and three N-terminal flanking positions (Fig. 3B). The enrichment was most prominent for the charged residues and less pronounced for proline. Another feature of gatekeeper motifs that was highlighted in this study is the use of multiple gatekeepers: nearly 75% of all aggregation nucleating regions in the human proteome uses two or more gatekeepers. The type of gatekeeper

used varies between single and multiple gatekeeper motifs: when using one single gatekeeper residue, proline is used most often, but its usage decreases with the introduction of more gatekeepers. Using multiple gatekeepers may be a protection mechanism against mutation: redundancy in the gatekeeper motif reduces the risk of initiating aggregation by a single point mutation.

#### GLOBAL NET CHARGE AND STERIC HINDRANCE PROTECT AGAINST AGGREGATION

In addition to safeguarding the flanks of aggregation nuclei, charged residues and structure breakers such as

glycine and proline provide protection on the overall protein sequence. For instance, the use of multiple structure breakers to oppose aggregation was also found in Huntingtin, from which the aggregation is associated to Huntington's disease, where the polyglutamine stretch is flanked by a proline-rich region that keeps aggregation in check [39]. Other studies show examples of prolines [40] and glycines [41] that were evolutionary conserved to modulate aggregation. In the investigation of three highly conserved prolines in fibronectin type III domains (Fig. 3C), no obvious structural or functional role could be appointed to these conserved residues. The stability of alanine mutations of these three prolines in the 10<sup>th</sup> domain of human fibronectin was similar to that of the wild-type domain, but the aggregation rate of the mutant proteins was significantly higher than that of the original domain [40]. Analogous results were obtained in the study of conserved glycine residues in human muscle acylphosphatase (AcP, Fig. 3D) [41]: mutating these glycines to alanine does not affect stability more than mutating non-conserved positions, but it does accelerate amyloid formation of AcP. Furthermore, an earlier extensive mutation study of the same enzyme already demonstrated that the aggregation behaviour of AcP could be modified by the mutation of single amino acids. More specifically, the authors showed an inverse correlation between the net charge of the protein and its aggregation rate [42]. This anti-correlation between charge and aggregation is also illustrated in amyotrophic lateral sclerosis, where the majority of disease-related SOD1 mutations reduce the net charge of the protein [43]. Oliveberg *et al.* performed a computational analysis of 100 ALS-associated mutations in SOD1 and showed that in comparison with other well-described disease related genes and mutations, the charge bias of SOD1 is significantly higher. In a similar comparison of SOD1 with all disease-related proteins in the SwissProt database with more than 50 causal mutations, the average charge difference of SOD1 mutations was ranked second. The generality of these protection mechanisms has been shown in mutational studies of several proteins, where the introduction of charged residues, proline and glycine resulted in reduced aggregation kinetics or compromised stability of the formed aggregates [34, 42, 44, 45]. The protection against aggregation provided by these charged residues and structure breakers is also employed in intrinsically disordered proteins, where a higher proline content [46] and higher net charge [27, 47] contributes to lower aggregation.

In addition to the involvement in the pathology of misfolding disorders, protein aggregation also poses a problem *in vitro*, in biotechnology and biomedical research. Building further on the aforementioned observations of opposing aggregation with charges, Liu and colleagues set out to design supercharged versions of naturally occurring proteins [48]. By replacing solvent-exposed residues of a monomeric (green fluorescent protein, GFP), a dimeric (glutathione-S-transferase, GST) and a tetrameric protein (streptavidin, SAV) with charged amino acids, they demonstrated that by supercharging proteins it is feasible to obtain correctly folded variants of the natural protein. The design of GST and SAV mutants was performed with an automated mutagenesis strategy: residues were ranked by increasing solvent exposure calculated from the crystallographic structure, and then the highest ranked residues were replaced by

lysine for positive supercharging, or by glutamate or aspartate for negative supercharging. These supercharged variants displayed their native functionality *in vitro* but also remained soluble in conditions that normally cause the proteins to aggregate. This approach might solve some of the unwanted behavior of *de novo* designed proteins, and may contribute to the adaptation of natural proteins to thrive in non-natural conditions, such as increased temperature or the presence of denaturing chemical additives [49]. These combined results suggest a strong evolutionary pressure on the flanks of aggregation-prone regions, and confirms the use of structural gatekeepers as a universal mechanism against aggregation.

## CHAPERONE BINDING IS MODULATED BY GATEKEEPER RESIDUES

However different in mechanism, most chaperones display a remarkable resemblance in substrate specificity and prefer binding to hydrophobic stretches flanked by positive charges. This has been shown by affinity studies for Hsp70 [50, 51], Hsp90 [52], and many other chaperones [53-55]. Although Hsp60 substrate specificity studies on GroEL have not revealed such clear charge preferences, it is suggested that proteins with negative charges fold rapidly by repulsion forces in the negatively charged cage [56]. Application of the substrate specificity preferences for DnaK and trigger factor developed by Bukau and co-workers [50, 54] on the aggregating segments of the *Escherichia coli* proteome showed that together these two chaperones target almost 100% of the strongly aggregating sequences in *E. coli* [12]. This suggests that chaperones recognize aggregation-prone regions by the double criterion of having a hydrophobic stretch flanked by (mostly positive) charges. The high prevalence of these motifs in proteomes in the various kingdoms of life suggest that the evolutionary pressure on proteomes to counteract aggregation with charges and structure breakers also shaped the specificity for chaperones to recognize these patterns. These findings are in accordance with the observation that intrinsically disordered proteins, which have significantly lower aggregation propensities than globular proteins [27], also bind less to chaperones [57].

## GATEKEEPER MUTATIONS CONTRIBUTE TO HUMAN DISEASE

Although gatekeeper residues appear to be very effective in guarding stretches of aggregation-prone residues, they also imply a risk for disease development. As seen in several examples, such as mutations of tau [58], the Alzheimer beta-peptide [59] and  $\alpha$ -synuclein [60], mutating a single amino acid can substantially change aggregation propensity and can have dramatic effects on disease etiology. The TANGO algorithm was used to study difference in aggregation caused by known human disease mutations and neutral single nucleotide polymorphisms (SNPs) from the UniProt database [11]. The two main observations of this analysis were that i) the distribution of differences in the TANGO aggregation scores for disease mutations showed more extreme differences and a smaller fraction of neutral changes than the distribution for the neutral SNPs; and ii) the fraction of disease mutations that cause a significant increase of protein

aggregation due to the disruption of a gatekeeper motif was almost twice as large as the fraction of these mutations found among SNPs (3.5% of the disease mutations *versus* 1.9% of the SNPs). These findings suggest that indeed gatekeeper residues are crucial for correct protein function and that disruption of the gatekeeper pattern introduces a risk of disease.

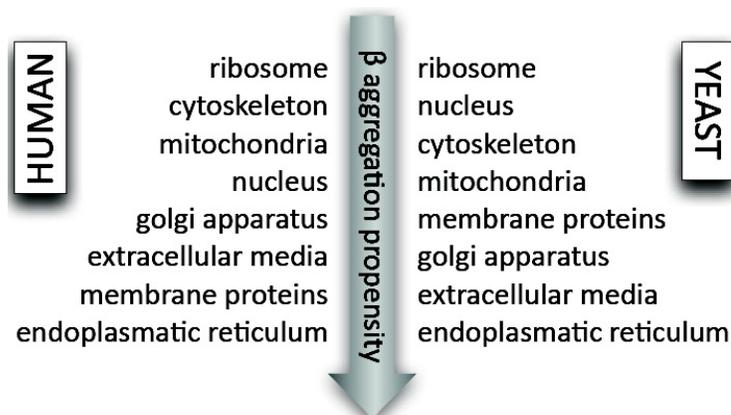
### NEGATIVE SELECTION AGAINST UNWANTED SELF-ASSOCIATION IN CORRECTLY FOLDED PROTEINS

There are some examples where proteins with native structure, more specifically  $\beta$ -sheet rich proteins, can form aggregates through intermolecular interactions of peripheral  $\beta$ -strands [20]. These external  $\beta$ -strands propose a risk for possible self-interaction and thus  $\beta$ -aggregation, but the placement of  $\beta$ -bulges, superposition of short loops, helices or distorted  $\beta$ -strands on the peripheral strand, and other ways to distort the  $\beta$ -structure are used to avoid inducing aggregation of peripheral strands with those of other molecules [20]. Two types of (functional) self-interactions of identical or homologous sequences that can be found in nature are homo-oligomeric complex formation and domain repeats in multidomain proteins. Especially in the former case, formation of functional homo-oligomers and non-functional aggregates are competing processes, as has been shown in studies of the C-Src SH3 domain [61]. Using protein-protein interaction data of fly, yeast and worm, Chen & Dokholyan showed proteins that have native self-interactions patterns (such as homo-oligomeric complex formation) have overall lower aggregation scores than proteins without these patterns [62], suggesting negative selection for aggregation in these proteins. Dobson and co-workers investigated the multidomain constructs of immunoglobulin domains in human cardiac titin and the ability of these homologous domains to co-aggregate [63]. Their conclusion was that the efficiency of co-aggregation lowers with decreasing sequence identity, with a lower bound at 30-40% sequence identity. Further computational analysis of homologous

domains in large multidomain proteins (i.e. the immunoglobulin and fibronectin type III superfamilies) showed that the sequence identity between repeats remains largely below this threshold. Comparison of the sequence identity between adjacent and non-adjacent domain pairs also revealed that there is a higher evolutionary pressure on adjacent domains: sequence identity between adjacent pairs is significantly lower.

### DETAILED FEATURE ANALYSIS OF AGGREGATING PROTEINS REVEALS ADDITIONAL SELECTION AGAINST AGGREGATION

Besides the apparent evolutionary pressure related to folding, gatekeeper patterns, chaperone binding and native self-interaction, various studies have revealed additional evidence for selection against aggregation-prone segments. Simple patterns that favor aggregation, such as the alternation of polar and non-polar stretches are rare in natural proteins [64]. This is a first example that not only the amino acid composition itself is a determinant for aggregation propensity, but also the order of the residues. Further proof was provided in a detailed study using horse heart apomyoglobin (apoMb) [26]. The core of the amyloid fibrils formed by apoMb is the region spanning from residue 7 to 18. Independent from the full length protein, the N-terminal region of apoMb (residue 1-29) is soluble at neutral pH but self-assembles into fibrils at pH 2. Keeping the same amino acid composition and length, four scrambled versions of the N-terminus were designed and their aggregation properties were investigated. The naturally occurring sequence is at the lower boundary of aggregation. Comparing the aggregation profile of the scrambled sequence with that of 745 peptides from the globin family homologous to the apoMb N-terminus showed that the former had significantly higher aggregation tendencies than their natural counterparts, confirming that the prevention of aggregation has been a driving force in protein evolution. Another piece of evidence that corroborates this evolutionary pressure was provided by investigating the aggregation propensities of essential versus



**Fig. (4).** Ranking of the aggregation propensity in different subcellular regions in human and yeast. The average aggregation propensity of proteins in different subcellular locations in *Homo sapiens* and *Saccharomyces cerevisiae* are very similar: the aggregation propensity of intracellular districts such as the nucleus and ribosome is much lower than that of secreted proteins or those located in the endoplasmatic reticulum. Adapted from [10] and [67].

non-essential proteins in *Saccharomyces cerevisiae* and *Caenorhabditis elegans* [62]. Essential genes were defined as those genes of which the knockdown led to lethality. Both in yeast and worm it was shown that essential proteins have lower aggregation propensity than non-essential ones, which is consistent with a higher evolutionary pressure on essential proteins.

### EXPRESSION LEVELS AND SUBCELLULAR LOCALIZATION ARE OTHER DETERMINANTS IN PROTECTION AGAINST AGGREGATION

As almost all proteins can be driven to aggregate when overexpressed *in vitro*, the high divergence in the expression levels of proteins in the cell is another determinant in the risk for aggregation [17]. Vendruscolo *et al* examined the *in vivo* expression levels, as measured by DNA microarray technology, of 12 human proteins of which experimentally determined aggregation rates were available. The expression levels of these human genes were anti-correlated with the aggregation rates of the corresponding proteins *in vitro* [65]. This suggests that polypeptide chains have co-evolved with their cellular environments to be soluble as far as is needed to effectively perform their functional role. These results are in accordance with previous observations that even small perturbations of expression levels can have dramatic pathological consequences in misfolding diseases [66]. In eukaryotes, not only the individual expression levels of proteins but also the overall biochemical properties in different cellular compartments can vary greatly. Studies on the differences in aggregation property between proteins of different subcellular locations in the yeast [67] and human proteome [10] agree on the observation that the aggregation propensity of secreted and ER proteins is on average higher than that of intracellular districts such as the nucleus and ribosome (Fig. 4). This evolutionary pressure against aggregation in cellular organelles is expected because on the one hand overall protein concentrations are high in these compartments, and on the other hand it has been shown that these compartments contain a large portion of unfolded molecules [68].

### CONCLUSION

Since the biophysical properties underlying the correct folding of globular proteins and the formation of protein aggregates are alike, the two processes are inescapably linked. The combination of the generality of protein aggregation propensity of globular proteins and the putative detrimental effects of protein aggregation on the cell has resulted in negative selective pressure to minimize aggregation. In this review we have described the intrinsic features of protein sequence and structure that keep aggregation in check. The main contributor to the avoidance of aggregation is correct folding: aggregation nuclei are buried within the hydrophobic core of globular proteins. However, during the lifetime of a protein (partial) unfolding cannot always be avoided. Charge repulsion and steric hindrance are used to disrupt the formation of intermolecular  $\beta$ -sheets by placing so-called *structural gatekeepers* (aspartate, glutamate, lysine, proline) at the flanks of aggregation nuclei. The use of these charged residues and structure breakers to minimize aggregation is not only observed at the flanks of aggregation

nuclei, but also global net charge and conservation of well placed prolines and glycines can limit the aggregation propensity of a protein. The charged-hydrophobic-charged pattern that characterizes the regions with high aggregation propensity flanked by gatekeepers is recognized by molecular chaperones, and optimizes chaperone binding to potentially dangerous motifs. Negative selection of aggregation-prone regions in multimeric proteins and within protein families further illustrates the evolutionary pressure against unwanted self-association. Selective pressure within the cell can vary between the different cellular compartments, related to variability in concentration, partial unfolding, and presence of chaperones in these compartments. Modulation of the protection level against aggregation in these varying situations can be achieved by combining different types of protection. This type of redundancy in protection can also serve as a fail-safe if mutation disrupts one of the protection mechanisms.

### ACKNOWLEDGEMENTS

Joke Reumers was supported by the Institute for the Encouragement of Scientific Research and Innovation of Brussels (ISRIB), Belgium.

The VIB Switch laboratory was supported by a grant from the Federal Office for Scientific Affairs, Belgium (IUAP P6/43) and the Fund for Scientific Research (FWO Vlaanderen).

### REFERENCES

- [1] Stefani M. Protein misfolding and aggregation: new examples in medicine and biology of the dark side of the protein world. *Biochim Biophys Acta* 2004; 1739(1): 5-25.
- [2] Chiti F, Dobson CM. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem* 2006; 75: 333-66.
- [3] Dobson CM. Principles of protein folding, misfolding and aggregation. *Semin Cell Dev Biol* 2004; 15(1): 3-16.
- [4] Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* 2003; 424(6950): 805-8.
- [5] Conchillo-Sole O, de Groot NS, Aviles FX, *et al*. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* 2007; 8: 65.
- [6] Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* 2004; 22(10): 1302-6.
- [7] Pawar AP, Dubay KF, Zurdo J, *et al*. Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *J Mol Biol* 2005; 350(2): 379-92.
- [8] Tartaglia GG, Cavalli A, Pellarin R, Caflisch A. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci* 2005; 14(10): 2723-34.
- [9] Trovato A, Seno F, Tosatto SC. The PASTA server for protein aggregation prediction. *Protein Eng Des Sel* 2007; 20(10): 521-3.
- [10] Monsellier E, Ramazzotti M, Taddei N, Chiti F. Aggregation propensity of the human proteome. *PLoS Comput Biol* 2008; 4(10): e1000199.
- [11] Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F. Protein sequences encode safeguards against aggregation. *Hum Mutat* 2009; 30(3): 431-7.
- [12] Rousseau F, Serrano L, Schymkowitz JW. How evolutionary pressure against protein aggregation shaped chaperone specificity. *J Mol Biol* 2006; 355(5): 1037-47.
- [13] Monsellier E, Chiti F. Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep* 2007; 8(8): 737-42.

- [14] Knaupp AS, Bottomley SP. Serpin polymerization and its role in disease—the molecular basis of alpha1-antitrypsin deficiency. *IUBMB Life* 2009; 61(1): 1-5.
- [15] Bukau B, Weissman J, Horwich A. Molecular chaperones and protein quality control. *Cell* 2006; 125(3): 443-51.
- [16] Hartl FU, Hayer-Hartl M. Converging concepts of protein folding *in vitro* and *in vivo*. *Nat Struct Mol Biol* 2009; 16(6): 574-81.
- [17] Dobson CM. Protein misfolding, evolution and disease. *Trends Biochem Sci* 1999; 24(9): 329-32.
- [18] Clark LA. Protein aggregation determinants from a simplified model: cooperative folders resist aggregation. *Protein Sci* 2005; 14(3): 653-62.
- [19] Watters AL, Deka P, Corrent C, *et al*. The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. *Cell* 2007; 128(3): 613-24.
- [20] Richardson JS, Richardson DC. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci USA* 2002; 99(5): 2754-9.
- [21] Rousseau F, Schymkowitz JWH, Wilkinson HR, Itzhaki LS. Three-dimensional domain swapping in p13suc1 occurs in the unfolded state and is controlled by conserved proline residues. *Proc Natl Acad Sci USA* 2001; 98(10): 5596-601.
- [22] Liu Y, Eisenberg D. 3D domain swapping: As domains continue to swap. *Protein Sci* 2002; 11(6): 1285-99.
- [23] Liu Y, Gotte G, Libonati M, Eisenberg D. Structures of the two 3D domain-swapped RNase A trimers. *Protein Sci* 2002; 11(2): 371-80.
- [24] Uversky VN, Fink AL. Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochim Biophys Acta* 2004; 1698(2): 131-53.
- [25] Sanchez IE, Tejero J, Gomez-Moreno C, Medina M, Serrano L. Point mutations in protein globular domains: contributions from function, stability and misfolding. *J Mol Biol* 2006; 363(2): 422-32.
- [26] Monsellier E, Ramazzotti M, de Laureto PP, *et al*. The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution. *Biophys J* 2007; 93(12): 4382-91.
- [27] Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L. A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J Mol Biol* 2004; 342(1): 345-53.
- [28] Abeln S, Frenkel D. Disordered flanks prevent peptide aggregation. *PLoS Comput Biol* 2008; 4(12): e1000241.
- [29] Uversky VN, Fernandez A, Fink AL. Structural and conformational prerequisites for amyloidogenesis. In: Uversky VN, Fink AL, Eds. *Protein misfolding, aggregation, and conformational diseases*. Berlin: Springer Verlag 2006.
- [30] Hammarstrom P, Jiang X, Hurshman AR, Powers ET, Kelly JW. Sequence-dependent denaturation energetics: A major determinant in amyloid disease diversity. *Proc Natl Acad Sci USA* 2002; 99: 16427-32.
- [31] Wang Q, Johnson JL, Agar NY, Agar JN. Protein aggregation and protein instability govern familial amyotrophic lateral sclerosis patient survival. *PLoS Biol* 2008; 6(7): e170.
- [32] Lindberg MJ, Bystrom R, Boknas N, Andersen PM, Oliveberg M. Systematically perturbed folding patterns of amyotrophic lateral sclerosis (ALS)-associated SOD1 mutants. *Proc Natl Acad Sci USA* 2005; 102(28): 9754-9.
- [33] Chiti F, Bucciantini M, Capanni C, *et al*. Solution conditions can promote formation of either amyloid protofilaments or mature fibrils from the HypF N-terminal domain. *Protein Sci* 2001; 10(12): 2541-7.
- [34] Calloni G, Zoffoli S, Stefani M, Dobson CM, Chiti F. Investigating the effects of mutations on protein aggregation in the cell. *J Biol Chem* 2005; 280(11): 10607-13.
- [35] Ramirez-Alvarado M, Merkel JS, Regan L. A systematic exploration of the influence of the protein stability on amyloid fibril formation *in vitro*. *Proc Natl Acad Sci USA* 2000; 97(16): 8979-84.
- [36] Chiti F, Taddei N, Bucciantini M, *et al*. Mutational analysis of the propensity for amyloid formation by a globular protein. *EMBO J* 2000; 19(7): 1441-9.
- [37] Otzen DE, Oliveberg M. Salt-induced detour through compact regions of the protein folding landscape. *Proc Natl Acad Sci USA* 1999; 96(21): 11746-51.
- [38] Otzen DE, Kristensen O, Oliveberg M. Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: a structural clue to amyloid assembly. *Proc Natl Acad Sci USA* 2000; 97(18): 9907-12.
- [39] Dehay B, Bertolotti A. Critical role of the proline-rich region in Huntingtin for aggregation and cytotoxicity in yeast. *J Biol Chem* 2006; 281(47): 35608-15.
- [40] Steward A, Adhya S, Clarke J. Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily. *J Mol Biol* 2002; 318(4): 935-40.
- [41] Parrini C, Taddei N, Ramazzotti M, *et al*. Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation. *Structure* 2005; 13(8): 1143-51.
- [42] Chiti F, Calamai M, Taddei N, *et al*. Studies of the aggregation of mutant proteins *in vitro* provide insights into the genetics of amyloid diseases. *Proc Natl Acad Sci USA* 2002; 99: 16419-26.
- [43] Sandelin E, Nordlund A, Andersen PM, Marklund SS, Oliveberg M. Amyotrophic lateral sclerosis-associated copper/zinc superoxide dismutase mutations preferentially reduce the repulsive charge of the proteins. *J Biol Chem* 2007; 282(29): 21230-6.
- [44] Calamai M, Tartaglia GG, Vendruscolo M, Chiti F, Dobson CM. Mutational Analysis of the Aggregation-Prone and Disaggregation-Prone Regions of Acylphosphatase. *J Mol Biol* 2009; 387: 965-74.
- [45] Fowler SB, Poon S, Muff R, *et al*. Rational design of aggregation-resistant bioactive peptides: reengineering human calcitonin. *Proc Natl Acad Sci USA* 2005; 102(29): 10105-10.
- [46] Tompa P. Intrinsically unstructured proteins. *Trends Biochem Sci* 2002; 27(10): 527-33.
- [47] Uversky VN. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 2002; 11(4): 739-56.
- [48] Lawrence MS, Phillips KJ, Liu DR. Supercharging proteins can impart unusual resilience. *J Am Chem Soc* 2007; 129(33): 10110-2.
- [49] Vendruscolo M, Dobson CM. Chemical biology: More charges against aggregation. *Nature* 2007 Oct 4; 449(7162): 555.
- [50] Rudiger S, Germeroth L, SchneiderMergener J, Bukau B. Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. *EMBO J* 1997; 16(7): 1501-7.
- [51] Rudiger S, Mayer MP, Schneider-Mergener J, Bukau B. Modulation of substrate specificity of the DnaK chaperone by alteration of a hydrophobic arch. *J Mol Biol* 2000; 304(3): 245-51.
- [52] Xu W, Yuan X, Xiang Z, *et al*. Surface charge and hydrophobicity determine ErbB2 binding to the Hsp90 chaperone complex. *Nat Struct Mol Biol* 2005; 12(2): 120-6.
- [53] Knoblauch NT, Rudiger S, Schonfeld HJ, *et al*. Substrate specificity of the SecB chaperone. *J Biol Chem* 1999; 274(48): 34219-25.
- [54] Patzelt H, Rudiger S, Brehmer D, *et al*. Binding specificity of *Escherichia coli* trigger factor. *Proc Natl Acad Sci USA* 2001; 98(25): 14244-9.
- [55] Schlieker C, Weibezahn J, Patzelt H, *et al*. Substrate recognition by the AAA+ chaperone ClpB. *Nat Struct Mol Biol* 2004; 11(7): 607-15.
- [56] Tang Y-C, Chang H-C, Roeben A, *et al*. Structural features of the GroEL-GroES nano-cage required for rapid folding of encapsulated protein. *Cell* 2006; 125(5): 903-14.
- [57] Hegyi H, Tompa P. Intrinsically disordered proteins display no preference for chaperone binding *in vivo*. *PLoS Comput Biol* 2008; 4(3): e1000017.
- [58] von Bergen M, Barghorn S, Li L, *et al*. Mutations of tau protein in frontotemporal dementia promote aggregation of paired helical filaments by enhancing local beta-structure. *J Biol Chem* 2001; 276(51): 48165-74.
- [59] Hardy J. Testing times for the amyloid cascade hypothesis. *Neurobiol Aging* 2002; 23(6): 1073-4.
- [60] Conway KA, Harper JD, Lansbury PT, Jr. Fibrils formed *in vitro* from alpha-synuclein and two mutant forms linked to Parkinson's disease are typical amyloid. *Biochemistry* 2000; 39(10): 2552-63.
- [61] Ding F, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. *J Mol Biol* 2002; 324(4): 851-7.
- [62] Chen Y, Dokholyan NV. Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. *Mol Biol Evol* 2008; 25(8): 1530-3.

- [63] Wright CF, Teichmann SA, Clarke J, Dobson CM. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature* 2005; 438(7069): 878-81.
- [64] Broome BM, Hecht MH. Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. *J Mol Biol* 2000; 296(4): 961-8.
- [65] Tartaglia GG, Pechmann S, Dobson CM, Vendruscolo M. Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends Biochem Sci* 2007; 32(5): 204-6.
- [66] Lansbury PT, Lashuel HA. A century-old debate on protein aggregation and neurodegeneration enters the clinic. *Nature* 2006; 443(7113): 774-9.
- [67] Tartaglia GG, Caflisch A. Computational analysis of the *S. cerevisiae* proteome reveals the function and cellular localization of the least and most amyloidogenic proteins. *Proteins* 2007; 68(1): 273-8.
- [68] Hageman J, Vos MJ, van Waarde MA, Kampinga HH. Comparison of intra-organellar chaperone capacity for dealing with stress-induced protein unfolding. *J Biol Chem* 2007; 282(47): 34334-45.

---

Received: April 21, 2009

Revised: July 07, 2009

Accepted: July 09, 2009

© Reumers *et al.*; Licensee *Bentham Open*.

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.