Predicting aggregation prone sequences in proteins

De Baets Greet[1,2], Joost Schymkowitz[1,2,3] & Frederic Rousseau[1,2,3]

1 Switch Laboratory, VIB, University of Leuven, Leuven, Belgium
2 Switch Laboratory, Department of Cellular and Molecular Medicine, University of Leuven, B-3000 Leuven, Belgium
3 To whom correspondence should be addressed. E-mail: frederic.rousseau@switch.vib-kuleuven.be (F.R.); joost.schymkowitz@switch.vib-kuleuven.be (J.S.)

Key words/phrases: β-aggregation, APR, hydrophobicity, charge, β-sheet, thermodynamic stability, sequence-based, structure-based

Abstract

Due to its association with a diverse range of human diseases, the determinants of protein aggregation are studied intensively. It is generally accepted that the effective aggregation tendency of a protein depends on many factors such as folding efficiency towards the native state, thermodynamic stability of that conformation, intrinsic aggregation propensity of the polypeptide sequence and its ability to be recognized by the protein quality control system. The intrinsic aggregation propensity of a polypeptide sequence is related to the presence of short aggregation-prone regions (APRs) that self-associate to form intermolecular β-structured assemblies. These are typically short sequence segments (5-15 amino acids) that display high hydrophobicity, low net charge, and high tendency to form β-structures. As the presence of such APRs is a prerequisite for aggregation, a plethora of methods have been developed to identify APRs in amino acid sequences. In this review, the methodological basis of these approaches is discussed, as well as some practical applications.

Introduction

Misfolding of a polypeptide, either by deleterious mutations or stress conditions in the cell, can cause protein aggregation. For many years it has been investigated how monomer proteins stack into these aggregates, i.e. which interactions initiate the intermolecular association resulting in an aggregate. Several models have been proposed, such as β-aggregation, native aggregation, and 3D domain swapping. The latter posits that two or more protein chains exchange identical domains to form a strongly bound oligomer. β-aggregation refers to the formation of β-structure by exposure of short aggregation prone regions (APRs). No single model is likely to account for the properties of all aggregates formed from different (poly)peptides and under different conditions, but several lines of evidence suggest that β-aggregation is the most prevalent.

First of all, a study performed by Chiti *et al*. [1] illustrated that the aggregation rate of the α/β protein acylphosphatase is determined by two regions in the sequence. These regions have a high hydrophobicity and a high tendency to form β-sheet structure, pointing to aggregation driven by short APRs. Moreover, these regions are distinct from the folding nucleus, hinting at a competition between protein folding and aggregation. Next, the importance of these APRs to initiate aggregation was confirmed by several independent grafting studies where aggregation of an otherwise non-aggregating protein was induced through grafting of an APR from another protein. An example is the grafting of the mouse β2-microglobulin (β2M) with the APR present in human β2M. In contrast to the wild type mouse β2M, the chimera which contains an APR readily aggregates [2]. Moreover, it has been shown that the vast majority of proteins known to be associated with aggregation diseases contains an APR that determines the intrinsic aggregation propensity of a polypeptide [3].

These APRs are usually composed of 5-15 successively placed hydrophobic amino acids with a high beta-sheet propensity and a low net charge. Although most proteins possess one or several APRs, they are mostly protected from aggregation by being buried inside the hydrophobic core [4]. This points to another important determinant of effective aggregation namely protein stability. It is only upon destabilization of a protein through e.g. mutation that the APRs might be exposed, triggering aggregation (Figure 1). Examples are destabilizing mutations in SOD1, p53 and α-galactosidase resulting in protein aggregation.

Dependent on the degree of β-sheet organization, protein aggregation can refer to the formation of different macroscopic forms. The two extremes are amyloid fibrils and amorphous beta aggregates, with a whole range of morphologies in between (Figure 2) [5]. Amyloid fibrils are highly ordered and repetitive structures. In contrast, amorphous aggregates consist mostly of disordered polypeptide chains and show no macroscopic regularity using electron microscopy. Whether a protein will form an amorphous aggregate or an amyloid fibril depends on the amino acid sequence of the APR. In the case of amorphous aggregates, the only requirement is a short stretch with an overall high β -sheet propensity and neutral in charge [6]. In contrast, the less flexible amyloid fibrils are more position-specific with a very strict core and flanks that are more tolerant towards polar and charged residues [3].

Prediction of aggregation

As the presence of an APR is a requirement for the ubiquitous beta-aggregation mechanism, several methods have been developed to determine the intrinsic aggregation propensity of a protein by detecting the APRs in its sequence (Figure 3, step 1-3). The most common approaches evaluate either a) intrinsic amino-acid properties (sequence based) or b) the compatibility of the protein structural features with known amyloid fibril structures (structure based). As the structure-based methods are based on the structural information of amyloid fibrils, they are more specific for this type of aggregates. Machine-learning methods combining several predictors, e.g. AmylPred [7], form another approach. As these methods do not add additional physico-chemical or structural information, they are not discussed in this overview. For the alternative aggregation mechanisms such as 3D domain swapping and native protein aggregation, no methods have yet been developed.

General aggregation predictors

The main physico-chemical factors that promote aggregation of unfolded polypeptide chains have been characterized a decade ago: hydrophobicity, net charge, and propensity to form β-sheet and α-helical structure are correlated with aggregation propensity [8]. This original formula was extended with experimental variables such as protein concentration, solvent pH, and ionic strength to predict the absolute aggregation rates of unstructured peptides and natively unfolded proteins [9]. Based on these initial findings, several methods have been developed that generate aggregation propensity profiles, enabling the identification of regions with high intrinsic propensity for aggregation.

The **Zyggregator** method initially started purely from these principles, but later versions also included more sophisticated measures, such as the spatial relationship between aggregation prone residues and gatekeepers i.e. residues opposing aggregation. An upgraded version of this method is available which includes the protein flexibility and solvent accessibility. As such it tries to compensate for the fact that these APRs are under native conditions buried inside stable structural elements, unable to form the specific intermolecular interactions required for aggregation [10]. **TANGO**, a statistical thermodynamics algorithm, is another algorithm to identify the nucleation sites for aggregation by considering not only the factors described above, but also the competition between β-aggregate formation and other structural states such as α-helix, β-turn, β-strand and the random coil [6]. Another method is **SALSA** (Simple ALgorithm for Sliding Averages), which assumes a strong correlation between β-strand propensity and fibril formation. It calculates a mean β-strand propensity for each residue to identify the fibrillogenic hotspot [11]. On the other hand, **AGGRESCAN** identifies aggregation hot spots relying on an aggregation propensity scale for each of the amino acids [12]. This scale is based on the relative solubility of point mutants of amyloid β-peptide in *E. coli* [13]. Another method is **FoldAmyloid** that predicts the amyloid fibril-forming regions based on the mean packing density. Segments with a strong packing density are considered amyloidogenic [14].

Amyloid-specific predictors

High-resolution structural studies of fibrils from a number of different peptides have revealed that APRs associate in an intermolecular way through formation of a cross-β spine [15]. In this model, the APR has a tendency to pack into a β-sheet and the fibril grows as more segments of identical molecules stack into the β-sheet. Transmission electron microscopy revealed that amyloids are straight, unbranched fibrils with a diameter of 7-12 nm, made up by 2 to 6 protofilaments. Inside these protofilaments, intermolecular β-sheets perpendicular to the fibril axis are present, confirming the cross-β sheet motif. This motif is observed by X- ray reflections at 4.7 and 10 Å, corresponding to the spacing between β-strands and the distance between adjacent β-sheets, respectively [16]. Using microcrystal x-ray diffraction, Eisenberg and co-workers showed that amyloid protofibrils are stabilized by a steric zipper, indicating that the side chains of two identical β-sheets facing each other, intermesh in a zipper-like, tightly packed, highly complementary interface [15].

This wealth of structural information has been exploited by a) employing homology-modeling methods based on this structural data or b) by combining position-specific scoring matrices with such a homology modeling based scores. However, as these predictors are based on structural information of amyloid fibrils, they are, in contrast with sequence-based predictors, more specific for the amyloid fibril.

3D-profile method [17] and **PRE-AMYL** [18] are such structure-based methods that use the amyloid fibril structure as a template to define amino acid sequences compatible with the three-dimensional cross-β-spine structure. By determining the probability that a protein segment fits in this conformation, they identify APRs.

PASTA (Prediction of Amyloid STructure Aggregation) is another method based on the assumption that β-strands constituting the amyloid fibril have a preference for an in-register parallel or anti-parallel arrangement with minimal energy. Creating a dataset with these strictly defined secondary structures allowed the calculation of a pairing energy for each possible pair of residues, which is then used to score all possible stretches [19]. The Peptide Interaction Matrix Analyzer (**PIMA**) method is based on the same principle and threads each possible peptide stretch onto an in-register parallel or anti-parallel β-sheet structure [20]. **BETASCAN** is another algorithm based on β-strands pairing in the amyloid core [21].

Our own **Waltz** combines the sequence-based and structure based method by using a sequence based PSSM based on a dataset including both positive and negative peptides for fibril formation, a set of physicochemical properties, and a structure based PSSM [22].

Workflow of APR detection

Once the APRs are identified, it is crucial to investigate where the APRs reside in the protein structure (Figure 3, step 4). In most cases these contribute to the thermodynamic stability of a protein and are buried inside the protein core. Therefore, to estimate the effect of an amino acid change on the effective aggregation tendency, it is required to analyze the effect on a) intrinsic aggregation tendency using the aggregation predictors and b) protein stability using force fields [23,24]. In figure 3 (step 6), the effect of all known mutations in α-galactosidase on protein stability and

aggregation tendency was analyzed using SNPeffect [25] and their effect on both determinants was visualized using a MASS (Mutant Aggregation & Stability Spectrum) plot (Figure 3, step 6), i.e. a scatter plot of the effect on protein stability (ddG) versus the difference in aggregation tendency. This plots shows that the majority of the mutations does not alter the intrinsic aggregation propensity of a protein but decrease the thermodynamic stability (positive ddG) of a protein. Since protein stability is a cooperative effect dependent on many residues, whereas aggregation only depends on few residues, we can conclude that mutations are much more likely to trigger aggregation by affecting stability than by increasing intrinsic aggregation.

Limitations of current prediction methods

All algorithms discussed so far have been inspired by and tested against experimental data obtained *in vitro*, where the protein aggregates are in a buffered solution in the absence of other proteins. In contrast, the cell is a complex environment, so the question arises whether the predictors are also valid in this context. A study by Chiti *et al*. illustrated that in most cases there is an agreement between the predictions and *in vivo* experiments [26], justifying the use of these predictors.
In contrast, a recent study [27] using scrambled versions of the aggregating stretch of Huntington protein, illustrates that the algorithms do vary in their ability to correctly identify the aggregating ones. They propose several reasons for the under- and over-prediction. A possible reason for overprediction is the high peptide concentrations used in the experimental system to train the algorithms. Moreover, there is also a poor understanding of how a peptide contributes to the aggregation kinetics. Protein aggregation depends on both primary and secondary nucleation processes. In the primary pathway, aggregate formation results from interaction between soluble monomers. If this nucleation step is inefficient, aggregation will not occur.
Additionally, for most approaches the APRs identified computationally still need to become exposed by (partial) unfolding before they can actually nucleate protein aggregation. Therefore, 3D relationships that exist in the folded state are highly relevant to determine if a particular region is likely to become exposed in the first place. However, most methods do not take into account structural constraints and the modulatory context of the remaining protein. Therefore, there is a need for extension of these APR-detection methods with reliable methods to estimate protein stability.

Application of APR predictors

Define APR present in a protein

The above-described methods are very useful to estimate whether a protein of interest is prone to aggregation and to define which region is responsible for this aggregation propensity. Identification of these APRs makes it possible to design mutations that avoid aggregation. As there are several methods available, based on different assumptions, it seems wise to test them all. In the following section two proteins known to aggregate are analyzed.

Amyloid-beta (Aβ): Aβ40 versus Aβ42

Deposition in the brain of Aβ, which is generated from the amyloid precursor protein (APP), is one of the main hallmarks of Alzheimer's disease. Proteolytic cleavage of APP can result in peptides of different length and longer variants seem to have an increased aggregation propensity. In this example, we investigate how good different predictors predict i) the aggregation propensity of the Aβ peptide and ii) the increased aggregation tendency with C-terminal elongation. Two sequence-based methods (TANGO and Zyggregator) and two structure-based methods (Waltz and 3D profile) were used. In figure 4, you can notice that all predictors estimate that the Aβ peptide has some aggregation tendency. However, WALTZ fails to predict any aggregation propensity in the C-terminal part of the peptide and is not capable to detect an increased aggregation tendency upon C-terminal elongation. As this elongation induces amorphous aggregation, this could explain why WALTZ, which is more specific for amyloid fibrils, did not predict this observation.

Tumor suppressor p53 (p53)

p53 is a key regulator of the cell cycle and gets mutated in around 30% of all cancers. Previously, it has been shown that the DNA-binding domain of p53 is conformationally unstable and studies in our lab confirmed that destabilizing mutations in this domain can result in formation of cellular aggregates [28]. TANGO, Zyggregator, and 3D profile predict an APR around position 250 (APR1, sequence = ILTIITL). Zyggregator and 3D profile suggest also another APR around position 120 (APR2, sequence = SVTCTYS) (Figure 5). When analyzing where these APRs reside in the protein structure, it is clear that APR1 is buried inside the hydrophobic core and needs to become exposed to trigger aggregation. WALTZ does not predict any of these APRs, which is again due to the amorphous nature of p53 aggregates [28]. In general, we can conclude that WALTZ is specific for amyloid fibrils whereas the others predict APRs that can result in structures ranging from amorphous aggregates to amyloid fibrils.

Proteome wide analysis of aggregation to detect evolutionary imprints

Using the power of the aggregation prediction algorithms described in the previous section, the aggregation load of several proteomes has been analyzed in detail with different methods. In all these studies, a clear evolutionary pressure to counteract aggregation was apparent.

A first study illustrated that the vast majority of proteins in any proteome contains at least one and generally several APRs [29] that can nucleate aggregation by assembling into intermolecular β-structures. A first way to avoid misfolding and aggregation is proper folding into the compact structure [30]. However, in the case of intrinsically disordered polypeptides (IDPs), the whole backbone is exposed to the solvent, so folding cannot play its protective role. Therefore several specific sequence adaptations are present to maintain their solubility and prevent aggregation: these proteins have a high net charge and low hydrophobicity [31], a lower number of APRs [32], and a higher proline content [33].

Another way to prevent aggregation is interrupting the contiguous stretches of hydrophobic residues by placement of charged side chains acting as gatekeepers [34]. Rousseau and co-workers revealed a strong enrichment of charged residues (Arg, Lys, Glu, Asp) and proline at the flanks of these APRs [29]. Their study showed that 90% of all APRs are capped with at least one gatekeeper residue (Arg, Lys, Glu, Asp or

Pro), with a preference for positive charged residues at regions with the highest aggregation propensity. These gatekeepers counteract aggregation by i) charge repulsion (Arg, Lys, Glu, Asp); ii) being large and flexible (Arg and Lys); or iii) being incompatible with the beta-structure (Pro and Gly). It is important to note that gatekeepers do not avoid aggregation, but reduce the aggregation rate sufficiently to tip the balance towards native protein folding.

The importance of strategic placement of charges to avoid aggregation was already illustrated by the observation that supercharged proteins are remarkably resistant to aggregation and refold efficiently [35]. Recent studies also discovered that gatekeepers facilitate the recognition of APRs by molecular chaperones, such as Hsp70 [36]. This Hsp70 system can slow down the aggregation process by allowing polypeptide chains to fold or by directing them to degradation. The observation that mutations disrupting gatekeeper patterns are more frequently disease-associated mutation than neutral mutations also points to their functional role [37].

Moreover, the extent to which selective pressure minimizes aggregation tendency is determined by the biological context. Using the aforementioned methods, several interesting observations were made. First, proteins forming oligomeric complexes have a lower aggregation propensity than those operating in free form [38]. As they constantly interact with other polypeptides, they are at increased risk of aggregation and therefore the aggregation tendency should be minimized. Comparably, the sequence similarity of APRs between different subunits is minimized in multi-domain proteins, also illustrating the sequence- specificity of aggregation [39].
Second, essential proteins were found to have a lower aggregation propensity, emphasizing evolutionary pressure to minimize aggregation propensity [40].
Third, Monsellier and co-workers demonstrated that long proteins, with slower folding rates [41], have less pronounced aggregation peaks [42]. The aggregation propensity seems to be anti-correlated with organism complexity [43], and evolutionary protection mechanisms are more pronounced in thermophilic compared to mesophilic proteins [44]. It was also shown that aggregation propensity inversely correlates with gene expression [45] and with protein turnover rate [46].

Conclusion

As aggregation is a detrimental process for the cell, considerable time has been invested to investigate which parameters affect the effective aggregation propensity of the cell and to consequently use this knowledge to identify proteins prone to aggregate. Although several aggregation mechanisms have been identified, β-aggregation seems to be the most prevalent mechanism. It is based on the formation of β-structure by exposure of short aggregation prone regions (APRs). Nowadays, several methods are available to predict the presence of these APRs in a protein and it is advised to combine these to obtain a reliable result. Moreover, beside the presence of APRs, protein stability is another important determinant, which is often not taken into account. Development of next-generation aggregation predictors should therefore include a reliable estimation of thermodynamic stability [10].

Summary

- Beside 3D domain swapping and native aggregation, β-aggregation is the most prevalent.

- β-aggregation is driven by exposure of APRs forming an intermolecular β-structure
- APRs typically contain 5-15 successively placed hydrophobic amino acids with a high beta-sheet propensity and a low net charge.
- Thermodynamic stability is another important parameter affecting the effective aggregation propensity.
- Several methods, either structure- or sequence-based, predict the presence of these APRs.
- Presence of an APR is not the only prerequisite, therefore measures of protein stability should be taken into account.

References

[1] Chiti, F., Taddei, N., Baroni, F., Capanni, C., Stefani, M., Ramponi, G. and Dobson, C.M. (2002). Kinetic partitioning of protein folding and aggregation. Nat Struct Biol 9, 137-43.

[2] Ivanova, M.I., Sawaya, M.R., Gingery, M., Attinger, A. and Eisenberg, D. (2004). An amyloid-forming segment of beta2-microglobulin suggests a molecular model for the fibril. Proc Natl Acad Sci U S A 101, 10584-9.

[3] Lopez de la Paz, M. and Serrano, L. (2004). Sequence determinants of amyloid fibril formation. Proc Natl Acad Sci U S A 101, 87-92.

[4] Dobson, C.M. (2003). Protein folding and misfolding. Nature 426, 884-90.

[5] Rousseau, F., Schymkowitz, J. and Serrano, L. (2006). Protein aggregation and amyloidosis: confusion of the kinds? Curr Opin Struct Biol 16, 118-26.

[6] Fernandez-Escamilla, A.M., Rousseau, F., Schymkowitz, J. and Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 22, 1302-6.

[7] Frousios, K.K., Iconomidou, V.A., Karletidi, C.M. and Hamodrakas, S.J. (2009). Amyloidogenic determinants are usually not buried. BMC Struct Biol 9, 44.

[8] Chiti, F., Stefani, M., Taddei, N., Ramponi, G. and Dobson, C.M. (2003). Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature 424, 805-8.

[9] DuBay, K.F., Pawar, A.P., Chiti, F., Zurdo, J., Dobson, C.M. and Vendruscolo, M. (2004). Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. J Mol Biol 341, 1317-26.

[10] Tartaglia, G.G., Pawar, A.P., Campioni, S., Dobson, C.M., Chiti, F. and Vendruscolo, M. (2008). Prediction of aggregation-prone regions in structured proteins. J Mol Biol 380, 425-36.

[11] Zibaee, S., Makin, O.S., Goedert, M. and Serpell, L.C. (2007). A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone. Protein Sci 16, 906-18.

[12] Conchillo-Sole, O., de Groot, N.S., Aviles, F.X., Vendrell, J., Daura, X. and Ventura, S. (2007). AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinformatics 8, 65.

[13] Sanchez de Groot, N., Pallares, I., Aviles, F.X., Vendrell, J. and Ventura, S. (2005). Prediction of "hot spots" of aggregation in disease-linked polypeptides. BMC Struct Biol 5, 18.

[14] Galzitskaya, O.V., Garbuzynskiy, S.O. and Lobanov, M.Y. (2006). Prediction of amyloidogenic and disordered regions in protein chains. PLoS Comput Biol 2, e177.
[15] Sawaya, M.R. et al. (2007). Atomic structures of amyloid cross-beta spines reveal varied steric zippers. Nature 447, 453-7.
[16] Serpell, L.C., Sunde, M., Benson, M.D., Tennent, G.A., Pepys, M.B. and Fraser, P.E. (2000). The protofilament substructure of amyloid fibrils. J Mol Biol 300, 1033-9.
[17] Thompson, M.J., Sievers, S.A., Karanicolas, J., Ivanova, M.I., Baker, D. and Eisenberg, D. (2006). The 3D profile method for identifying fibril-forming segments of proteins. Proc Natl Acad Sci U S A 103, 4074-8.
[18] Zhang, Z., Chen, H. and Lai, L. (2007). Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. Bioinformatics 23, 2218-25.
[19] Trovato, A., Seno, F. and Tosatto, S.C. (2007). The PASTA server for protein aggregation prediction. Protein Eng Des Sel 20, 521-3.
[20] Bui, J.M., Cavalli, A. and Gsponer, J. (2008). Identification of aggregation-prone elements by using interaction-energy matrices. Angew Chem Int Ed Engl 47, 7267-9.
[21] Bryan, A.W., Jr., Menke, M., Cowen, L.J., Lindquist, S.L. and Berger, B. (2009). BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. PLoS Comput Biol 5, e1000333.
[22] Maurer-Stroh, S. et al. (2010). Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Methods 7, 237-42.
[23] Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. and Serrano, L. (2005). The FoldX web server: an online force field. Nucleic Acids Res 33, W382-8.
[24] Rohl, C.A., Strauss, C.E., Misura, K.M. and Baker, D. (2004). Protein structure prediction using Rosetta. Methods Enzymol 383, 66-93.
[25] De Baets, G., Van Durme, J., Reumers, J., Maurer-Stroh, S., Vanhee, P., Dopazo, J., Schymkowitz, J. and Rousseau, F. (2012). SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Res 40, D935-9.
[26] Belli, M., Ramazzotti, M. and Chiti, F. (2011). Prediction of amyloid aggregation in vivo. EMBO Rep 12, 657-63.
[27] Roland, B.P., Kodali, R., Mishra, R. and Wetzel, R. (2013). A serendipitous survey of prediction algorithms for amyloidogenicity. Biopolymers
[28] Xu, J. et al. (2011). Gain of function of mutant p53 by coaggregation with multiple tumor suppressors. Nat Chem Biol 7, 285-95.
[29] Rousseau, F., Serrano, L. and Schymkowitz, J.W. (2006). How evolutionary pressure against protein aggregation shaped chaperone specificity. J Mol Biol 355, 1037-47.
[30] Watters, A.L., Deka, P., Corrent, C., Callender, D., Varani, G., Sosnick, T. and Baker, D. (2007). The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. Cell 128, 613-24.

[31] Uversky, V.N. and Fink, A.L. (2004). Conformational constraints for amyloid fibrillation: the importance of being unfolded. Biochim Biophys Acta 1698, 131-53.
[32] Linding, R., Schymkowitz, J., Rousseau, F., Diella, F. and Serrano, L. (2004). A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. J Mol Biol 342, 345-53.
[33] Tompa, P. (2002). Intrinsically unstructured proteins. Trends Biochem Sci 27, 527-33.
[34] Otzen, D.E., Kristensen, O. and Oliveberg, M. (2000). Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: a structural clue to amyloid assembly. Proc Natl Acad Sci U S A 97, 9907-12.
[35] Lawrence, M.S., Phillips, K.J. and Liu, D.R. (2007). Supercharging proteins can impart unusual resilience. J Am Chem Soc 129, 10110-2.
[36] Van Durme, J., Maurer-Stroh, S., Gallardo, R., Wilkinson, H., Rousseau, F. and Schymkowitz, J. (2009). Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. PLoS Comput Biol 5, e1000475.
[37] Reumers, J., Schymkowitz, J. and Rousseau, F. (2009). Using structural bioinformatics to investigate the impact of non synonymous SNPs and disease mutations: scope and limitations. BMC Bioinformatics 10 Suppl 8, S9.
[38] Chen, Y. and Dokholyan, N.V. (2008). Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. Mol Biol Evol 25, 1530-3.
[39] Wright, C.F., Teichmann, S.A., Clarke, J. and Dobson, C.M. (2005). The importance of sequence diversity in the aggregation and evolution of proteins. Nature 438, 878-81.
[40] Tartaglia, G.G. and Caflisch, A. (2007). Computational analysis of the S. cerevisiae proteome reveals the function and cellular localization of the least and most amyloidogenic proteins. Proteins 68, 273-8.
[41] Ivankov, D.N., Garbuzynskiy, S.O., Alm, E., Plaxco, K.W., Baker, D. and Finkelstein, A.V. (2003). Contact order revisited: influence of protein size on the folding rate. Protein Sci 12, 2057-62.
[42] Monsellier, E., Ramazzotti, M., Taddei, N. and Chiti, F. (2008). Aggregation propensity of the human proteome. PLoS Comput Biol 4, e1000199.
[43] Tartaglia, G.G., Pellarin, R., Cavalli, A. and Caflisch, A. (2005). Organism complexity anti-correlates with proteomic beta-aggregation propensity. Protein Sci 14, 2735-40.
[44] Thangakani, A.M., Kumar, S., Velmurugan, D. and Gromiha, M.S. (2012). How do thermophilic proteins resist aggregation? Proteins 80, 1003-15.
[45] Tartaglia, G.G., Pechmann, S., Dobson, C.M. and Vendruscolo, M. (2007). Life on the edge: a link between gene expression levels and aggregation rates of human proteins. Trends Biochem Sci 32, 204-6.
[46] De Baets, G., Reumers, J., Delgado Blanco, J., Dopazo, J., Schymkowitz, J. and Rousseau, F. (2011). An evolutionary trade-off between protein turnover rate and protein aggregation favors a higher aggregation propensity in fast degrading proteins. PLoS Comput Biol 7, e1002090.

Legends of figures:

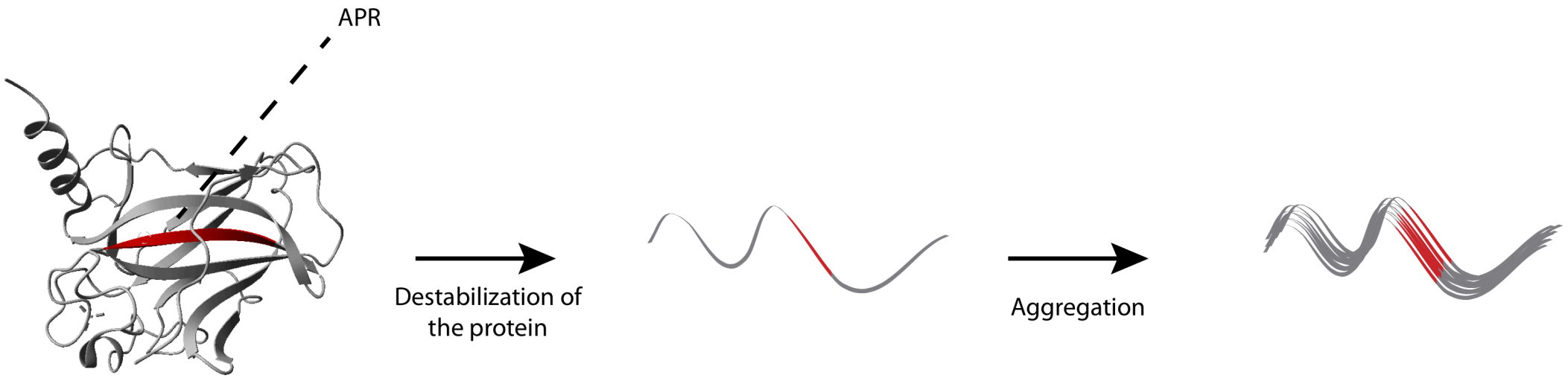
Figure 1: Schematic representation of protein aggregation through short stretches. In the folded state, the APR is buried into the globular native structure. Only upon destabilization, the protein exposes an aggregation nucleating stretch (APR). These APRs may align into an intermolecular β-sheet, nucleating the formation of a protein aggregate.

Figure 2: Transmission electron microscopy images of aggregating peptides. Selection of peptides displaying a wide range of morphologies: from a completely amorphous (left) to a highly ordered structure (right).

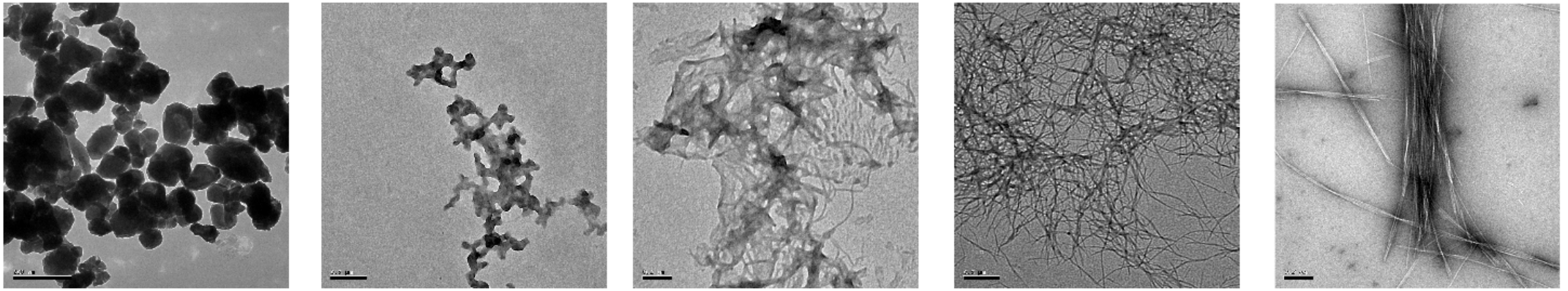
Figure 3: Workflow to determine whether a protein is prone to aggregation.

Figure 4: Predicted aggregation tendency of Aβ(40) and Aβ(42). A) TANGO and WALTZ, B) Zyggregator and C) 3D profile output for Aβ(40) and Aβ(42). A red line indicates the threshold for Zyggregator and 3D profile.

Figure 5: Predicted aggregation tendency of p53. A) TANGO and WALTZ, B) Zyggregator and C) 3D profile output for p53. A red line indicates the threshold for Zyggregator and 3D profile.


APR
Destabilization of
the protein
Aggregation

Degree of β-sheet organization



1) Retrieve amino acid sequence

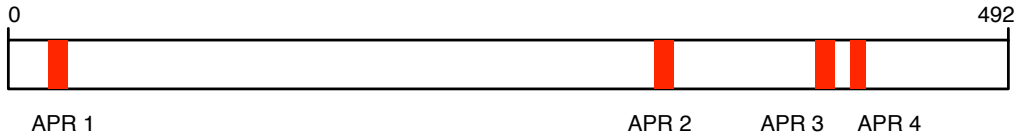
>AGAL_HUMAN
MQLRNPELHLGCALALRFLALVSWD....EWTSRLRSHINPTGTVLLQLENTMQMSLKDLL

2) Feed aggregation predictors

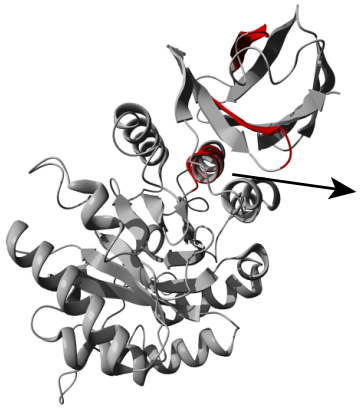
Waltz Tango

Betascan Zyggregator

3) Determine where the APRs are located in the sequence



4) Determine where the APRs are located in the protein structure

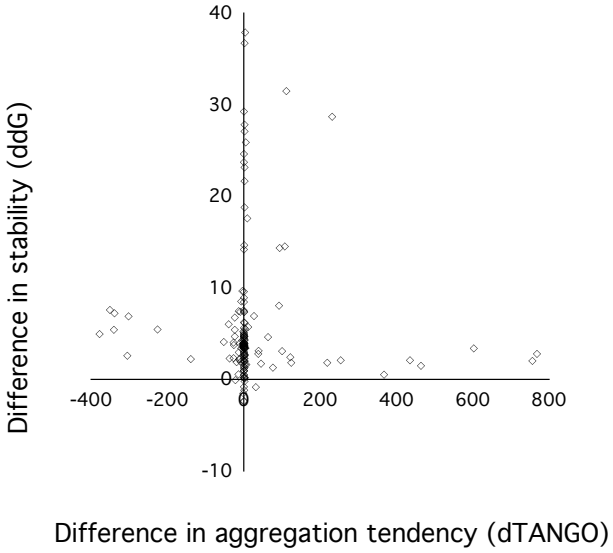


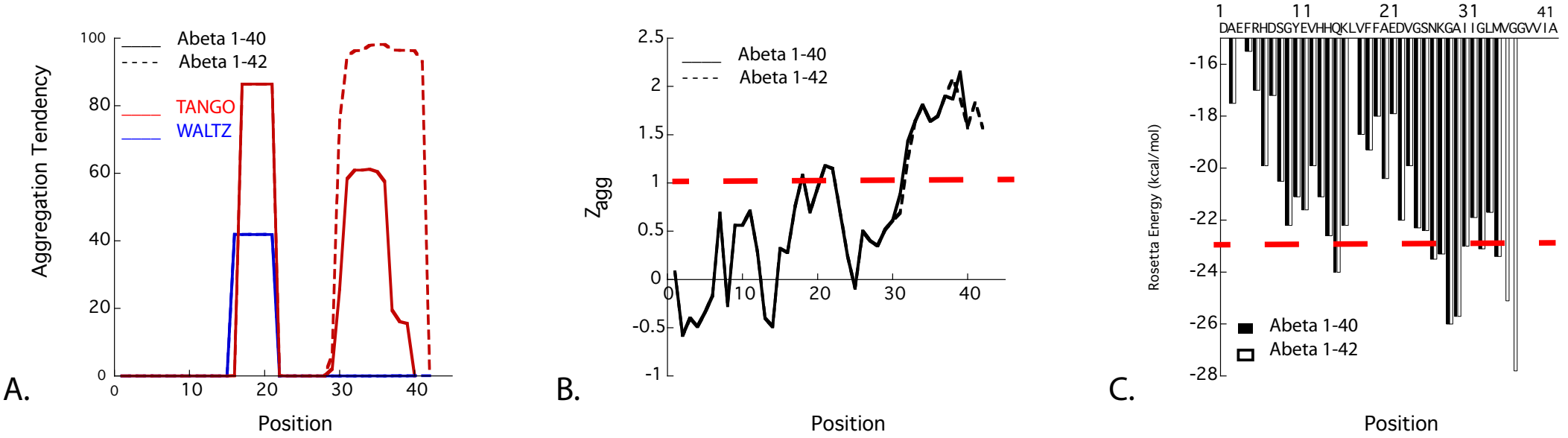
APR buried inside hydrophobic core

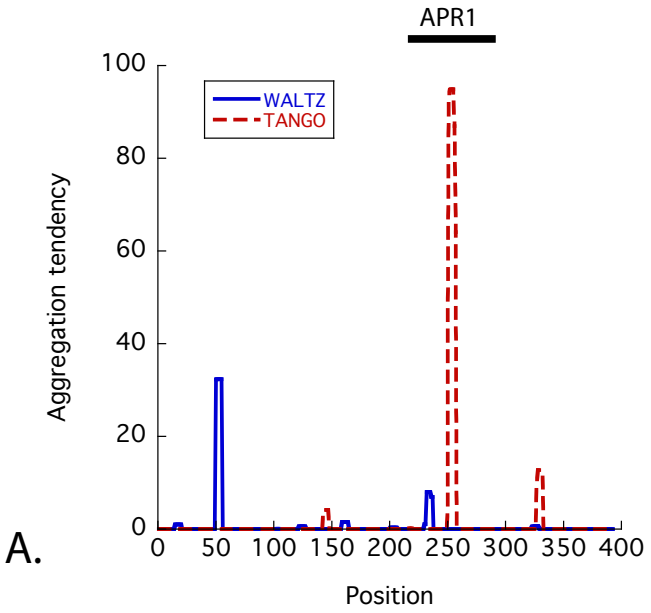
5) Retrieve stability of this region or estimate effect of a mutation on protein stability

Rosetta FoldX

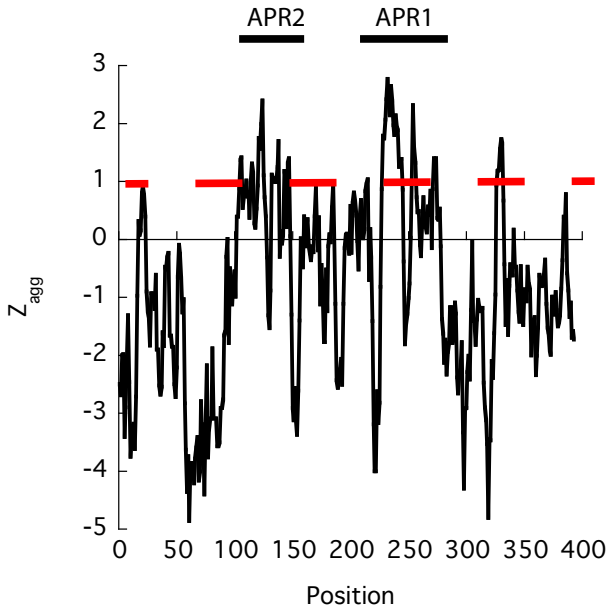
6) Analyze effect of mutations on both a) intrinsic aggregation tendency and b) protein stability



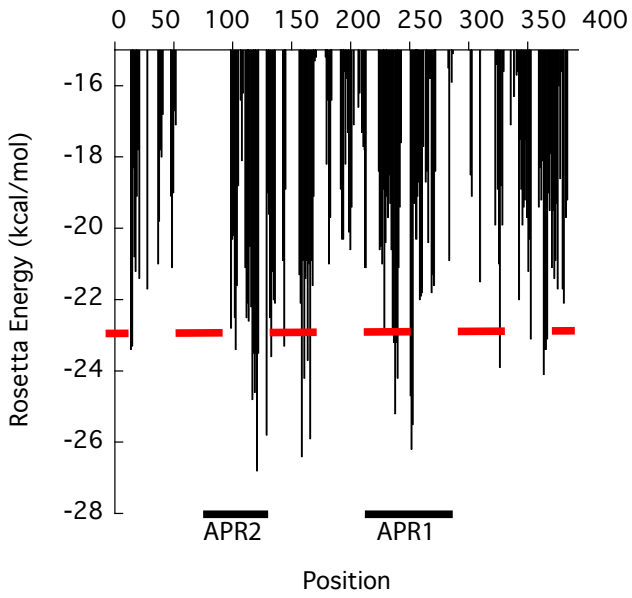

A.
Aggregation Tendency
Abeta 1-40
Abeta 1-42
TANGO
WALTZ
Position
B.
z_{agg}
Abeta 1-40
Abeta 1-42
Position
C.
Rosetta Energy (kcal/mol)
DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA
Abeta 1-40
Abeta 1-42
Position


APR1
WALTZ
TANGO
Aggregation tendency
100
80
60
40
20
0
0 50 100 150 200 250 300 350 400
Position

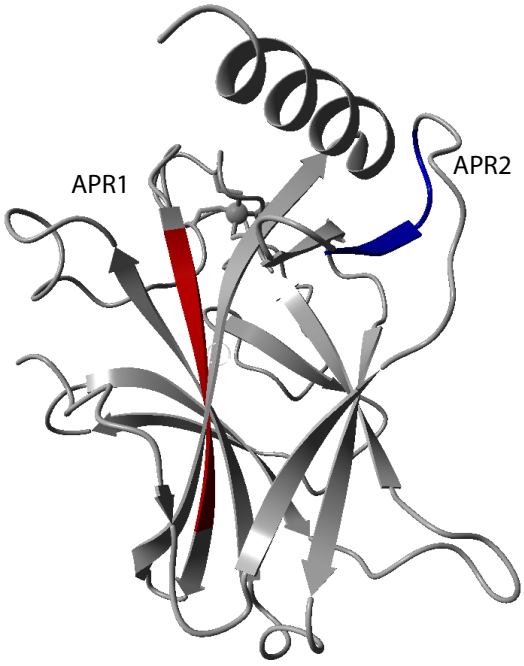
A.


APR2
APR1
Z_{agg}
3
2
1
0
-1
-2
-3
-4
-5
0 50 100 150 200 250 300 350 400
Position

B.


0 50 100 150 200 250 300 350 400
Rosetta Energy (kcal/mol)
-16
-18
-20
-22
-24
-26
-28
APR2
APR1
Position

C.


APR1
APR2

D.