Database of Peptides Susceptible to Aggregation as a Tool for Studying Mechanisms of Diseases of Civilization

Pawel P. Wozniak[1], Jean-Christophe Nebel[2], and Malgorzata Kotulska[1]*

[1] Wroclaw University of Technology, Faculty of Fundamental Problems of Technology, Department of Biomedical Engineering, Poland
[2] Kingston University, Faculty of Science, Engineering and Computing, School of Computing and Information Systems, United Kingdom

Abstract. We introduce a database containing peptides related to diseases arising from protein aggregation. The general database AmyLoad includes all experimentally studied protein fragments that could be involved in erroneous protein folding, leading to amyloid formation. The database has been extended since its first release with regard to new instances of peptides or their fragments. Moreover, information of related diseases have been added to all entries, whenever available. Currently the database includes all available peptides tested for their potential amyloid properties, obtained from diverse resources, creating the largest dataset available at one place. This enables comparison between properties of amyloid and non-amyloid peptides. We could also select candidates for the most pathogenic peptides, involved in several diseases related to protein aggregation. We also discuss a need for sub-databases of different structures, such as related to $\beta\gamma$-crystallins - a protein family occurring in the eye lens. Misfolding of these proteins may lead to various forms of cataract. Those freely available internet services can facilitate finding the link between a protein sequence, its propensity to aggregation and the resulting disease, as well as support research on their pharmacological treatment and prevention.

1 Introduction

Many diseases, especially neurodegenerative, result from protein fragments forming aggregates. This occurs when a cell environment fosters the partial unfolding of protein chains or their fragmentation, in a way that the parts prone to joining with other protein fragments are exposed. For the majority of proteins, considerable conformational rearrangement must have occurred to initiate the aggregation process. Such changes cannot take place in the typical tightly packed native protein conformation, due to the constraints of the tertiary structure. Thus, formation of a non-native partially unfolded conformation is required, presumably enabling specific intermolecular interactions, including electrostatic attraction,

* Corresponding author: malgorzata.kotulska@pwr.edu.pl

hydrogen bonding and hydrophobic contacts. This partial unfolding can be influenced by various factors, such as protein high concentration, high temperature, low pH, binding metals, or exposition to UV light.

Initially, the resulting molecules form clusters consisting of a few elements, which are called oligomers. Next, they grow into larger aggregates. Aggregation of proteins or their fragments may lead to amorphous (unstructured) clusters or amyloid (highly ordered) unbranched fibrils. Independently of the protein sequence and its original structure, aggregates always display a common cross-β structure. The distinctive structure of the steric zipper enables the selective detection of amyloids from amorphous aggregates using either a variety of microscopic techniques or fluorescence of probes with which they form compounds.

Amyloid fibrils have been observed in the brains of people suffering from Alzheimer's disease. They are also associated with Parkinson's disease, amyotrophic lateral sclerosis and Huntington's disease, as well as many other conditions, even non-neurodegenerative diseases such as type 2 diabetes and some types of cataract. Cells in tissues containing these fibrils exhibit very high mortality. However, the reasons for this cytotoxicity have not been resolved. In recent years the occurrence rate of diseases characterized by accumulation of protein deposits has increased significantly. These disorders are sometimes called diseases of civilization since they are more prevalent in developed countries where life expectancy is higher. Unfortunately, their mechanisms are still poorly understood. Although studies indicate that these diseases to some extent have a genetic basis, the influence of lifestyle cannot be excluded. Unfortunately dissolution of peptide aggregates is very difficult, especially for amyloids which are resistant to activity of proteolytic enzymes and chemical compounds due to the specific and highly ordered structure of their steric zipper.

Cataract is among the diseases associated with protein aggregation. Age-related cataract is a major burden on public health: this is the most common cause of blindness worldwide, affecting tens of millions of people. The lens fibre cells are essentially composed of crystallin proteins, which are among the most highly concentrated intracellular proteins in the human body. $\beta\gamma$-crystallins define a superfamily of crystallins sharing similar sequence and structure. Despite a large set of literature concerning the family of $\beta\gamma$-crystallins, there is no dedicated service containing all available genotypic and phenotypic data, as well as tools for resolving molecular mechanisms of the disease and supporting the development of potential pharmacological treatments.

Currently, it is believed that short peptide sequences of amyloidogenic properties (called hot-spots) can be responsible for aggregation of amyloid proteins. These 4-10 residue long fragments (typically hexapeptides) have a high propensity for strong interactions that lead to protein aggregation. Previous studies have suggested that amyloidogenic fragments may have regular characteristics, not only with regard to averaged physicochemical properties of their amino acids, but also the order of amino acids in the sequence. There have been attempts to predict the sequence of such peptides by computational modelling. Physics and chemistry based models have been used, including FoldAmyloid [1]. This method

is based on the density of the protein contact sites. Other methods perform threading a peptide on an amyloid fiber backbone, followed by determination of its energy and stability [2][3][4]. Statistical approaches include production of frequency profiles, such as the WALTZ method [5] and machine learning methods, which have been used by our team [6][7]. Some other approaches, mostly biophysical-based, enable classification of hot spots for non-amyloid aggregation. Recently, AGGRESCAN3D has been proposed to estimate more accurately aggregation propensity by performing 3D structure based analysis [8]. All of these methods, although promising, have faced difficulties due to limited amount of experimental data available for their construction and validation.

Knowledge in the field of diseases related to protein aggregation is still patchy, no global view of the problem is available and the link between the molecular level and the phenotype is still generally missing. In addition, although a large amount of information is available, its dispersion in separate publications, data sets and web services hampers research development. To fill in this gap, we proposed a new web service, AmyLoad [9], which is devoted to protein aggregation and can facilitate global research in the field. The service is focused on general amyloidogenic peptides. However studies on specific cataract related aggregation have led us to the idea of separate sub-database services, including more specific and detailed information, even if some is only predicted by software tools.

2 Results

AmyLoad is a website which gathers information about known, experimentally-derived, amyloidogenic and non-amyloidogenic amino acid protein fragments [9]. The data comes from literature studies and various data sources, which are WALTZ-DB [5], AmylFrag, AmylHex [10], and datasets used to validate such methods like TANGO [11] and AGGRESCAN [12]. Although these data sources contain protein fragments according to different specific features, the fact that one fragment can belong to more than one dataset makes these databases difficult to work with. In addition, data filtration is usually unavailable. In this category of datasets WALTZ-DB and AmylHex contain only protein fragments which are composed of six amino acids. AmylFrag possesses 45 literature-derived fragments which are longer than six residues. TANGO and AGGRESCAN are well known amyloidogenicity prediction methods, which are based on different sequence analysis algorithms. They were also validated using original experimental data. TANGO uses statistical mechanics algorithm based on the physicochemical principles of beta-sheet formation to predict protein aggregation. It was tested on a set of 179 peptides obtained from the literature and 71 new peptides derived from human disease-related proteins [11]. Alternatively, AGGRESCAN takes advantage of the aggregation-propensity scale for amino acids. It was trained on a database of 57 experimentally known amyloidogenic proteins [12]. These training data have been included into our database.

The website was built using the Django web framework and a MySQL database. Figure 1 presents the AmyLoad database tables and relations between them. The

most important table contains detailed information about each fragment such as name, residue sequence, aggregation ability, experimental conditions, and data about its first occurrence in the database: user and date. Each protein fragment is related to only one protein record stored in another table. While there can be more than one protein fragment related to a single protein record, the residue sequence and protein record id attributes are unique for each record of the protein fragment. Other tables store information about references related to the record of the fragment, datasets of origin, experimental methods used for their discovery, and their saves in private users sessions. Records of these tables are in many-to-many relation with protein fragment record. Information about registered users and administrators are stored securely taking advantage of the common user authentication system provided by Django.

The principal function of the AmyLoad website is data browsing and filtering. Currently, the data can be filtered according to protein name, aggregation propensity, residue sequence length, and sequence substring. Selected fragment records can be moved to the temporary session field where they can be saved for later studies. The data about selected fragments can be downloaded in several file formats, such as the CSV, SSV (Semicolon Separated Values), XML, and FASTA. The SSV and XML files contain all information gathered in the database for selected fragments. The CSV file includes only that information which is visible in the browsing table, i.e. protein name, fragment name, residue sequence, and aggregation propensity. The FASTA file contains information in the FASTA format, which is well known for those who work in bioinformatics. It is a text-based format which represents nucleotide or protein sequences as a single-letter code. Each sequence in the FASTA format can be preceded with an additional sequence information line which begins with the greater-than (>) symbol. In the AmyLoad website, that line contains the AmyLoad index, protein name, and fragment name.

New fragments can be submitted into the database in two ways by either filling in the website form or uploading a file in a proper XML format. The first option is recommended for the submission of a single fragment which has no more than one reference and dataset of origin. This is the most common manner of fragment submission. Taking advantage of the XML format, users can add at once multiple fragment records including potentially several references and datasets of origin each. The AmyLoad XML format is explained in details on the help page where examples and downloadable files are provided. Submitted fragments are visible to the general public only after reviewer approval. Until then, they are only accessible on the personal web page of the user who submitted them.

Following submission of protein sequences in the FASTA format, the AmyLoad website allows for their analysis with several implemented amyloidogenicity predictors, i.e. FoldAmyloid [1], AGGRESCAN, and FISH [7]. Implementations of all these methods were validated through comparison of their results with those of their original online implementations. FoldAmyloid and AGGRESCAN are based on sliding window algorithms. FoldAmyloid analyzes experimentally-

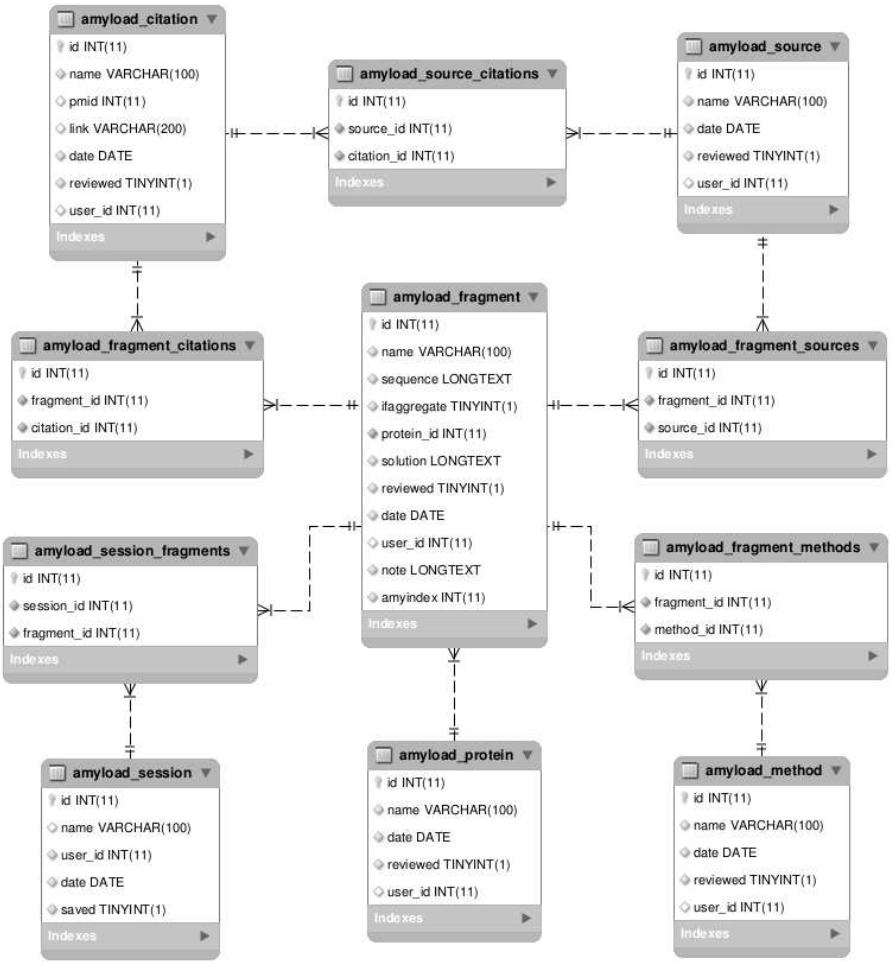

Fig. 1. AmyLoad database structure

derived expected probability of hydrogen bonds formation and expected packing density of amino acids in the sequence of interest [1]. AGGRESCAN, on the other hand, calculates the aggregation-propensity scale for residues derived from in vivo experiments [12]. Finally, FISH is a machine learning method which was created in our group, based on the site-specific co-localization of aminoacid pairs. The results of the analysis for any of the chosen methods are sent by email to users using a binary format, where 1 means that the associated residue in the submitted sequence belongs to the amyloidogenic fragment. Together with the implemented methods, AmyLoad allows its users to search its database for sequence fragments which occur in submitted FASTA sequences. The advantage of using the AmyLoad implementations instead of the original modelling tools is that users can run calculations for several submitted fragments at once. In addition, results of different methods are presented in the email in a comparable format which makes AmyLoad a simple consensus amyloidogenicity predictor. On the analysis website, references and links to the original implementations of FoldAmyloid, AGGRESCAN, FISH, and other 13 well known amyloidogenicity predictors are also provided.

There are three types of users interacting with the AmyLoad website. The first one is the common non-registered user who can use the entire analysis website, and browse and filter the data about sequence fragments gathered in the database. The second one is a registered user who can create temporary sessions and save them for later studies. Furthermore, only the registered user is able to submit new fragments into the database. Finally, there are the database administrators who have all the privileges of non-registered and registered users, together with the ability to review the submitted fragments. Only after the administrator-reviewer approval, a fragment becomes visible to the general public on the browsing web page.

Currently, there are 1477 unique entries in the AmyLoad database, which come from over 150 different proteins. Figure 2 shows the distribution of sequence lengths within the deposited fragments. The website contains also information about almost 100 references related to the amyloidogenicity topic. According to the literature, peptide fragments deposited in the AmyLoad were analysed by almost 20 different experimental methods such as electron microscopy or thioflavin dyeing.

Collecting all sequences diagnosed for amyloidogenicity enabled finding a pattern in the contents of amyloid and non-amyloid fragments. We tested several sets of peptides within different length ranges. Exemplary results are shown in Fig. 3 (hexapeptides which are the best studied with regard to their aggregation propensity) and in Fig. 4 (fragments of all lengths). The study showed that amyloid fragments are rich in valine (V), isoleucine (I), leucine (L) and phenyloalanine (F), which appeared in the instances of almost all lengths. It proves an excess of non-polar neutral aminoacids with high propensity to form beta strands. Interestingly, non-amyloidogenic peptides are characterized with increased contents of non-polar charged aminoacids, especially positively charged lysine (K) and arginine (R), and to lesser extent negatively charged aspartic acid

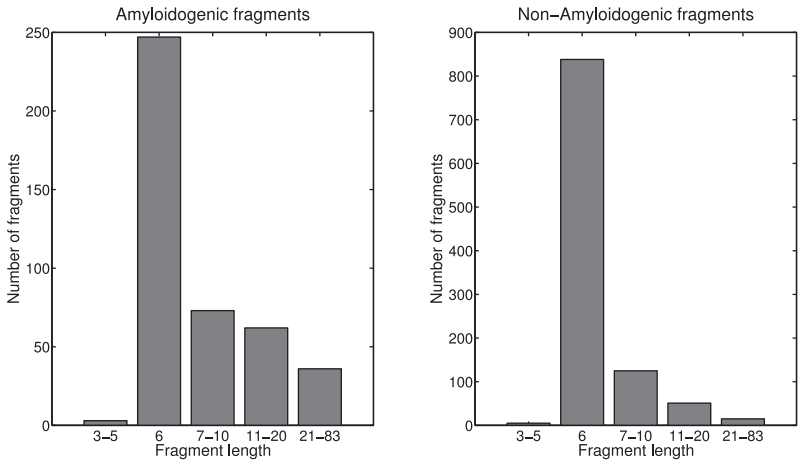

Fig. 2. Lengths of fragments deposited in AmyLoad

(D) and glutamic acid (E). Enrichement of peptides with charged aminoacids impacts on their non-amyloidogenic properties, which may be considered for applications towards changing the amyloidogenic properties of peptides.

The database has been extended by diseases related to amyloid fragments. It enabled to observe that majority of the fragments (44) are related to the Alzheimer's disease, then to prion diseases (26 fragments), type II diabetes (20 fragments), dialysis-related amyloidosis (16 fragments), and Amyotrophic Lateral Sclerosis (11 fragments). Interesingly, some of the amyloidogenic fragments have been shown as involved in up to three diseases. These include certain fragments of alpha synuclein - 8 fragments of this protein can lead either to Alzheimer's disease, or Parkinson's disease, or dementia with Lewy bodies. Another protein involved in 3 different diseases is τ-protein - with 6 fragments that can lead to Alzheimer's disease, Pick's disease, and progressive supranuclear palsy. However, a number of entries has not been associated with any disease, yet.

A general database, such as AmyLoad cannot contain sequences that have not been experimentally confirmed with regard to their misfolding properties, even though they may contribute to understanding of a disease. On the other hand, several protein features could be predicted with modelling tools, which would be very helpful in studying molecular mechanisms of the diseases of interest. Such information is produced by specific modelling tools and it is associated with some uncertainty, especially when predictors produce contradicting results. As a consequence it needs to be very carefully examined and tagged in a database. Although addition of modelling results could be considered for a general database such as AmyLoad, this would require an immense amount of work

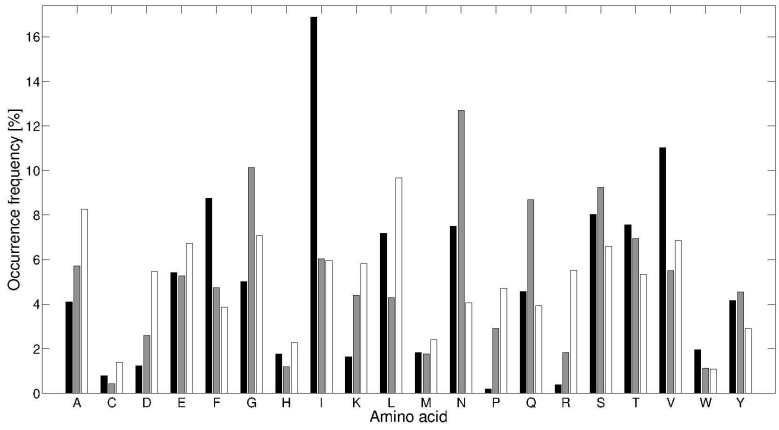

Fig. 3. Amino acid contents in hexapeptides collected in AmyLoad. Black bars denote amyloid fragments, grey non-amyloid, white bars statistical frequency as in Uniprot database

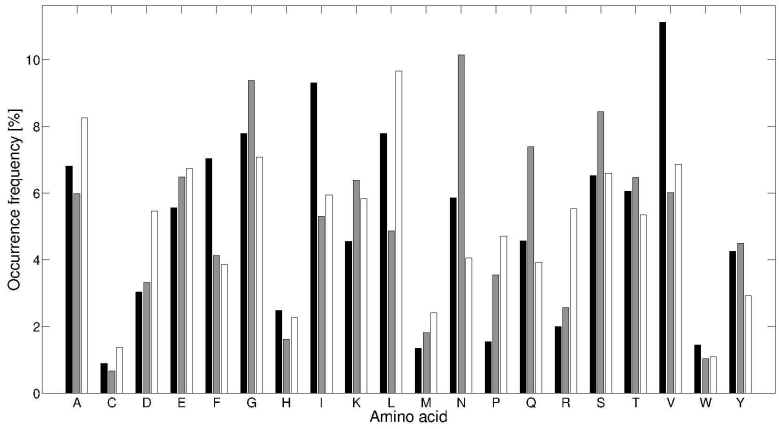

Fig. 4. Amino acid contents in all fragments collected in AmyLoad. Black bars denote amyloid fragments, grey non-amyloid, white bars statistical frequency as in Uniprot database

of the database curators to add these results with regard to every protein potentially aggregating, taking into account all available modelling tools. We believe that general databases should include less information, i.e. fewer fields, but from more reliable sources. Therefore, we decided to include only experimental data into AmyLoad and extend it into related subdatabases, dedicated to specific protein families that may be of research interest with regard to certain diseases. The first such database will be devoted to the $\beta\gamma$-crystallins and their involvement in cataract development. One of the specificities of these proteins is their aggregation pathway: many crystallins that are involved in the development of cataract can form amorphous aggregates rather than amyloids, others follow an amyloid aggregation path, whereas some may be involved in both paths. The sub-database of crystallins will separate and extend the number of their current entries in AmyLoad, while increasing the amount of available information about each of them.

3 Conclusions

We reported an internet database system dedicated to aggregating peptides, which may underlie several diseases of civilization, such as neurodegenerative diseases, diabetes type 2, and cataract. The general peptide database AmyLoad contains all currently known sequences of aminoacids whose propensity to amyloid aggregation has been published, based on experimental results. The entries also include some more specific information regarding each record.

Analysis of the amyloidogenic peptides, collected in the database showed a strong excess of neutral non-polar aminoacids with high propensity to for beta strands, such as valine, isoleucine, leucine and phenyloalanine. It appeared in the instances of all lengths. Interestingly, non-amyloidogenic peptides are characterized with increased contents of aminoacids with a positive charge and to lesser extent of negative charge. Since the enrichement of peptides with charged aminoacids impacts on non-amyloidogenic properties of peptides, it may be considered for applications towards changing the amyloidogenic properties of peptides.

AmyLoad may be too general for a specific family of proteins related to diseases in which aggregation may assume different forms, or modelling results would be crucial for further studies. Hence, more specialized sub-databases of different fields and including modelling results are required, such as the one containing $\beta\gamma$-crystallin proteins - underlying various forms of cataract. The sub-database should allow more detailed information than AmyLoad, focused more on different aggregate structures, leading to different disease phenotypes and different potential treatments for which pharmacophores could be designed based on available data.

Databases of aggregating proteins are needed to support further research in related diseases. We believe that our freely available internet service will facilitate the identification of the link between a protein sequence, its propensity to aggregation and the resulting disease, and discovering molecular events behind

development of diseases related to protein aggregation, as well as their pharmacological treatment and prevention.

The AmyLoad database, as well as other tools, could be found at Comprec server: `http://comprec-lin.iiar.pwr.edu.pl/amyload/`

Acknowledgements This work was in part supported by the grant N N519 643540 from National Science Centre of Poland.

References

1. Garbuzynskiy, SO., Lobanov, MY., Galzitskaya, OV. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Bioinformatics 26(3):326–32 (2010).
2. Goldschmidt, L., Tenga, PK., Riek, R., Eisenberg, D.: Identifying the amylome, proteins capable of forming amyloid-like fibrils. PNAS 107:3487–3492 (2010).
3. Bryan, AW Jr., O'Donnell, CW., Menke, M., Cowen, LJ., Lindquist, S., Berger, B.: STITCHER: Dynamic assembly of likely amyloid and prion -structures from secondary structure predictions. Proteins 80:410–420 (2011).
4. O'Donnell, CW., Waldispühl, J., Lis, M., Halfmann, R., Devadas, S., Lindquist, S., Berger, B.: A method for probing the mutational landscape of amyloid structure. Bioinformatics 27:i34–42 (2011).
5. Beerten, J., Van Durme, J., Gallardo, R., Capriotti, E., Serpell, L., Rousseau, F., Schymkowitz, J.: WALTZ-DB: a benchmark database of amyloidogenic hexapeptides. Bioinformatics 31(10):1698–700 (2015).
6. Stanislawski, J., Kotulska, M., Unold, O.: Machine learning methods can replace 3D profile method in classification of amyloidogenic hexapeptides. BMC Bioinformatics 14:21 (2013).
7. Gasior, P., Kotulska, M.: FISH Amyloid - a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of aminoacids. BMC Bioinformatics 15:54 (2014).
8. Zambrano, R., Jamroz, M., Szczasiuk, A., Pujols, J., Kmiecik, S., Ventura, S.: AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. Nucleic Acids Research (2015), doi: 10.1093/nar/gkv359
9. Wozniak, PP., Kotulska, M.: AmyLoad - website dedicated to amyloidogenic protein fragments. Bioinformatics (2015), doi: 10.1093/bioinformatics/btv375. `http://comprec-lin.iiar.pwr.edu.pl/amyload/`
10. Thompson, MJ., Sievers, SA., Karanicolas, J., Ivanova, MI., Baker, D., Eisenberg, D.: The 3D profile method for identifying fibril-forming segments of proteins. Proc. Natl. Acad. Sci. U.S.A. 103(11), 4074–8 (2006).
11. 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. 22(10), 1302–6 (2004).
12. Conchillo-Sol, O., de Groot, NS., Avils, FX., Vendrell, J., Daura, X., Ventura, S.: AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinformatics, 8:65 (2007).