Taylor & Francis
Taylor & Francis Group

Amyloid
The Journal of Protein Folding Disorders

ISSN: 1350-6129 (Print) 1744-2818 (Online) Journal homepage: http://www.tandfonline.com/loi/iamy20

Mining databases for protein aggregation: a review

Paraskevi L. Tsiolaki , Katerina C. Nastou , Stavros J. Hamodrakas & Vassiliki A. Iconomidou

To cite this article: Paraskevi L. Tsiolaki , Katerina C. Nastou , Stavros J. Hamodrakas & Vassiliki A. Iconomidou (2017): Mining databases for protein aggregation: a review, Amyloid, DOI: 10.1080/13506129.2017.1353966

To link to this article: http://dx.doi.org/10.1080/13506129.2017.1353966

Published online: 18 Jul 2017.

AMYLOID, 2017
https://doi.org/10.1080/13506129.2017.1353966

Taylor & Francis
Taylor & Francis Group

REVIEW ARTICLE

Mining databases for protein aggregation: a review

Paraskevi L. Tsiolaki*, Katerina C. Nastou*, Stavros J. Hamodrakas and Vassiliki A. Iconomidou

Section of Cell Biology and Biophysics, Department of Biology, School of Sciences, National and Kapodistrian University of Athens, Athens, Greece

ABSTRACT

Protein aggregation is an active area of research in recent decades, since it is the most common and troubling indication of protein instability. Understanding the mechanisms governing protein aggregation and amyloidogenesis is a key component to the aetiology and pathogenesis of many devastating disorders, including Alzheimer's disease or type 2 diabetes. Protein aggregation data are currently found "scattered" in an increasing number of repositories, since advances in computational biology greatly influence this field of research. This review exploits the various resources of aggregation data and attempts to distinguish and analyze the biological knowledge they contain, by introducing protein-based, fragment-based and disease-based repositories, related to aggregation. In order to gain a broad overview of the available repositories, a novel comprehensive network maps and visualizes the current association between aggregation databases and other important databases and/or tools and discusses the beneficial role of community annotation. The need for unification of aggregation databases in a common platform is also addressed.

Abbreviations: AD: Alzheimer's disease; AL: amyloidosis light chain amyloidosis; FTD: frontotemporal dementias; PD: Parkinson's disease

ARTICLE HISTORY
Received 2 February 2017
Revised 6 July 2017
Accepted 7 July 2017

KEYWORDS
Protein aggregation; amyloid; amyloidosis; amyloidogenesis; database

Introduction

From finding novel anti-amyloid drugs, to engineering solid biomaterials or designing successful biopharmaceutical products, a broad spectrum of questions remain to be answered, regarding the principles governing protein aggregation or amyloidogenicity [1–5]. The abnormal deposition of protein aggregates results in the disruption of the normal function of various tissues and organs and causes the onset and progression of a wide spectrum of diseases [6–8]. While the causal role of aggregation in disorders has not been established [9], the widespread phenomenon of protein aggregation interferes even with biotechnological routines [10]. Importantly, the same disease-related proteins (or peptides) may assemble to form amyloid fibrils *in vitro* [11] or during their physiological roles [12–14].

Protein aggregation indicates the biological processes by which stable complexes stem from an abnormal protein assembly. Under destabilizing conditions, proteins, unable to gain their proper native three-dimensional structure, are led to the formation of insoluble misfolded protein aggregates [15]. Aggregates are recorded as either "amorphous" or "ordered amyloid aggregates". Namely, amyloid fibrils refer to highly ordered protein aggregates characterized by typical cross-β diffraction patterns [16,17], unique morphologies [18] and specific tinctorial properties [19]. Their remarkable stability is the result of a rigid core structure, enhanced by cooperative hydrogen bonding [20].

The genesis of highly ordered amyloid aggregates is reported as amyloidogenesis or amyloidogenicity, while the underling mechanisms of misfolding have long been studied with experimental [21], computational [22] and theoretical methods [23]. A wealth of detailed information about the structural conversion from a soluble protein into elongated β-sheet protein aggregates introduced different theories of protein fibrillation [24,25]. Putative models propose destabilization and unfolding of the native structure, triggering a cascade of time-dependent events where protected "aggregation-prone" regions become exposed and nucleate the aggregation process [26–28]. Other experimental kinetic studies revealed a mechanism analogous with the process of crystal growth, since intermediate species gradually polymerize to form oligomers, substructures such as filaments or protofibrils and subsequent mature amyloid fibrils [29].

Currently, a vast amount of information about protein aggregation and amyloidogenicity is deposited in separate online collections, but has not received yet a lot of attention in the field of protein aggregation research. Complete and up-to-date databases are essential for the biological and medicinal research [30,31]. In abstracted terms, biological databases embody repositories of organized biological knowledge, which are invaluable for the researchers, who attempt to quickly find the appropriate resources for resolving their research questions. The availability of specialized resources has changed the way biologists study the phenomenon of

CONTACT Vassiliki A. Iconomidou veconom@biol.uoa.gr Section of Cell Biology and Biophysics, Department of Biology, School of Sciences, National and Kapodistrian University of Athens, Athens 15701, Greece
*These authors contributed equally.

protein misfolding, by broadening the focus from a single protein assay to the systematic analysis of aggregation.

In this review, we outline the repositories that are relevant with this field of research, by mining all freely and publicly available datasets and databases of amyloidogenic proteins, "aggregation-prone" peptides, aggregation disorders and amyloidoses. We also elucidate the interconnections between separate databases, by creating a network, defined by our own classification scheme, and discuss the challenges for developing a new and unified database of protein aggregation.

Protein aggregation databases

The scattered knowledge of protein aggregation data designates the outstanding need for introducing protein aggregation databases to the research community.

The need for protein aggregation databases

The phenomenon of protein aggregation and the subsequent formation of highly ordered aggregates attracted considerable attention and became an area of intense research [32]. Computational work over the past 15 years or so, assisted in the development of various tools, in an effort to predict and analyze the aggregation profile of proteins [33–35]. Since *in silico* methods were at the time based on selected data from random molecular databases or the ample existing literature, an imperative need for aggregation-related data to be organized, accessible, and easily retrieved, arises. Hence, the availability of data stored in ideally designed protein aggregation databases changed the way researchers study these aggregation phenomena and contributed to significant advances in biological, medical and – even – clinical processes.

Universal protein databases, such as UniProtKB [36], contain high-quality computationally analyzed records, enriched with automatic annotation and classification, but the extraction of information can be complex due to the large amount of data and the diversity of protein families. Commonly, fully annotated entries in general databases are tagged with controlled "vocabulary keywords" that can be used to retrieve particular subsets of entries [37]. However, advanced computational techniques should accompany this search procedure, in order to track down entries related, for example, to "amyloid formation" or "amyloidosis". Therefore, it emerges the essential need for specialized databases dedicated to protein aggregation.

Classification of protein aggregation repositories

A list of protein aggregation repositories was prepared and is presented in Table 1 and Table 2. Available details on repositories with experimental conditions are gathered in Table 3. Figure 1 maps all the interconnections between protein aggregation databases, analyzed and discussed below.

Amyloidogenic datasets

Amyloidogenic datasets were the primary repositories of protein aggregation data, originally helping to facilitate the development, validation or testing of new algorithms. Nowadays, with the avalanche of genomic and proteomic sequences generated in the postgenomic era, many of these datasets are deposited in various reference databases [38]. However, a vast amount of data can only be obtained through literature searches. Besides their exploitation in experimental research, collecting these data contributes to the study of amyloidogenic or "aggregation-prone" proteins, since they can act as scaffolds for developing powerful and efficient computational methods.

An early example of a collection of amyloidogenic peptides was designed by Lopez de la Paz and Serrano in 2003 [22], by systematically replacing residues of the *de novo* designed amyloidogenic peptide STVIIE, with all known natural amino acids, in an attempt to elucidate the principles of amyloidogenicity. This dataset has proved extremely important for experimental researchers, a fact established from almost 274 citations. Two additional, individual sets of amyloidogenic peptides, AmylHex and AmylFrag, were gathered carefully to evaluate the 3D profile method [39]. The first dataset included peptides from the aforementioned Lopez de la Paz dataset, together with protein fragments derived from insulin, β_2-microglobulin, amylin (also called IAPP) and tau, whereas the second exploited experimentally verified amyloidogenic peptides, identified by various researchers in the literature [40]. Another example, AmyloBase [41], a currently

Table 1. Amyloidogenic datasets of protein aggregation.

Dataset	Dataset Description	Reference
Lopez de la Paz Collection	Variations – mutants of the STVIEE peptide	[22]
AmylHex	158 peptides; 67 amyloid-forming and 91 non-amyloid-forming peptides experimentally verified	[39]
AmylFrag	45 amyloidogenic fragments of proteins experimentally verified	[39]
AmyloBase	Experimentally determined kinetic data about oligomer or amyloid fibril formation	[41]
Hexpepset	2452 hexpeptides; 1226 amyloid-forming and 1226 non-amyloid-forming	[42]
TANGO Dataset	71 peptides derived from human disease-related proteins (prion protein, lysozyme and β_2-microglobulin)	[65]
AGGRESCAN Dataset	160 natively globular proteins; 51 intrinsically disordered proteins; 38 soluble proteins, overexpressed in bacteria; 121 proteins forming inclusion bodies, overexpressed in bacteria; 57 amyloidogenic proteins	[44]
AMYLPRED Dataset	12 amyloidogenic proteins with know 3D structures, exposing experimentally or computationally verified peptides	[27,45]

The majority of the datasets play a key role in the development, validation or testing of the reported algorithms, while others published independently as large-scale proteomic analyses. A brief description for each dataset is provided.

unavailable collection, was described as a nascent list of experimentally determined data, regarding oligomer or amyloid formation, which was created to support the development of future amyloid prediction methods. It should also be mentioned that datasets based on various rational concepts were created for many remarkable aggregation prediction methods (Pafig [42], TANGO [43], AGGRESCAN [44], AMYLPRED [45] and AMYLPRED2 [27]). It should be noted that, although datasets are usually hidden components of computational methods, they can prove to be extremely valuable assets to researchers. Well-known amyloidogenic datasets are listed in Table 1.

Protein-based databases

Protein-based libraries of aggregation and amyloidogenicity, introduced soon after the onset of protein aggregation research, are consistent protein repositories that improved the data access to specialized users.

Although amyloid fibrils from different proteins share common ultrastructures, the amino acid composition and native structures of the proteins, associated with amyloid diseases, have been found to be highly variable [46]. A significant first attempt to accumulate information of amyloidogenic proteins was the fibril_one online database, published back in 2002 [47]. The main idea behind the

Table 2. Classification of protein aggregation databases.

Category	Database	Database Contents	URL	Availability	User Annotation	Reference
Protein-based	fibril_one	250 mutations and experimental conditions, diseases associated with 22 amyloidogenic proteins	http://www.bioinformatics.leeds.ac.uk/group/online/fibril_one	N/A	N	[47]
	ZipperDB	Amyloidogenic profiles of proteins in 76 different genomes	https://services.mbi.ucla.edu/zipperdb/	A	N	[40]
	AMYPdb	Amyloid precursor families of 600 organisms and their amino acid sequence signatures	http://amypdb.genouest.org/e107_plugins/amypdb_project/project.php	A	Y	[49]
	CreateFibril	Three-dimensional fibril models of HETs, Aβ and amylin	http://amyloid.cs.mcgill.ca/database/index.html	A	N	[51]
Fragment-based	WALTZ-DB	Experimentally characterized amyloid-forming and non-amyloid-forming hexapeptides	http://waltzdb.switchlab.org/	A	N	[58]
	CPAD	Peptides related to amorphous or amyloid aggregation	http://www.iitm.ac.in/bioinfo/CPAD/	A	Y	[61]
	AmyLoad	Amyloidogenic and non-amyloidogenic protein fragments, experimentally or computationally characterized	http://comprec-lin.iiar.pwr.edu.pl/amyload	A	Y	[63]
Disease-based	Mutations in Hereditary Amyloidosis	Mutations in hereditary amyloidoses related with 8 amyloidogenic proteins	www.amyloidosismutations.com	A	Y	[67]
	ALBase	Nucleotide and amino acid sequences of immunoglobulin light chains of patients with AL amyloidosis	http://albase.bumc.bu.edu/aldb/	A	N	[71]
	AlzGene	Genetic association studies in the field of Alzheimer's disease	http://www.alzgene.org/	A	N	[72]
	PDGene	Genetic association studies in the field of Parkinson's disease	http://www.pdgene.org/	A	N	[73]
	PDbase	Gene and genetic variations of Parkinson's disease	http://bioportal.kobic.re.kr/PDbase/	N/A	N	[74]
	AD&FTDMD	Mutations of the Alzheimer's disease and frontotemporal disorders, reported in the literature, directly submitted or communicated at scientific meetings	http://www.molgen.ua.ac.be/admutations/	A	Y	[75]
	Amyloidosis Foundation	Informal data on selected forms of amyloidoses	http://www.amyloidosis.org/	A	N	–
	ProADD	Information for 600 proteins involved in 12 aggregation diseases	http://bicmku.in/ProADD	A	N	[76]

Databases are categorized as protein-based repositories, fragment-based repositories and disease-based repositories. The contents, the availability (N/A: not available; A: available) and the corresponding URL of each database are provided. User annotation availability is also included (Y: yes or N: no).

Table 3. Contents of aggregation datasets and databases with *in vitro* experimentally verified data.

	Features		
Aggregation Repositories	Methodology	Additional Data	Availability
Lopez de la Paz Collection	Transmission electron microscopy, Fourier transform infrared spectroscopy, X-ray fibre diffraction, circular dichroism	Buffer name and concentration, peptide concentration, ionic strength, pH, temperature	A
AmyloBase	–	Kinetic details (seconds of lag phase, seconds of $t^{1/2}$, exp. value), mutation type, pH, temperature, ionic strength, solvent additives, solvent cofactors, protein/peptide concentration	N/A
fibril_one	–	–	N/A
WALTZ-DB	Transmission electron microscopy, Fourier transform infrared spectroscopy, circular dichroism, ProteoStat dye	–	A
CPAD	Fluorescence, light scattering, immunofluorescence, turbidity measurements, thioflavin T dye, Congo red dye, thioflavin S dye	Mutation type, temperature range, pH range, buffer name and concentration, ionic name and strength, protein/peptide concentration, solvent additives	A
AmyLoad	Transmission electron microscopy, X-ray fibre diffraction, Fourier transform infrared spectroscopy, circular dichroism, ProteoStat dye, Congo red dye	Peptide concentration, pH, ionic strength	A

Features of experimental-derived repositories of protein aggregation are listed below. The availability of each repository is provided (N/A: not available; A: available).

creation of this collection was based on how different mutations affect fibrillogenesis of known amyloidogenic proteins. Nearly 250 mutations and 50 experimental conditions associated with 22 proteins constitute fibril_one, a database riddled with information, yet currently unavailable to the public.

The concept of deciphering the factors responsible to drive a protein into the amyloid state allowed Goldschmidt et al. in 2009 to introduce the amylome, the "universe" of amyloidogenic proteins, by predicting the aggregation profile of every protein in 76 genomes. Notably, following UniProt's footsteps, ZipperDB [40] was created to store a large-scale amyloidogenic analysis on different genomes. All database entries are proteins segregated in computationally predicted, overlapping peptide segments. Structural modelling was used in order to evaluate the possibility that a particular protein stretch can form a tightly packed interface with an identical stretch and participate in the formation of a steric zipper [48]. This computationally derived repository is a unique approach, which uses solely structural information to predict protein stretches, "prone" to give rise to an amyloid spine. The user can browse the database for each genome or submit queries using a protein's name, a peptide sequence or a genome's name. Nonetheless, an account is necessary to perform any of the prior actions.

On the contrary, AMYPdb [49], a database of 31 selected amyloid precursor families, serves the purpose of *in silico* collecting all amyloidogenic peptide signatures among nearly 600 organisms. The idea that the propensity of a protein to aggregate into typical amyloid fibrils varies greatly, depending on the amino acid sequence or the cellular environment, enabled the development of a database, based on known PROSITE protein patterns [50]. Each precursor entry in AMYPdb is thoroughly analyzed and annotated, including a brief description, information about the primary and the secondary structure and finally, a list of sequences matching the PROSITE pattern of the precursor protein. Among the available functions, an advanced search engine allows the complex navigation on the stored data, while implemented methods enable the calculation of several physicochemical parameters or the amyloidogenic profile of a query protein. The main advantage of AMYPdb is the benefit to create customized working sets of proteins, authorizing the interactive operation of the user. This repository is eventually a computational effort to extract accurate data of potential amyloidogenic proteins among different organisms.

Meanwhile, the prominence of amyloid deposits in many diseases raised the question of how misfolded proteins interact at atomic level, and thus, much effort has been spent on elucidating the three-dimensional structure of amyloid fibrils. Amyloid assemblies can be extremely diverse as a result of amyloid polymorphism, a common phenomenon even for fibrils formed from an identical sequence. An extremely specialized database of classified fibrillar shapes and hyperstructures of proteins well known to self-assemble into amyloid fibrils was constructed by Smaoui et al. [51]. CreateFibril database is basically a computationally calculated collection of stable polymorphic fibril models of HETs, Aβ and amylin proteins, along with their structural energy landscapes, produced by the CreateFibril tool [51].

Fragment-based databases

It has been generally established that short protein segments can nucleate the fibrillation of polypeptide chains, since they intrinsically exhibit the tendency to drive a native protein to the amyloid state [52]. The moment this observation emerged, the research interest redirected from the study of native proteins into short protein stretches [53–56]. Computational approaches for investigating the aggregation profile of proteins and detecting "aggregation-prone" segments in proteins have been explicitly reviewed [33–35,57]. Several biological databases are dedicated to such short

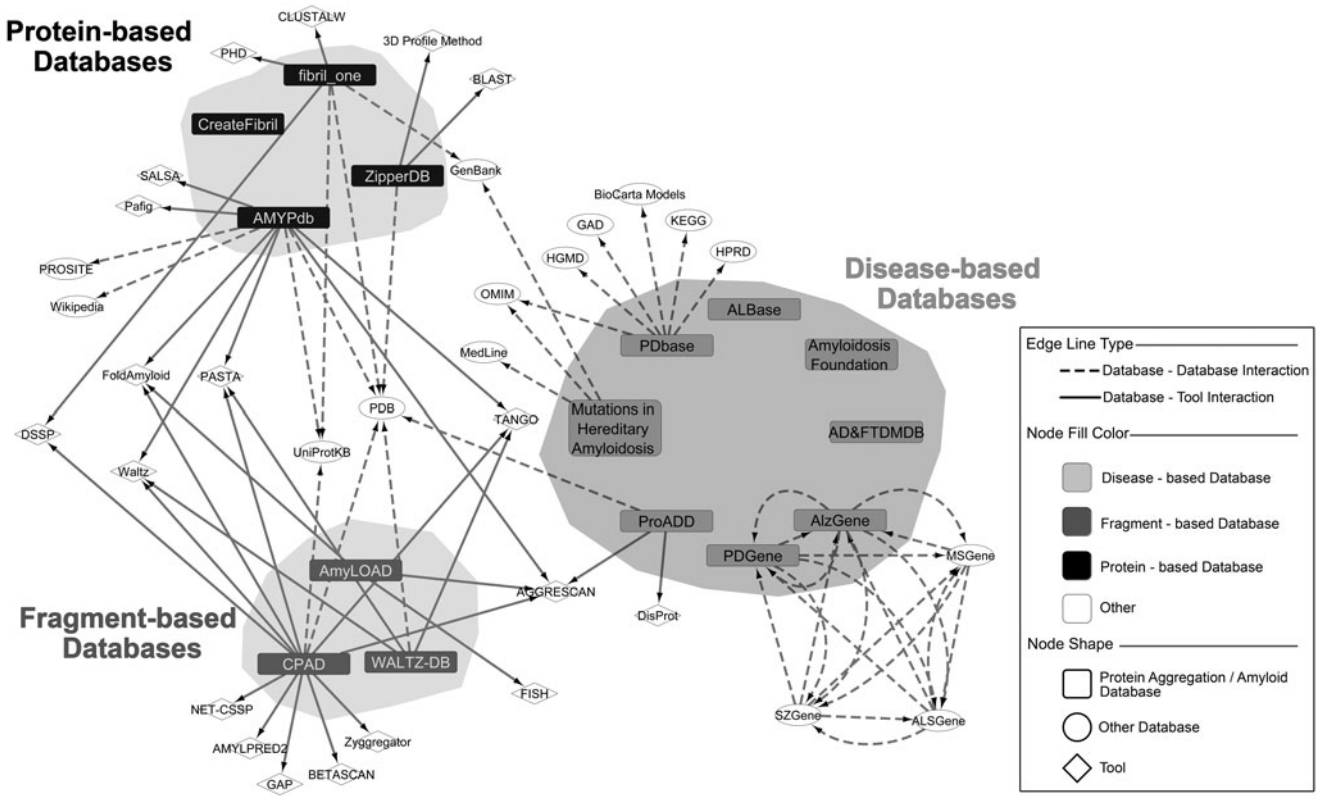

Figure 1. A directed network representing interconnections between protein aggregation databases with other databases and computational tools. Each database or computational tool is represented as a node, whereas interactions between neighbouring nodes are visualized as edges. Incoming and outcoming edges are used to describe the directed flow of information within the network. Protein aggregation databases are sorted in three distinct categories, as described in Classification of Protein Aggregation Repositories Section. Specifically, databases dedicated to amyloid disorders and other aggregation diseases are shown in light grey (group on the right), databases that provide information regarding "aggregation-prone" segments in protein sequences in dark grey (group on the bottom) and databases that contain amyloidogenic proteins in black (group on the top). The node shape is used to discern the different components of the network. Protein aggregation databases are depicted as rectangles (■), whereas other databases are depicted as circles (●). Noteworthy, prediction algorithms, presented as diamonds (◆), are computational tools utilized or implemented explicitly in the creation of aggregation databases and can be either relevant to protein aggregation or provide other important utilities for the repositories in which they are found, despite not being aggregation specific. Aggregation prediction algorithms presented in this figure are extensively reviewed in other works [33–35]. Dashed lines reveal the connections between databases, while solid lines are used to depict the connections between aggregation databases and computational tools. Two elements are considered connected when a hyperlink exists from the source element to the target. In particular, dashed lines connect aggregation databases with other databases and solid lines connect aggregation databases with computational tools. The visualization of the protein aggregation databases network was executed using the open source program Cytoscape [77]. Databases and tools that are not mentioned in the main text are referenced in Table 4.

segments; we describe these fragment-based repositories thoroughly below and in Table 2.

WALTZ-DB [58] is the first elaborate collection of amyloidogenic and non-amyloidogenic hexapeptides, inspired and created by the developers of Waltz algorithm [59]. This database assembles experimentally characterized peptides in one extended and detailed collection. The collection contains hexapeptides derived from different sources, yet annotated in detail, regarding their amyloidogenic capacities. The available experimental data, together with the structural characterization following Eisenberg's classification [60], are a useful addition for peptide specialists. The accessibility on experimental details and/or experimental material, accompanying each peptide entry, allows researchers to independently evaluate the quality of data (Table 3). Among the advantages of WALTZ-DB is the choice of the user to filter information, by using the search engine and particularizing their query.

Moving a step further, Thangakani et al., recognizing the expanding interest of researchers in the field of protein aggregation, developed the Curated Protein Aggregation Database (CPAD) [61]. This database ventured to assemble current information on peptide aggregation and thus is a collection of peptides that are related either to fibril formation or amorphous aggregation. Particularly, an accurate distinction between amyloid-forming peptides of different lengths and amorphous or non-amyloid-forming peptides is available, whereas additional structural information of peptides with known 3D structures is provided, with external links to PDB [62].

A great supplement in fragment-based databases is a repository aiming at bringing together amyloidogenic sequences from all major sources, in a common platform. AmyLoad consolidated a great majority of the currently available amyloidogenic and non-amyloidogenic sequences in an effort to collect all fragment data of amyloidogenesis [63]. Developers made a special effort to merge validation datasets of different prediction methods (AGGRESCAN [64], TANGO [65], AmylHex and AmylFrag [40]), along with data obtained from the literature. Entries are easily accessible and information is carefully organized, while interconnections with other available databases and

computational methods compose a user-friendly interface (see Database interconnections in Figure 1). For instance, for a given peptide, researchers can obtain in-depth information, such as the experimental preparation of the sample or the name of the disease related with each record. A major benefit for the research community is the availability of the experimental methodology, used for the verification of amyloidogenicity for each peptide entry (Table 3).

Disease-based databases

According to Sipe et al., amyloidogenic proteins can either give rise to distinct amyloidoses, or can play a pathological role in neurodegenerative or endocrine diseases that are actually not classified as amyloidosis, from a clinical viewpoint [8]. Amyloidoses constitute a remarkably heterogeneous group of diseases that have been studied extensively and undergone various classifications (localized or systemic, primary or secondary, mesenchymal or parenchymal) [66].

Hereditary amyloidosis is the type of disease, caused by inheriting a particular gene mutation. Mutations in Hereditary Amyloidosis database [67] merges the need of the clinical and scientific community for accessing information about new mutations and phenotypes in hereditary amyloidosis. The database contains information about genes, related to different forms of amyloidosis, and is enhanced with a brief description of known clinical phenotype or ethnicity, allowing a more complete understanding of the relationship between genes and the trait of a disease. External reference resources to Medline [68] and OMIM [69] assist in "spanning" clinical and available online knowledge on such amyloidoses.

A few databases have been designed to accumulate knowledge for specific disorders related to amyloid formation. Light chain amyloidosis (or AL amyloidosis) is the most common type of systemic amyloidosis, mainly associated with plasma cell dyscrasia [70]. The dominance of this systemic form in the developed world triggered the creation of the ALBase [71], a repository of the primary sequences of immunoglobulin light chains of patients with AL amyloidosis, containing 4364 nucleotide and amino acid sequences of immunoglobulin light chains. AlzGene [72], PDGene [73] and PDbase [74] are disease-based databases dedicated to the Alzheimer's disease (AD) and Parkinson's disease (PD). The aforementioned neurodegenerative diseases have high prevalence worldwide and are responsible for a great number of deaths globally in recent years. AlzGene and PDGene gather and comprehensively catalogue all genetic association studies in the field of Alzheimer's and Parkinson's disease, respectively, whereas PDbase present an in-depth collection of molecular characteristics, governing aggregation in Parkinson's disease. Another approach in Alzheimer's disease is available through the Alzheimer Disease & Frontotemporal Dementia Mutation Database (AD&FTDMD) [75], a platform that aims at collecting all known gene mutations related, in general, to the Alzheimer's disease and frontotemporal dementias (FTD). Finally, Amyloidosis Foundation [60], known mainly as a forum rather than a scientific database, contains informal recorded information on selected forms of amyloidoses and raises awareness for these rare forms of diseases to the general public.

While a range of overwhelming human diseases is known to be associated with the formation of highly organized protein aggregates, deposition of amorphous protein aggregates could also be related to the onset of other aggregation diseases. The tremendous impact of several amyloidoses worldwide, though, concealed the presence of other significant disorders, caused by protein aggregation. In contrast to the specialized amyloid databases above, ProADD holds details of around 600 proteins involved in aggregation diseases [76]. Proteins in ProADD are annotated and categorized as aggregating proteins or intrinsically disordered proteins, based on data gathered from literature or by utilizing available prediction algorithms. Unfortunately, the usefulness of this library is limited, as protein data are not currently available.

Interconnections between protein aggregation repositories

In order to further analyze the functional "interactions" between the above databases, we constructed a directed graph, utilizing Cytoscape [77]. A graph or network (interchangeably used terms) consists of a set of nodes, representing entities of interest, and a set of edges, signifying specific relationships between them. Namely, a directed graph has ordered pairs of edges with directions from a certain node to another and is mainly used to describe procedures with a specific flow of information.

Following a systematic approach, protein aggregation databases and their implemented or integrated tools were collected. Data were divided as protein aggregation/amyloid databases (rectangle node shape), other databases (circle node shape) and tools (diamond node shape). A number of well-known repositories and computational methods, not strictly related to protein aggregation, were intentionally gathered, so as to visualize the way repositories, dedicated to protein aggregation, implement other databases, methods or tools. Incoming and outcoming arrows were used to map the context-specific information propagation (Figure 1).

Figure 1, illustrating connections between repositories and computational tools, provides a frame for the interpretation of links among protein aggregation repositories. This network-based representation is evidently presenting that there are no direct links between any of the databases dedicated to protein aggregation. AlzGene and PDGene are the only exception, since they depict extreme interconnectivity with other databases, compiling data from genetic association studies [78,79]. Fragment-based databases (dark grey rectangles) are strongly interconnected with aggregation prediction tools (Waltz, FoldAmyloid, PASTA, TANGO, etc.), a reasonable finding as predicted aggregation rates are the basis of discovering short protein segments. Protein-based collections (black rectangles) implement only a few prediction methods; however, they form connections with major protein databases (UniProtKB, PDB). In turn, databases related to aggregation diseases (light grey rectangles) exhibit extremely low

Table 4. Detailed interconnections between aggregation databases and other databases or tools, derived from the analysis of the directed network (Figure 1).

Aggregation Database	Other Database	Computational Tool
fibril_one	PDB [62] UniProtKB [36] GenBank [80]	CLUSTALW [81] PHD [82] DSSP [83]
ZipperDB	PDB [62]	Blast [84] 3D Profile method [39]
AMYPdb	Wikipedia PROSITE [50] UniProtKB [36] PDB [62]	Salsa [85] Pafig [42] Fold Amyloid [86] Waltz [59] PASTA [87] TANGO [43] AGGRESCAN [64]
WALTZ-DB	PDB [62]	TANGO [43] PASTA [87] Waltz
CPAD	UniProtKB [36] PDB [62]	NET-CSSP [88] GAP [89] AMYLPRED2 [27] BETASCAN [90] Zyggregator [91] DSSP [83] Waltz [59] PASTA [87] TANGO [43] AGGRESCAN [64] FoldAmyloid [86]
AmyLoad		AGGRESCAN [64] FoldAmyloid [86] FISH Amyloid [92]
Mutations in Hereditary Amyloidosis	GenBank [80] MedLine [68] OMIM [69]	
AlzGene	PDGene [93] SZGene [78] MSGene ALSGene [93]	
PDGene	AlzGene [72] SZGene [78] MSGene ALSGene [93]	
PDbase	BioCarta Models [79] GAD [94] HGMD [95] HPRD [96] KEGG [97] OMIM [69]	
ProADD	PDB [62] DisProt [98]	AGGRESCAN [64]

Each aggregation database, highlighted in bold, occupies the first column and is related to other databases and computational tools, listed in the second and third column, respectively. Apparently, the association between aggregation databases is poor, with the exception of AlzGene and PDGene, since only two aggregation databases are included in the second column.

connectivity with the entire network, since only three out of seven collections have external links to other computational tools or databases. UniProtKB [36] and PDB [62], two well-annotated and comprehensive resources, are the databases with external links from the majority of protein and fragment-based databases, since three (3) and six (6) incoming edges were recorded, respectively. However, neither of these databases has explicit or implicit links to amyloid-related repositories (no outcoming edges) (Table 4).

The graphical visualization of the aforementioned databases (Figure 1, Table 4) emphasizes the necessity to concentrate additional knowledge regarding protein aggregation and consequently create connections between them and other online resources, in order to make data mining more efficient for the scientific community. Our graph is believed to contribute towards putting forward alternative strategies in merging data between different groups of repositories.

Moving to aggregation-related data, more rigorous studies should take place, since data include a considerable amount of redundancy. Although it is unlikely that one database will be the collection of choice in all circumstances, a major restriction for researchers is that only few repositories keep key details on experimental conditions of protein aggregation (Table 3). A problem that needs to be solved is the heterogeneity of the records, even within the same group of aggregation databases. A potential format standardization would sustain the data usefulness and would allow the effortless comparison among data from different sources.

Advancements in computationally based or experimental-based databases from various research groups will upgrade the quality of data and will accelerate protein aggregation research worldwide. It is also expected that aggregation repositories will be valuable "partners" for the experimental community, when they will be able to provide a common protein aggregation platform of non-redundant data in a uniform way.

User annotation of protein aggregation

A stimulating and a particularly useful feature available in some of the aforementioned repositories is the choice of the community annotation [49,61,63,67,75]. In general, the primary goal of a scientific database is the accurate and comprehensive demonstration of the data, along with the effortless accessibility for the scientific community. The ever-increasing demand of experimental data, though, necessitates the constant update of the available knowledge, a challenging attempt for specialized or "small-scale" repositories, such as databases dedicated to aggregation and amyloidogenicity. To address this issue, apart from the common database curation, the user annotation feature permits researchers to outsource their experimental or computational results and eventually supports the compilation of the available biological knowledge. Namely, five out of thirteen databases, reviewed herein, implement – to some extent – the user "interaction" and allow the submission of relevant data. The synergy between the database curators and the scientific community is a fruitful attempt to improve specialized repositories dedicated to protein aggregation. The user annotation availability for every database is presented in Table 2.

Conclusions

A field that is receiving renewed emphasis, both computationally and experimentally, is the elucidation of the molecular mechanisms that drive protein aggregation and amyloidogenesis. A variety of protein aggregation databases recently emerged, ranging from general collections of amyloidogenicity, to specialized disease databases, bearing various utilities. This review provides a brief analysis of available protein aggregation repositories that fulfil diverse needs and can be divided into four distinct categories, namely amyloidogenic datasets, protein-based repositories, databases dedicated to "aggregation-prone" fragments of proteins and, finally, disease-based collections (Table 1 and Table 2). A more clear representation of the association among the protein aggregation repositories is graphically displayed herein (Figure 1). Apparently, interconnections between various collections are scarce and thus, there are several limitations in the combined mining of the databases for information regarding protein aggregation (Table 4). Contrary to computational tools, databases require

continuous annotation and as a result, an existing problem is the need of incorporating protein aggregation data into ideally designed databases. The experimental and the computational community should universally join forces by exploiting user annotation, during the study of amyloidogenicity. It is hoped that, in years to come, databases containing protein aggregation data will undergo a unification and standardization, following the example of major biological databases such as UniProtKB, where protein sequence information, in its entirety, can be accessed through a single database.

Acknowledgements

The authors sincerely thank the Editor in Chief for properly handling this manuscript and both the Associate Editor and the anonymous reviewers for their very useful and constructive criticism, which helped to considerably improve the manuscript. We also thank the University of Athens for support.

Disclosure statement

The authors declare no conflict of interest

Funding

This project was financially supported by the Greek State Scholarships Foundation, through the program "Research Projects for Excellence IKY/Siemens" (2015–2017).

ORCID

Paraskevi L. Tsiolaki http://orcid.org/0000-0002-9525-6433
Katerina C. Nastou http://orcid.org/0000-0003-3611-5726
Stavros J. Hamodrakas http://orcid.org/0000-0001-6280-1645
Vassiliki A. Iconomidou http://orcid.org/0000-0002-9472-5146

References

[1] Aguzzi A, O'Connor T. Protein aggregation diseases: pathogenicity and therapeutic perspectives. Nat Rev Drug Discov. 2010;9:237–248.
[2] Eisele YS, Monteiro C, Fearns C, et al. Targeting protein aggregation for the treatment of degenerative diseases. Nat Rev Drug Discov. 2015;14:759–780.
[3] Frokjaer S, Otzen DE. Protein drug stability: a formulation challenge. Nat Rev Drug Discov. 2005;4:298–306.
[4] Mitraki A. Protein aggregation from inclusion bodies to amyloid and biomaterials. Adv Protein Chem Struct Biol. 2010;79:89–125.
[5] Cherny I, Gazit E. Amyloids: not only pathological agents but also ordered nanomaterials. Angew Chem Int Ed Engl. 2008;47:4062–4069.
[6] Sipe JD, Benson MD, Buxbaum JN, et al. Nomenclature 2014: amyloid fibril proteins and clinical classification of the amyloidosis. Amyloid. 2014;21:221–224.
[7] Soto C. Unfolding the role of protein misfolding in neurodegenerative diseases. Nat Rev Neurosci. 2003;4:49–60.
[8] Sipe JD, Benson MD, Buxbaum JN, et al. Amyloid fibril proteins and amyloidosis: chemical identification and clinical classification International Society of Amyloidosis 2016 Nomenclature Guidelines. Amyloid. 2016;23:209–213.
[9] Ross CA, Poirier MA. Opinion: what is the role of protein aggregation in neurodegeneration? Nat Rev Mol Cell Biol. 2005;6:891–898.
[10] Agrawal NJ, Kumar S, Wang X, et al. Aggregation in protein-based biotherapeutics: computational studies and tools to identify aggregation-prone regions. J Pharm Sci. 2011;100:5081–5095.
[11] Fandrich M. Structure and formation of amyloid fibrils. Acta Histochem. 2003;105:379.
[12] Hammer ND, Wang X, McGuffie BA, et al. Amyloids: friend or foe? J Alzheimers Dis. 2008;13:407–419.
[13] Iconomidou VA, Vriend G, Hamodrakas SJ. Amyloids protect the silkmoth oocyte and embryo. FEBS Lett. 2000;479:141–145.
[14] Iconomidou VA, Chryssikos GD, Gionis V, et al. Amyloid-like fibrils from an 18-residue peptide analogue of a part of the central domain of the B-family of silkmoth chorion proteins. FEBS Lett. 2001;499:268–273.
[15] Fandrich M. On the structural definition of amyloid fibrils and other polypeptide aggregates. Cell Mol Life Sci. 2007;64:2066–2078.
[16] Sunde M, Blake C. The structure of amyloid fibrils by electron microscopy and X-ray diffraction. Adv Protein Chem. 1997;50:123–159.
[17] Geddes AJ, Parker KD, Atkins ED, et al. "Cross-beta" conformation in proteins. J Mol Biol. 1968;32:343–358.
[18] Eichner T, Radford SE. A diversity of assembly mechanisms of a generic amyloid fold. Mol Cell. 2011;43:8–18.
[19] Ashkenazi Y, Hersko C, Gafni J, et al. Chemical aspects of amyloid-Congo red binding. Isr J Med Sci. 1967;3:572–574.
[20] Sunde M, Serpell LC, Bartlam M, et al. Common core structure of amyloid fibrils by synchrotron X-ray diffraction. J Mol Biol. 1997;273:729–739.
[21] 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;99:16419–16426.
[22] Lopez de la Paz M, Serrano L. Sequence determinants of amyloid fibril formation. Proc Natl Acad Sci USA. 2004;101:87–92.
[23] Knowles TP, Waudby CA, Devlin GL, et al. An analytical solution to the kinetics of breakable filament assembly. Science. 2009;326:1533–1537.
[24] Knowles TP, Vendruscolo M, Dobson CM. The amyloid state and its association with protein misfolding diseases. Nat Rev Mol Cell Biol. 2014;15:384–396.
[25] Kelly JW. Mechanisms of amyloidogenesis. Nat Struct Biol. 2000;7:824–826.
[26] Castillo V, Ventura S. Amyloidogenic regions and interaction surfaces overlap in globular proteins related to conformational diseases. PLoS Comput Biol. 2009;5:e1000476.
[27] Tsolis AC, Papandreou NC, Iconomidou VA, et al. A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins. PLoS One. 2013;8:e54175.
[28] Nelson R, Eisenberg D. Recent atomic models of amyloid fibril structure. Curr Opin Struct Biol. 2006;16:260–265.
[29] Chiti F, Dobson CM. Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem. 2006;75:333–366.
[30] Apweiler R, Bairoch A, Wu CH. Protein sequence databases. Curr Opin Chem Biol. 2004;8:76–80.
[31] Xenarios I, Eisenberg D. Protein interaction databases. Curr Opin Biotechnol. 2001;12:334–339.
[32] Chiti F, Webster P, Taddei N, et al. Designing conditions for *in vitro* formation of amyloid protofilaments and fibrils. Proc Natl Acad Sci USA. 1999;96:3590–3594.
[33] Hamodrakas SJ. Protein aggregation and amyloid fibril formation prediction software from primary sequence: towards controlling the formation of bacterial inclusion bodies. FEBS J. 2011;278:2428–2435.
[34] Ahmed AB, Kajava AV. Breaking the amyloidogenicity code: methods to predict amyloids from amino acid sequence. FEBS Lett. 2013;587:1089–1095.

[35] De Baets G, Schymkowitz J, Rousseau F. Predicting aggregation-prone sequences in proteins. Essays Biochem. 2014;56: 41–52.
[36] Boutet E, Lieberherr D, Tognolli M, et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. Methods Mol Biol. 2016;1374:23–54.
[37] UniProt C. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2007;35:D193–D197.
[38] Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
[39] Thompson MJ, Sievers SA, Karanicolas J, et al. The 3D profile method for identifying fibril-forming segments of proteins. Proc Natl Acad Sci USA. 2006;103:4074–4078.
[40] Goldschmidt L, Teng PK, Riek R, et al. Identifying the amylome, proteins capable of forming amyloid-like fibrils. Proc Natl Acad Sci USA. 2010;107:3487–3492.
[41] Belli M, Ramazzotti M, Chiti F. Prediction of amyloid aggregation *in vivo*. EMBO Rep. 2011;12:657–663.
[42] Tian J, Wu N, Guo J, et al. Prediction of amyloid fibril-forming segments based on a support vector machine. BMC Bioinformatics. 2009;10:S45.
[43] Linding R, Schymkowitz J, Rousseau F, et al. A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. J Mol Biol. 2004;342:345–353.
[44] de Groot NS, Castillo V, Grana-Montes R, et al. AGGRESCAN: method, application, and perspectives for drug design. Methods Mol Biol. 2012;819:199–220.
[45] Frousios KK, Iconomidou VA, Karletidi CM, et al. Amyloidogenic determinants are usually not buried. BMC Struct Biol. 2009;9:44.
[46] Citron M, Oltersdorf T, Haass C, et al. Mutation of the beta-amyloid precursor protein in familial Alzheimer's disease increases beta-protein production. Nature. 1992;360:672–674.
[47] Siepen JA, Westhead DR. The fibril_one on-line database: mutations, experimental conditions, and trends associated with amyloid fibril formation. Protein Sci. 2002;11:1862–1866.
[48] Sawaya MR, Sambashivan S, Nelson R, et al. Atomic structures of amyloid cross-beta spines reveal varied steric zippers. Nature. 2007;447:453–457.
[49] Pawlicki S, Le Bechec A, Delamarche C. AMYPdb: a database dedicated to amyloid precursor proteins. BMC Bioinformatics. 2008;9:273.
[50] Sigrist CJ, Cerutti L, Hulo N, et al. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinformatics. 2002;3:265–274.
[51] Smaoui MR, Poitevin F, Delarue M, et al. Computational assembly of polymorphic amyloid fibrils reveals stable aggregates. Biophys J. 2013;104:683–693.
[52] Teng PK, Eisenberg D. Short protein segments can drive a non-fibrillizing protein into the amyloid state. Protein Eng Des Sel. 2009;22:531–536.
[53] Tsiolaki PL, Louros NN, Hamodrakas SJ, et al. Exploring the 'aggregation-prone' core of human Cystatin C: a structural study. J Struct Biol. 2015;191:272–280.
[54] Louros NN, Tsiolaki PL, Zompra AA, et al. Structural studies and cytotoxicity assays of 'aggregation-prone' IAPP(8-16) and its non-amyloidogenic variants suggest its important role in fibrillogenesis and cytotoxicity of human amylin. Biopolymers. 2015;104:196–205.
[55] Louros NN, Tsiolaki PL, Griffin MD, et al. Chameleon 'aggregation-prone' segments of apoA-I: a model of amyloid fibrils formed in apoA-I amyloidosis. Int J Biol Macromol. 2015;79:711–718.
[56] Tsiolaki PL, Hamodrakas SJ, Iconomidou VA. The pentapeptide LQVVR plays a pivotal role in human cystatin C fibrillization. FEBS Lett. 2015;589:159–164.
[57] Trainor K, Broom A, Meiering EM. Exploring the relationships between protein sequence, structure and solubility. Curr Opin Struct Biol. 2017;42:136–146.
[58] Beerten J, Van Durme J, Gallardo R, et al. WALTZ-DB: a benchmark database of amyloidogenic hexapeptides. Bioinformatics. 2015;31:1698–1700.
[59] Maurer-Stroh S, Debulpaep M, Kuemmerer N, et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Meth. 2010;7:237–242.
[60] Eisenberg D, Jucker M. The amyloid state of proteins in human diseases. Cell. 2012;148:1188–1203.
[61] Thangakani AM, Nagarajan R, Kumar S, et al. CPAD, curated protein aggregation database: a repository of manually curated experimental data on protein and peptide aggregation. PLoS One. 2016;11:e0152949.
[62] Rose PW, Prlic A, Altunkaya A, et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. 2017;45:D271–D281.
[63] Wozniak PP, Kotulska M. AmyLoad: website dedicated to amyloidogenic protein fragments. Bioinformatics. 2015;31: 3395–3397.
[64] 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.
[65] Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, et al. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol. 2004;22: 1302–1306.
[66] Pepys MB. Amyloidosis. Annu Rev Med. 2006;57:223–241.
[67] Rowczenio DM, Noor I, Gillmore JD, et al. Online registry for mutations in hereditary amyloidosis including nomenclature recommendations. Hum Mutat. 2014;35:E2403–24E2412.
[68] Wood EH. MEDLINE: the options for health professionals. J Am Med Inform Assoc. 1994;1:372–380.
[69] Hamosh A, Scott AF, Amberger JS, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33:D514–D517.
[70] Sanchorawala V. Light-chain (AL) amyloidosis: diagnosis and treatment. Clin J Am Soc Nephrol. 2006;1:1331–1341.
[71] Bodi K, Prokaeva T, Spencer B, et al. AL-Base: a visual platform analysis tool for the study of amyloidogenic immunoglobulin light chain sequences. Amyloid. 2009;16:1–8.
[72] Bertram L, McQueen MB, Mullin K, et al. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet. 2007;39:17–23.
[73] Lill CM, Roehr JT, McQueen MB, et al. Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: the PDGene database. PLoS Genet. 2012;8: e1002548.
[74] Yang JO, Kim WY, Jeong SY, et al. PDbase: a database of Parkinson's disease-related genes and genetic variation using substantia nigra ESTs. BMC Genomics. 2009;10:S32.
[75] Cruts M, Theuns J, Van Broeckhoven C. Locus-specific mutation databases for neurodegenerative brain diseases. Hum Mutat. 2012;33:1340–1344.
[76] Shobana R, Pandaranayaka EP. ProADD: a database on protein aggregation diseases. Bioinformation. 2014;10:390–392.
[77] Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–2504.
[78] Allen NC, Bagade S, McQueen MB, et al. Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. Nat Genet. 2008;40:827–834.
[79] Buchel F, Rodriguez N, Swainston N, et al. Path2Models: large-scale generation of computational models from biochemical pathway maps. BMC Syst Biol. 2013;7:116
[80] Benson DA, Karsch-Mizrachi I, Lipman DJ, et al. GenBank. Nucleic Acids Res. 2000;28:15–18.

[81] Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23:2947–2948.
[82] Rost B. Review: protein secondary structure prediction continues to rise. J Struct Biol. 2001;134:204–218.
[83] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–2637.
[84] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol. 1990;215:403–410.
[85] Zibaee S, Makin OS, Goedert M, et al. 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. 2007;16:906–918.
[86] Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Bioinformatics. 2010;26:326–332.
[87] Walsh I, Seno F, Tosatto SC, et al. PASTA 2.0: an improved server for protein aggregation prediction. Nucleic Acids Res. 2014;42:W301–W307.
[88] Kim C, Choi J, Lee SJ, et al. NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation. Nucleic Acids Res. 2009;37:W469–W473.
[89] Thangakani AM, Kumar S, Nagarajan R, et al. GAP: towards almost 100 percent prediction for beta-strand-mediated aggregating peptides with distinct morphologies. Bioinformatics. 2014;30:1983–1990.
[90] Bryan AW Jr, Menke M, Cowen LJ, et al. BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. PLoS Comput Biol. 2009;5:e1000333.
[91] Tartaglia GG, Vendruscolo M. The Zyggregator method for predicting protein aggregation propensities. Chem Soc Rev. 2008;37:1395–1401.
[92] 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. 2014;15:54.
[93] Lill CM, Abel O, Bertram L, et al. Keeping up with genetic discoveries in amyotrophic lateral sclerosis: the ALSoD and ALSGene databases. Amyotroph Lateral Scler. 2011;12:238–249.
[94] Becker KG, Barnes KC, Bright TJ, et al. The genetic association database. Nat Genet. 2004;36:431–432.
[95] Stenson PD, Ball EV, Mort M, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003;21:577–581.
[96] Keshava Prasad TS, Goel R, Kandasamy K, et al. Human Protein Reference Database—2009 update. Nucleic Acids Res. 2009;37:D767–D772.
[97] Kanehisa M, Araki M, Goto S, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008;36:D480–D484.
[98] Vucetic S, Obradovic Z, Vacic V, et al. DisProt: a database of protein disorder. Bioinformatics. 2005;21:137–140.