FEBS openbio

FEBS PRESS science publishing by scientists

METHOD

PDB_Amyloid: an extended live amyloid structure list from the PDB

Kristóf Takács[1], Bálint Varga[1] and Vince Grolmusz[1,2]

1 PIT Bioinformatics Group, Eötvös University, Budapest, Hungary
2 Uratim Ltd., Budapest, Hungary

Keywords

amyloid; amyloid-precursor; human chorionic gonadotropin; PDB; thymidylate synthase; web server

Correspondence

V. Grolmusz, PIT Bioinformatics Group,
Eötvös University, H-1117 Budapest,
Hungary
E-mail: grolmusz@pitgroup.org

(Received 24 May 2018, revised 21 August 2018, accepted 3 September 2018)

doi:10.1002/2211-5463.12524

The Protein Data Bank (PDB) contains more than 135 000 entries at present. From these, relatively few amyloid structures can be identified, since amyloids are insoluble in water. Therefore, most amyloid structures deposited in the PDB are in the form of solid state NMR data. Based on the geometric analysis of these deposited structures, we have prepared an automatically updated web server, which generates a list of the deposited amyloid structures, and also entries of globular proteins that have amyloid-like substructures of given size and characteristics. We have found that by applying only appropriately selected geometric conditions, it is possible to identify deposited amyloid structures and a number of globular proteins with amyloid-like substructures. We have analyzed these globular proteins and have found proof in the literature that many of them form amyloids more easily than many other globular proteins. Our results relate to the method of Stanković *et al.* [Stanković I *et al.* (2017) IPSI BgD Tran Int Res 13, 47–51], who applied a hybrid textual-search and geometric approach for finding amyloids in the PDB. If one intends to identify a subset of the PDB for certain applications, the identification algorithm needs to be re-run periodically, since in 2017 on average 30 new entries per day were deposited in the data bank. Our web server is updated regularly and automatically, and the identified amyloid and partial amyloid structures can be viewed or their list can be downloaded from the following website https://pitgroup.org/amyloid.

The Protein Data Bank (PDB) is a continually developing public resource of spatial structures of proteins and nucleic acids [1]. Today the database contains more than 135 000 structures. The geometric properties of these molecules can be analyzed by bioinformatical tools, and one may infer significant new relations in these very complex macromolecular structures through such analyses [2–9]. Here, we are interested in amyloid structures in the Protein Data Bank.

Amyloids are misfolded protein aggregates that are present in numerous biological structures including the cellular surface of a number of microorganisms [10,11], where they have a role in host–pathogen interaction; the silkmoth chorion and some fish choria, forming protective films [12]; the immune system of certain insects, helping in the encapsulation of pathogens and parasites [13]; healthy human pituitary secretory granules, for storing peptide hormones [14]; and human amyloidoses and several neurodegenerative diseases [15]. Amyloid structures sometimes show prion-like infective properties [15–17]. Cerebral β-amyloid plaques have long been considered to be biomarkers of Alzheimer's disease [18–21], although more recently their validity has been challenged by several authors

Abbreviations

hCG, human chorionic gonadotropin; PDB, Protein Data Bank; TS, thymidylate synthase A.

[22–24].Mechanisms of amyloid formation are reviewed in [25,26]. Amyloid fibers are formed from parallel β-sheets, with hydrogen bonds between the parallel strands. It is widely accepted that amyloid formation requires the presence of a nucleus or a seed of amyloid-forming segments with exposed edges of β-sheet structures [25–28].

Since amyloid fibers are insoluble in water, until very recently there were no high-resolution structures deposited in the RCSB PDB [1]. Today, one can find several dozen atomic-resolution amyloid structures in the PDB, and this dataset has opened up the possibility of the analysis and the data mining of the properties of these misfolded proteins, using their high-resolution spatial structure.

The first step in this direction is the identification of the amyloid structures in the PDB.

Amyloid and amyloid-precursor molecules have been collected and predicted using protein-sequencing data in numerous articles (e.g. in the AMYPdb resource [29], or in [30]). We are interested in the analysis of the spatial protein structures for finding amyloid and amyloid-precursor molecules, rather than in the analysis of residue-sequence properties of proteins of unknown three-dimensional structure.

In a remarkable piece of work, Stanković *et al.* [31] screened the PDB for amyloid structures by applying the following procedure: (a) by a textual search, those PDB entries were selected that contain the word ‘amyloid’ or any of another 38 words describing amyloid precursors; (b) helical structures, identified by torsion angles, were discarded; and (c) parallel, near-linear fragments of length at least four residues were identified; structures without these fragments were also discarded.

In the present work, we prepared an automatically updated list of amyloid and potentially amyloidogenic structures from the PDB, applying only the geometric properties of β-sheets; consequently, we did not use any textual search, referring to the annotations of the PDB entries. By this choice, we intended to identify not only the aggregated amyloid entries and known precursors, but also those globular proteins that contain small, locally amyloid-like substructures. We assumed that these globular proteins may also be amyloidogenic, i.e. they can more easily turn into amyloid fibers than globular proteins without these structural elements.

Since the PDB grows very quickly – in 2017, every day on average 30 new structures were deposited – we needed to construct an automatically updated web server, which periodically examines the new PDB entries and includes the newly deposited amyloid and potentially amyloidogenic structures. Consequently, our list does not give a snapshot of the amyloid structures in the PDB at a given time as with other efforts, but rather presents a live list of these structures. Our online resource is available at https://pitgroup.org/amyloid/.

Materials and methods

Here we describe the selection method that generates the Extended Amyloid List at https://pitgroup.org/amyloid/.

In contrast with Stanković *et al.* [31], we did not make any selection through a textual search in the annotation fields of the PDB files. Instead, we attempted to collect the minimal set of geometric rules, which already return the amyloids found in [31], plus novel, globular proteins with possible amyloidogenic substructures.

We constructed three rules, (a)–(c), in which, informally, rule (a) assures the parallelism of the β-sheets; rule (b) excludes the structures with large curvature, e.g. locally parallel helices; and rule (c) ensures that the approximately straight and parallel fragments are sufficiently long, compared with the total length of the chain. More formally, the following rules were applied:

(a) For finding parallel β-sheets. Stanković *et al.* [31] selected parallel chain segments by requiring the distance difference between the closest C_α atoms of the fragment to be less than 1.5 Å. Instead of this condition, we have applied a limit to the standard deviation σ between the closest C_α atoms of the fragment such that its value is less than 1.5 Å. We think that this approach is more tolerant of singular, random errors in the structure, while it is strict enough to characterize the parallel polypeptide chains in the amyloid structures. More technically, our condition can be re-phrased as follows. Let us consider two separate chains of the structures, A and B, both identified as β-sheets. Next, we compute the array $C(A)_{\text{dist}}$, which contains the minimum distance for every C_α atom of chain A from the closest C_α that is located in chain B. Next, we identify the maximal subchains F of A, satisfying $\sigma(C(F)_{\text{dist}}) \leq 1.5$, while every distance in vector $C(F)_{\text{dist}}$ is required to be between 2 and 15 Å.

(b) Excluding structures with large curvature. Stanković *et al.* [31] excluded helical structures from consideration. We apply a locally verifiable angular condition for the fragments F as follows. Fragment F, which satisfies the conditions in (a), needs also to satisfy the condition that the angles of each of three consecutive C_α atoms, averaged for the fragment F, need to be between 110° and 180°. In other words, these angles, on average, should be obtuse angles between 110° and 180°.

(c) Condition for the minimum length of parallel fragments. $\text{len}(F) \geq \text{len}(\text{A})/7$, where F denotes the same as in rule (a), and $\text{len}(X)$ denotes the length of chain X, measured in residues.

The specific parameters for the conditions above were selected for including all multi-chain amyloid structures that were also found in Stanković *et al*. [31]. We did not aim to find amyloid-like structures containing only one single polypeptide chain since the amyloid structures contain a large number of approximately parallel fibers, each consisting of different chains. While the PDB contains partial amyloid structures with one single chain (e.g. 1HZ3), these structures will not be listed within our results, since they lack the characteristic property of nearly parallel, distinct polypeptide chains. The search for distinct parallel chains is useful for disallowing single chains with long parallel subsections, for example, β-barrel structures, such as the bacterial porin structure 4RLC. Since amyloid aggregates always consist of a large number of distinct, approximately parallel polypeptide chains, our condition is not restrictive.

Results and discussion

Amyloid structures

We have found that our list at https://pitgroup.org/amyloid/ contains all amyloid structures with at least two polypeptide chains that are listed by Stanković *et al.* [31]. For example, the classical amyloid structures of 2KIB, 2N0A, 5KO0, 2LBU and 2LMN are all present in the list.

Possible amyloidogenic structures

Here, we review four non-amyloid proteins that were found by our screening algorithm and that are listed at https://pitgroup.org/amyloid/. We also give literature evidence showing links to the amyloid formation of these molecules. These findings are witness to the power of our algorithm, but clearly we cannot review here the more than 500 structures presented on the webpage https://pitgroup.org/amyloid/.

- 1HCN: Human chorionic gonadotropin (hCG) (Fig. 1A). It is a placenta-produced human hormone and applied in numerous pregnancy tests and in legal and illegal drug products, including physical performance-enhancing and weight-loss preparations. It is reported to increase β-amyloid levels in rats [32] and to increase β-cleavage of an amyloid-precursor protein [33]. Protein hCG also has a role in amyloid β precursor protein expression and modulation in human cells [34], and in protein folding regulation in endoplasmic reticulum [35]. We believe that these roles of hCG are closely related to particular geometric properties of its parallel β-sheets.
- 1BSF: Thymidylate synthase A (TS) from *Bacillus subtilis* (Fig. 1B). Thymidylate synthase has an important role in DNA synthesis, and its aggregational properties have been studied for a long

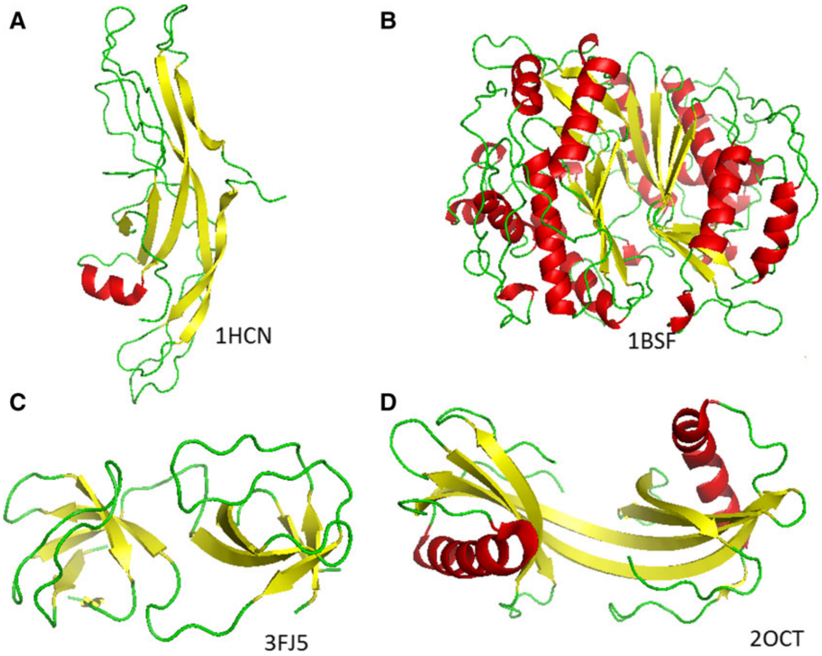

Fig. 1. Protein structures 1HCN, 1BSF, 3FJ5 and 2OCT, with partial amyloid-like substructures. All of these entries are documented as amyloidogenic in the literature.

time [36]. The human TS is a primary target of cancer chemotherapy, most importantly by 5-fluorouracil, a strong-binding TS inhibitor, applied widely in colon, esophageal, stomach, pancreatic, breast and cervical cancers. In Fig. 1B, it can be clearly seen that the parallel β-sheets are hidden in the dimeric structure. TS also has a monomeric form with distinct function, and the dimeric and the monomeric forms are in equilibrium in humans [37]. Therefore, the hidden β-sheets in the monomeric form may become accessible and may play a role in aggregation processes.

- 3FJ5: Tyrosine kinase c-Src (Fig. 1C). It has a role in the mitogen-activated protein kinase pathway, and in the development of breast cancers in animals and humans [38,39]. It has been shown that the SH3 domain of this protein aggregates to form amyloid fibrils at mild acidic pH values [40]. It is suggested that amyloid-associated microgliosis is strengthened by tyrosine kinase c-Src activity [41,42]. It has also been noted that mitogen-activated protein kinase signaling cascade dysfunction in fibroblasts is specific to Alzheimer's disease [43].
- 2OCT: Stefin B (cystatin B) tetramer (Fig. 1D), an intracellular thiol protease inhibitor. It is known to form amyloid fibrils *in vitro* [44]; its role in amyloidogenesis is detailed in [45] and [46].

Conclusions

We have demonstrated the validity of three geometric structural selection rules, which identify amyloid fibrils and plaques in the PDB. Additionally, these rules find non-amyloid soluble proteins, among which we have identified several amyloidogenic ones by scanning the literature. We believe that the great majority of the soluble proteins in the list show also – mostly still undocumented – amyloidogenic properties.

Acknowledgements

KT and BV were partially funded by the VEKOP-2.3.2-16-2017-00014 program, supported by the European Union and the State of Hungary, co-financed by the European Regional Development Fund, and by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002). VG was partially funded by grant NKFI-126472 of the National Research, Development and Innovation Office of Hungary.

Author contributions

VG initiated the study and analyzed results. BV created the web interface and the update mechanism. KT designed and programmed the geometric filtering algorithm and fine-tuned the geometric constraints.

Conflict of interest

The authors declare no conflict of interest.

Data availability

The automatically updated web page is available at https://pitgroup.org/amyloid/. The page contains the list of the PDB entries found by our program; each entry is given in graphical form, hyperlinked to the structures at the RCSB PDB site https://www.rcsb.org/pdb.

The Python source code of the software program that generates the PDB_Amyloid list is available at http://uratim.com/amyloid/amyloid_pit.zip.

The page https://pitgroup.org/amyloid/ contains not only the graphical representation of the proteins found but also a list of their PDB codes at https://pitgroup.org/apps/amyloid/amyloid_list.

References

1 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE (2000) The protein data bank. *Nucleic Acids Res* **28**, 235–242.

2 Iván G and Grolmusz V (2014) On dimension reduction of clustering results in structural bioinformatics. *Biochim Biophys Acta* **1844**, 2277–2283.

3 Ivan G, Szabadka Z and Grolmusz V (2010) A hybrid clustering of protein binding sites. *FEBS J* **277**, 1494–1502.

4 Ivan G, Szabadka Z, Ordog R, Grolmusz V and Naray-Szabo G (2009) Four spatial points that define enzyme families. *Biochem Biophys Res Commun* **383**, 417–420.

5 Ivan G, Szabadka Z and Grolmusz V (2007) Being a binding site: characterizing residue composition of binding sites on proteins. *Bioinformation* **2**, 216–221.

6 Ivan G, Szabadka Z and Grolmusz V (2010) Cysteine and tryptophan anomalies found when scanning all the binding sites in the protein data bank. *Int J Bioinform Res Appl* **6**, 594–608.

7 Szabadka Z and Grolmusz V (2007) High throughput processing of the structural information in the protein data bank. *J Mol Graph Model* **25**, 831–836.

8 Szabadka Z and Grolmusz V (2006) Building a structured PDB: The RS-PDB database. In *Proceedings of the 28th IEEE EMBS Annual International Conference*, New York, NY, Aug 30–Sept 3, 2006, pp. 5755–5758

9 Ordog R, Szabadka Z and Grolmusz V (2008) Analyzing the simplicial decomposition of spatial protein structures. *BMC Bioinformat* **9** (Suppl. 1), S11.

10 Gebbink MFBG, Claessen D, Bouma B, Dijkhuizen L and Wösten HAB (2005) Amyloids – a functional coat for microorganisms. *Nat Rev Microbiol* **3**, 333–341.

11 Blanco LP, Evans ML, Smith DR, Badtke MP and Chapman MR (2012) Diversity, biogenesis and function of microbial amyloids. *Trends Microbiol* **20**, 66–73.

12 Iconomidou VA and Hamodrakas SJ (2008) Natural protective amyloids. *Curr Protein Pept Sci* **9**, 291–309.

13 Falabella P, Riviello L, Pascale M, Di Lelio I, Tettamanti G, Grimaldi A, Iannone C, Monti M, Pucci P, Tamburro AM *et al.* (2012) Functional amyloids in insect immune response. *Insect Biochem Mol Biol* **42**, 203–211.

14 Maji SK, Perrin MH, Sawaya MR, Jessberger S, Vadodaria K, Rissman RA, Singru PS, Nilsson KP, Simon R, Schubert D *et al.* (2009) Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science* **325**, 328–332.

15 Soto C, Estrada L and Castilla J (2006) Amyloids, prions and the inherent infectious nature of misfolded protein aggregates. *Trends Biochem Sci* **31**, 150–155.

16 Caughey B, Baron GS, Chesebro B and Jeffrey M (2009) Getting a grip on prions: oligomers, amyloids, and pathological membrane interactions. *Annu Rev Biochem* **78**, 177–204.

17 Aguzzi A and Rajendran L (2009) The transcellular spread of cytosolic amyloids, prions, and prionoids. *Neuron* **64**, 783–790.

18 Alzheimer A (1907) Uber eine eigenartige erkrankung der hirnrinde. *Allgemeine Zeitschrife Psychiatrie* **64**, 146–148.

19 Prescott JW, Guidon A, Doraiswamy PM, Choudhury KR, Liu C, Petrella J, and Alzheimer's Disease Neuroimaging Initiative (2014) The Alzheimer structural connectome: changes in cortical network topology with increased amyloid plaque burden. *Radiology* **273**, 175–184.

20 Buyong Ma and Ruth Nussinov (2002) Stabilities and conformations of Alzheimer's beta-amyloid peptide oligomers (Abeta 16-22, Abeta 16-35, and Abeta 10-35): Sequence effects. *Proc Natl Acad Sci USA* **99**, 14126–14131.

21 Zheng J, Jang H, Ma B, Tsai C-J and Nussinov R (2007) Modeling the Alzheimer Abeta17-42 fibril architecture: tight intermolecular sheet-sheet association and intramolecular hydrated cavities. *Biophys J* **93**, 3046–3057.

22 Pimplikar SW (2009) Reassessing the amyloid cascade hypothesis of Alzheimer's disease. *Int J Biochem Cell Biol* **41**, 1261–1268.

23 Holmes C, Boche D, Wilkinson D, Yadegarfar G, Hopkins V, Bayer A, Jones RW, Bullock R, Love S, Neal JW *et al.* (2008) Long-term effects of Aβ 42 immunisation in Alzheimer's disease: follow-up of a randomised, placebo-controlled phase I trial. *Lancet* **372**, 216–223.

24 Hyman BT (2011) Amyloid-dependent and amyloid-independent stages of Alzheimer disease. *Arch Neurol* **68**, 1062–1064.

25 Eisenberg D and Jucker M (2012) The amyloid state of proteins in human diseases. *Cell* **148**, 1188–1203.

26 Richardson JS and Richardson DC (2002) Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci USA* **99**, 2754–2759.

27 Nelson R, Sawaya MR, Balbirnie M, Madsen AØ, Riekel C, Grothe R and Eisenberg D (2005) Structure of the cross-β spine of amyloid-like fibrils. *Nature* **435**, 773.

28 Lee J, Culyba EK, Powers ET and Kelly JW (2011) Amyloid-β forms fibrils by nucleated conformational conversion of oligomers. *Nat Chem Biol* **7**, 602.

29 Béchec AL, Pawlicki S and Delamarche C (2008) Amypdb: a database dedicated to amyloid precursor proteins. *BMC Bioinformat* **9**, 273. https://doi.org/10.1186/1471-2105-9-273

30 Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F and Vendruscolo M (2008) Prediction of aggregation-prone regions in structured proteins. *J Mol Biol* **380**, 425–436.

31 Stanković I, Hall MB and Zarić SD (2017) Construction of amyloid PDB files database. *IPSI BgD Tran Int Res* **13**, 47–51.

32 Berry A, Tomidokoro Y, Ghiso J and Thornton J (2008) Human chorionic gonadotropin (a luteinizing hormone homologue) decreases spatial memory and increases brain amyloid-β levels in female rats. *Horm Behav* **54**, 143–152.

33 Saberi S, Du YP, Christie M and Goldsbury C (2013) Human chorionic gonadotropin increases β-cleavage of amyloid precursor protein in SH-SY5Y cells. *Cell Mol Neurobiol* **33**, 747–751.

34 Porayette P, Gallego MJ, Kaltcheva MM, Vadakkadath Meethal S and Atwood CS (2007) Amyloid-β precursor protein expression and modulation in human embryonic stem cells: a novel role for human chorionic gonadotropin. *Biochem Biophys Res Comm* **364**, 522–527.

35 Ruddon RW, Sherman SA and Bedows E (1996) Protein folding in the endoplasmic reticulum: lessons from the human chorionic gonadotropin β subunit. *Protein Sci* **5**, 1443–1452.

36 Agarwalla S, Gokhale RS, Balaram P and Santi DV (1996) Covalent tethering of the dimer interface annuls aggregation in thymidylate synthase. *Protein Sci* **5**, 270–277.

37 Genovese F, Ferrari S, Guaitoli G, Caselli M, Costi MP and Ponterini G (2010) Dimer–monomer equilibrium of human thymidylate synthase monitored by fluorescence resonance energy transfer. *Protein Sci* **19**, 1023–1030.

38 Guy CT, Muthuswamy SK, Cardiff RD, Soriano P and Muller WJ (1994) Activation of the c-src tyrosine kinase is required for the induction of mammary tumors in transgenic mice. *Genes Dev* **8**, 23–32.

39 Biscardi JS, Ishizawar RC, Silva CM and Parsons SJ (2000) Tyrosine kinase signalling in breast cancer: epidermal growth factor receptor and c-src interactions in breast cancer. *Breast Cancer Res* **2**, 203.

40 Bacarizo J, Martinez-Rodriguez S, Manuel Martin-Garcia J, Andujar-Sanchez M, Ortiz-Salmeron E, Neira JL and Camara-Artigas A (2014) Electrostatic effects in the folding of the SH3 domain of the c-Src tyrosine kinase: pH-dependence in 3D-domain swapping and amyloid formation. *PLoS ONE* **9**, e113224.

41 Dhawan G and Combs CK (2012) Inhibition of Src kinase activity attenuates amyloid associated microgliosis in a murine model of Alzheimer's disease. *J Neuroinflammation* **9**, 117.

42 Dhawan G, Floden AM and Combs CK (2012) Amyloid-β oligomers stimulate microglia through a tyrosine kinase dependent mechanism. *Neurobiol Aging* **33**, 2247–2261.

43 Zhao W-Q, Ravindranath L, Mohamed AS, Zohar O, Chen GH, Lyketsos CG, Etcheberrigaray R and Alkon DL (2002) MAP kinase signaling cascade dysfunction specific to Alzheimer's disease in fibroblasts. *Neurobiol Dis*, **11**, 166–183.

44 Žerovnik E, Pompe-Novak M, Škarabot M, Ravnikar M, Muševič I and Turk V (2002) Human stefin B readily forms amyloid fibrils in vitro. *Biochim Biophys Acta* **1594**, 1–5.

45 Jenko Kokalj S, Gunčar G, Štern I, Morgan G, Rabzelj S, Kenig M, Staniforth RA, Waltho JP, Zerovnik E and Turk D (2007) Essential role of proline isomerization in stefin b tetramer formation. *J Mol Biol* **366**, 1569–1579.

46 Škerget K, Taler-Verčič A, Bavdek A, Hodnik V, Čeru S, Tušek-Žnidarič M, Kumm T, Pitsi D, Pompe-Novak M, Palumaa P *et al.* (2010) Interaction between oligomers of stefin b and amyloid-β in vitro and in cells. *J Biol Chem* **285**, 3201–3210.