On the Border of the Amyloidogenic Sequences: Prefix Analysis of the Parallel Beta Sheets in the PDB_Amyloid Collection

Kristóf Takács[a], Vince Grolmusz[a,b,∗]

[a]*PIT Bioinformatics Group, Eötvös University, H-1117 Budapest, Hungary*
[b]*Uratim Ltd., H-1118 Budapest, Hungary*

Abstract

The Protein Data Bank (PDB) today contains more than 153,000 entries with the 3-dimensional structures of biological macromolecules. Using the rich resources of this repository, it is possible identifying subsets with specific, interesting properties for different applications. Our research group prepared an automatically updated list of amyloid- and probably amyloidogenic molecules, the PDB_Amyloid collection, which is freely available at the address `http://pitgroup.org/amyloid`. This resource applies exclusively the geometric properties of the steric structures for identifying amyloids. In the present contribution, we analyze the starting (i.e., prefix) subsequences of the characteristic, parallel beta-sheets of the structures in the PDB_Amyloid collection, and identify further appearances of these length-5 prefix subsequences in the whole PDB data set. We have identified this way numerous proteins, whose normal or irregular functions involve amyloid formation, structural misfolding, or anticoagulant properties, simply by containing these prefixes: including the T-cell receptor (TCR), bound with the major histocompatibility complexes MHC-1 and MHC-2; the p53 tumor suppressor protein; a mycobacterial RNA polymerase transcription initialization complex; the human bridging integrator protein BIN-1; and the tick anti-coagulant peptide TAP.

Running head: On the Borders of the Amyloidogenic Sequences

Keywords: PDB, amyloid, amyloid-precursor, amyloidogenic proteins, web-server, prefix, suffix

---

∗Corresponding author

Email addresses: `takacs@pitgroup.org` (Kristóf Takács), `grolmusz@pitgroup.org` (Vince Grolmusz)

Introduction

Amyloids are misfolded protein aggregates, which are present in numerous biological organisms as structural building blocks or immunological agents [1, 2, 3, 4, 5, 6, 7]. In humans, the amyloid formation is frequently associated with neurodegenerative diseases and abnormal metabolic conditions [8, 9, 10, 11].

The structural studies of amyloid aggregates were considered to be difficult until recently, since being aggregates, they cannot be crystallized and measured by X-ray diffractometry. With the recent developments of solid-state NMR and cryo-electron microscopy, dozens of amyloid structures were deposited in the Protein Data Bank (PDB) [12, 13] in the past several years.

With the more than 100 amyloid structures among the PDB's 156 thousand entries, it is now possible to define structural characteristics, which well-describe amyloid structures. One good approach was made by [13], where the authors, with the application of a combination of textual search and specific geometric conditions, successfully retrieved the known amyloid structures from the PDB.

In a recent work of ours [14], we have defined a geometric set of constraints, by which we selected all the amyloid molecules, found by the method of [13], plus numerous globular proteins, with partial amyloid-like substructures. We emphasize that we were using geometric constraints for the β-sheet regions in the coordinate sections of the PDB files, without *any* textual search in the annotation section of those files. Since the annotation sections of PDB files are known to be less reliable than the coordinate sections, this technique increases the reliability of our results, and, additionally, helps in devising the proper definition of the amyloid-like structures. The resulting selection of the PDB entries, called the PDB_Amyloid list, is available as an automatically and regularly updated list of PDB entries, at the site `http://pitgroup.org/amyloid/`. Since, on the average, around 30 new PDB entries are deposited every day, the "automatic update" feature is clearly necessary for this service. The PDB_Amyloid list contains more than 640 entries today.

The geometric constraints, applied in [14], are as follows:

(i) First, parallel β-sheet segments are identified. The parallel segments need to be on separate polypeptide chains, their distance needs to be between 2 and 15 Å, and the standard deviation of their distance needs to be less than 1.5 Å.

(ii) Second, the large curvature parallel segments are excluded;

(iii) Third, the parallel segments need to cover at least the one-seventh of the length of the whole chain.

For a more detailed mathematical description of the constraints above, we refer to the original article [14].

We note that the requirement of considering only segments on separate polypeptide chains efficiently excludes hairpin and β-barrel structures, and also partial molecular structures, labeled as "amyloids", but lacking the repeated, parallel β-sheets in the PDB-deposited files.

Prefixes

The more than 640 PDB entries, available at `http://pitgroup.org/amyloid/`, yield a rich set of amyloid-related molecular structures. The list of the globular proteins (i.e., not the misfolded, aggregated amyloid structures) have a special feature: These molecules remained soluble, but they have partial sub-structures, satisfying the conditions (i), (ii) and (iii) above. We believe that the sequence-borders of the β-sheets in these structures have specific roles in the prevention of the transitions to amyloid state: in the globular proteins, these border-regions may prevent or regulate the formation of aggregated amyloid structures from the protein.

In the present contribution, we consider the prefixes of the parallel β-sheets of the entries of the PDB_Amyloid list `http://pitgroup.org/amyloid/`. These prefixes are the starting subsections, where the order of the residues is the default N-terminus through C-terminus.

Here we identify the most frequently found prefixes in the PDB_Amyloid list, and then we search for them in the *complete* Protein Data Bank. We identify numerous interesting hits, which have proven connections to amyloid formation. We stress that the hits, analyzed below, are found by sequence-searches in the whole PDB, without using *any* additional structural constraints, where the sequences we searched for were the prefixes, identified in the PDB_Amyloid list.

Methods

First, the defining parallel β-sheets, satisfying the conditions (i), (ii) and (iii) in the Introduction, of the PDB_Amyloid list `http://pitgroup.org/amyloid/`, were identified.

Next, we collected the length-5 prefixes of the form XXYYY, where the first three residues of the parallel β-sheet are YYY, and the last two residues, preceding the parallel section of the β-sheet, are XX.

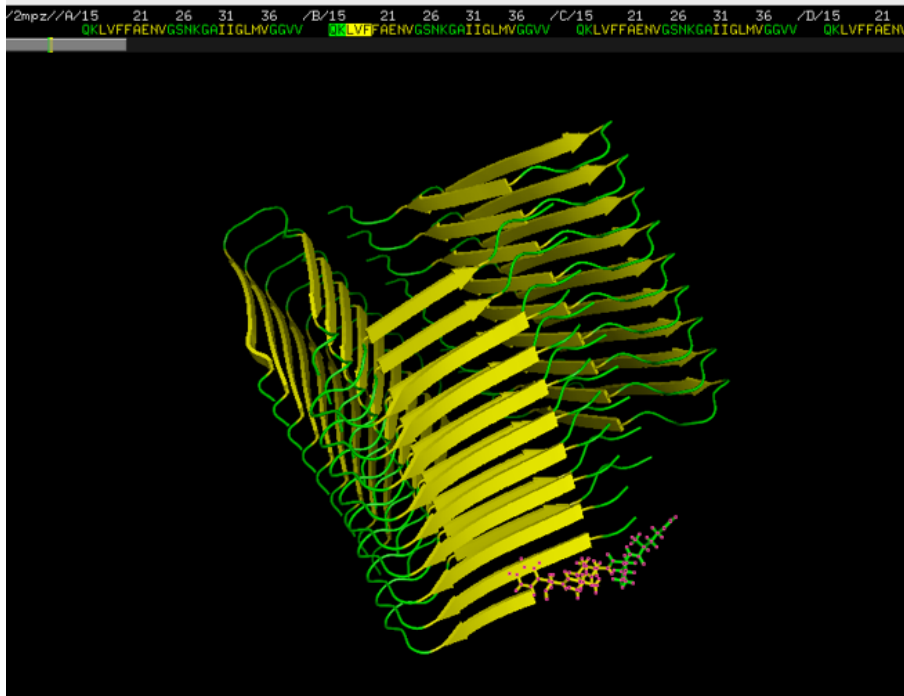Figure 1 depicts the GLN-LYS-LEU-VAL-PHE (QKLVF) prefix from the amyloid structure of the PDB entry 2MPZ.

Figure 1: PDB entry 2MPZ, depicted with PyMol. Yellow color denotes β-sheets. The QKLVF prefix, where QK is green (it is not a part of the parallel β-sheet) and LVF is yellow (i.e., LVF are the first three residues of the parallel β-segment), is emphasized at the right bottom of the figure, while its corresponding sequence at the top center.

Next, we have counted the number of appearances of the prefixes and the suffixes in the parallel β-sheet segments in the PDB_Amyloid list http://pitgroup.org/amyloid/. Note that one PDB entry may contain more than one identical prefixes (like in the case of PDB entry 2MPZ, shown in Figure 1). Therefore, the prefix and suffix counts contain multiplicities of two types: (i) multiple appearances in the very same PDB entry, or (ii) multiple appearances in different – and possibly homologous – PDB entries. Instead of introducing an unnecessarily complex homogeneity-corrected counting method for the prefix- and suffix appearances in the PDB_Amyloid list, we just count their raw, uncorrected number of appearances. Since the inclusion or exclusion of the protein structures in the PDB mostly relate to the interest of researchers depositing the structures, and do not carry a statistical or biological meaning. Moreover, we do not count the appearances in the whole PDB, just in the amyloid-like sublist of PDB_Amyloid. These counts (either corrected or uncorrected) can only be used informally, and do not show the frequency of these subsequences in the protein structures in Nature.

4

Discussion and results

In what follows, we consider the PDB_Amyloid list and note if the prefix appears in structures, described by the application of NMR spectroscopy (both solid and liquid phase), or by X-ray diffractometry.

The QKLVF Prefix

The QKLVF prefix (i.e., GLN, LYS, LEU, VAL, PHE) appears 77 times in the NMR-identified members of the PDB_Amyloid list, in the following PDB-structures: 2LMN, 2LMO, 2LMP, 2LMQ, 2LNQ, 2MPZ. The 2MPZ structure is depicted in Figure 1. We are interested in the appearances of the QKLVF subsequence in the whole PDB, and we intend to identify the protein structures, which have the proven potential to turn to amyloids.

Among the numerous β-amyloid hits, which are not reviewed here, several interesting appearances of the QKLVF subsequence in globular proteins are in the T-cell receptors (TCR), bound with the major histocompatibility complexes MHC-1 and MHC-2, in the PDB entries 4P5T, 4OZF, 3QIU, 3QIW, 1BD2, 2IAN, 2IAM, 2IAL, 4WW1, 4WW2, 5XOT. Very interestingly, the misfolded MHC molecules in activated T-cells have a signaling function [15]. Additionally, the MHC molecule is known to misfold in several cases when in complex with TCR, and then it is is degraded by housekeeping enzymes [?]. These articles show that the normally globular MHC molecules are known to misfold if in complex with the TCR molecule, containing the QKLVF subsequence.

The GEYFT Prefix

The GEYFT prefix (GLY, GLU, TYR, PHE, THR) appears 48 times in the NMR-identified members of the PDB_Amyloid list, in the PDB entries 1OLG, 1SAE, 1SAF, 1SAK, 1SAL, 3SAK. These are non-amyloid structures. One of them, 1OLG is depicted in Figure 2.
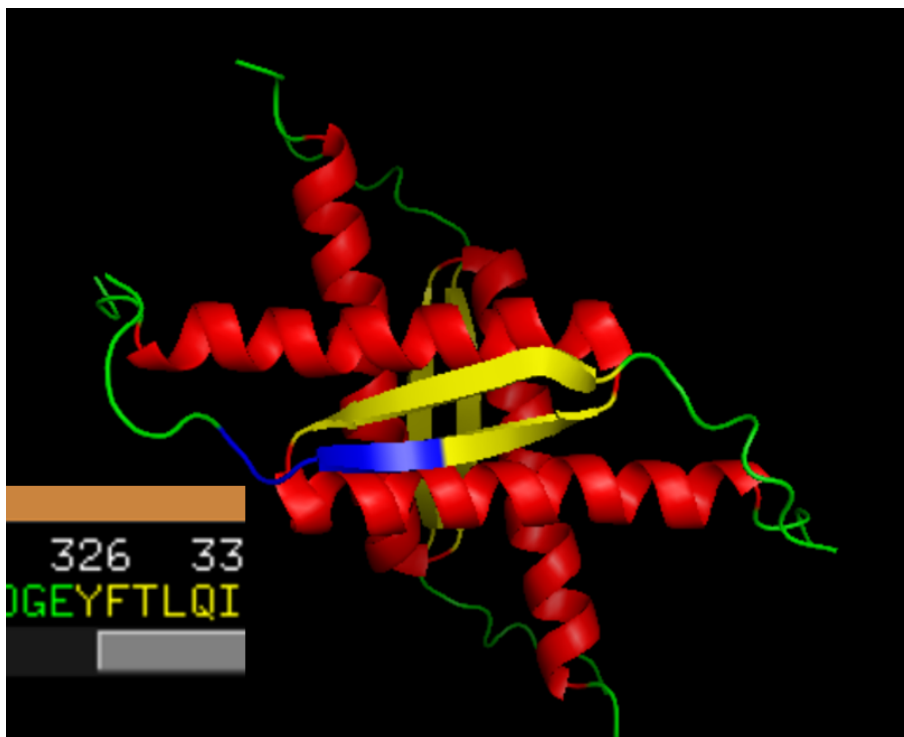
Figure 2: PDB entry 1OLG, depicted with PyMol. Yellow color denotes β-sheets. The GEYFT prefix, where GE are green (it is not a part of the parallel β-sheet) and YFT are yellow (i.e., YFT are the first three residues of the parallel β-segment), is emphasized at the left middle section of the figure, while its corresponding sequence at the lower left corner.

Numerous GEYFT appearances in the PDB_Amyloid list and also in the whole PDB are in p53 structures. p53 is a major tumor suppressor protein, whose gene is mutated in half of the human cancers [16, 17, 18], and both its mutational deficiency in humans and the knock-out of its gene in mice imply early on-set cancers [19, 20]. It is very surprising that p53 mutations have a tendency of prion-like, contagious amyloid transitions: it is found that the amyloid-like aggregation plays a role in the loss of the p53 function in several organisms and cell types [21, 22, 23].

We note that identifying non-amyloid p53 structures in the PDB_Amyloid list shows the power of the methods by which the PDB_Amyloid list was created [14]: p53, an important non-amyloid structure with amyloidogenic properties is found in the list. We also note that numerous appearances of the GEYFT sequence in the whole PDB are also found in the p53 proteins.

Many GEYFT prefixes in the whole PDB are found in *Mycobacterium* (either *tuberculosis* or *smegmatis*) RNA polymerase transcription initialization complexes (e.g., 6DVC, 6JCX, 6JCY, 5ZX2). While it is not documented that these initialization complexes form amyloids, other bacterial transcriptional regulators

do form amyloids. The *Bacillus subtilis* HeID, an RNA polymerase interacting helicase forms amyloids, as it was reported recently in [24]. Another finding that a mycobacterial global transcriptional factor, CarD, also forms amyloids, both *in vivo* and *in vitro* [25]. Therefore, it would not be surprising if the GEYFT-containing mycobacterial RNA polymerase transcription initialization complexes also formed amyloid fibrils.

The HQKLV Prefix

The HQKLV prefix (HIS, GLN, LYS, LEU, VAL) appears 25 times in the NMR-identified members of the PDB_Amyloid list, in the PDB entries 2LMN, 2LMO, 2LMP, 2LMQ; these are all β-amyloid fibrils. If we search for the HQKLV subsequence in the whole PDB, we find numerous amyloid structures and some human amphiphysins: human amphiphysin isoform 1 (PDB codes 3SOG, 4ATM), and human BIN1/amphiphysin II (2FIC). Interestingly, the HQKLV subsequence appears in these pure α-helix BIN1-structures as the part of the helix (Figure 3).
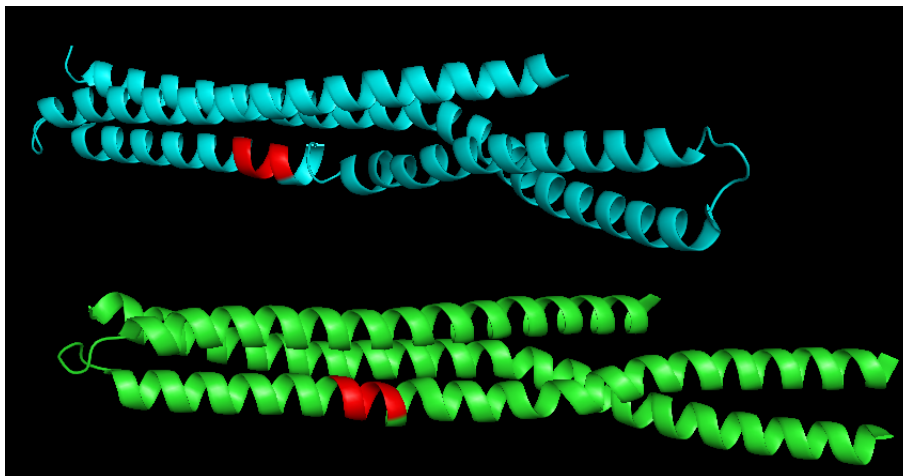

Figure 3: PDB entry 2FIC: the BAR domain of the human Bin1/amphiphysin II, depicted with PyMol. The red colored sections of the α-helices correspond to the HQKLV subsequence.

BIN1 is not known to form amyloid-fibrils, but it is well-known to relate to late-onset Alzheimer's-disease: its gene is the second most important risk locus for Alzheimer's disease (after APOE: apolipoprotein E) [26], it is related to increased susceptibility for Alzheimer's disease [?]. More recently, it was shown that BIN1 regulates BACE trafficking and β-amyloid production. Therefore, we may conjecture that the HQKLV subsequence plays a role in amyloid-formation, even if it is in an α-helix in BIN1 structures (Figure 3).

The GGERA Prefix

The GGERA prefix (GLY. GLY. GLU, ARG, ALA) appears 106 times in the X-ray crystallography-identified members of the PDB_Amyloid list, in the PDB entries 1DW9, 1DWK, 2IU7, 2IV1, 2IVQ, 4Y42; these are all bacterial cyanases. If we search for the GGERA subsequence in the whole PDB, the most interesting hit is the structure 1TCP: this is a tick anticoagulant peptide (TAP). The position of the GGERA subsequence is depicted in Figure 4.
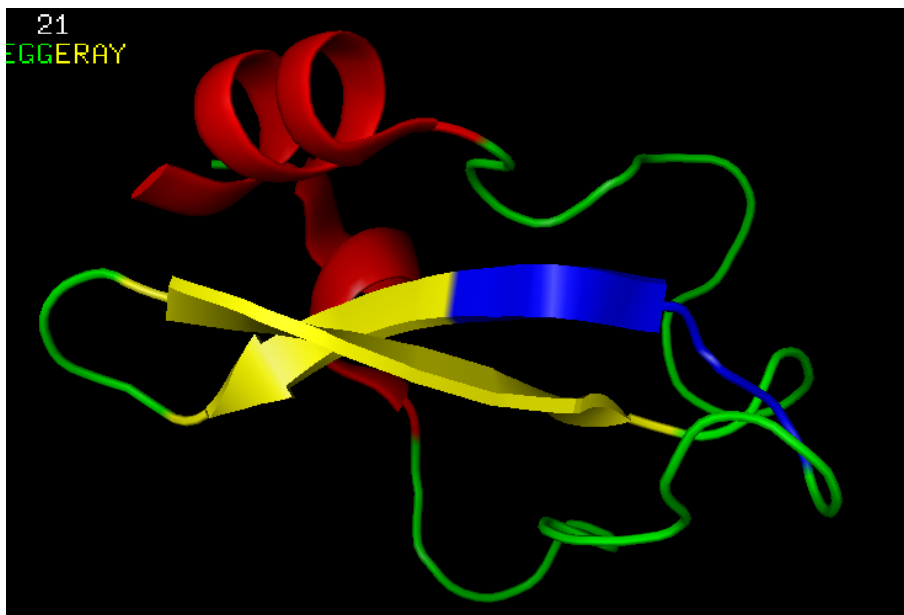

Figure 4: PDB entry 1TCP: the tick anticoagulant peptide (TAP), depicted with PyMol. The blue-colored section correspond to the GGERA prefix: GG preceeding the first three residues, ERA, in the β-sheet.

The β-sheet prefixes, listed above, were all related to prion- or amyloid-formation. Here, GGERA is found in an anti-coagulant molecule: the tick anti-coagulant peptide. Therefore, the GGERA subsequence is

- the prefix of the parallel β-sheet sections of several soluble proteins (cyanases) from the PDB_Amyloid list, therefore the β-sheet, which starts with the GGERA sequence, is similar to those in the amyloid-structures, by satisfying properties (i), (ii) and (iii), listed in the Introduction;

- but the cyanases 1DW9, 1DWK, 2IU7, 2IV1, 2IVQ, 4Y42 are all soluble proteins.

Consequently, since GGERA also appears in the anti-coagulant 1TCP, it may have anti-amyloidogenic properties.

Conclusions

By searching for the prefixes of the parallel β-sheet sections of the entries in the PDB_Amyloid list in `http://pitgroup.org/amyloid/`, we were able to find numerous proteins in the whole PDB, from which only very recently were shown that they relate to the amyloid-formation. We conjecture that the prefixes listed may have structural roles in these amyloidogenic properties.

Data availability

The automatically updated PDB_Amyloid web page is available at `http://pitgroup.org/amyloid/`. The list of the PDB codes of the PDB_Amyloid can be viewed and downloaded at `http://pitgroup.org/apps/amyloid/amyloid_list`.

Author contributions:

VG initiated the study, analyzed the results and wrote the paper. KT identified the parallel β-sheet segments in the spatial protein structures, and the prefixes in those β-sheets, satisfying the constraints, and described their appearances.

References

[1] Martijn FBG Gebbink, Dennis Claessen, Barend Bouma, Lubbert Dijkhuizen, and Han AB Wösten. Amyloids - a functional coat for microorganisms. *Nature Reviews Microbiology*, 3(4):333–341, 2005.

[2] Luz P Blanco, Margery L Evans, Daniel R Smith, Matthew P Badtke, and Matthew R Chapman. Diversity, biogenesis and function of microbial amyloids. *Trends in Microbiology*, 20(2):66–73, 2012.

[3] Vassiliki A Iconomidou and Stavros J Hamodrakas. Natural protective amyloids. *Current Protein and Peptide Science*, 9(3):291–309, 2008.

[4] Patrizia Falabella, Lea Riviello, Mariarosa Pascale, Ilaria Di Lelio, Gianluca Tettamanti, Annalisa Grimaldi, Carla Iannone, Maria Monti, Piero Pucci, Antonio Mario Tamburro, et al. Functional amyloids in insect immune response. *Insect Biochemistry and Molecular Biology*, 42(3):203–211, 2012.

[5] Samir K Maji, Marilyn H Perrin, Michael R Sawaya, Sebastian Jessberger, Krishna Vadodaria, Robert A Rissman, Praful S Singru, K Peter R Nilsson, Rozalyn Simon, David Schubert, et al. Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science*, 325 (5938):328–332, 2009.

[6] Nora Taricska, Daniel Horvath, Dora K Menyhard, Hanna Akontz-Kiss, Masahiro Noji, Masatomo So, Yuji Goto, Toshimichi Fujiwara, and Andras Perczel. The route from the folded to the amyloid state: Exploring the potential energy surface of a drug-like miniprotein. *Chemistry (Weinheim an der Bergstrasse, Germany)*, January 2020. ISSN 1521-3765. doi: 10. 1002/chem.201905181.

[7] Daniel Horvath, Dora K Menyhard, and Andras Perczel. Protein aggregation in a nutshell: The splendid molecular architecture of the dreaded amyloid fibrils. *Current protein & peptide science*, 20:1077–1088, 2019. ISSN 1875-5550. doi: 10.2174/1389203720666190925102832.

[8] Alois Alzheimer. Uber eine eigenartige erkrankung der hirnrinde. *Allgemeine Zeitschrife Psychiatrie*, 64:146–148, 1907.

[9] Jeffrey W. Prescott, Arnaud Guidon, P Murali Doraiswamy, Kingshuk Roy Choudhury, Chunlei Liu, Jeffrey Petrella, and For the Alzheimer's Disease Neuroimaging Initiative . The Alzheimer Structural Connectome: Changes in cortical network topology with increased amyloid plaque burden. *Radiology*, page 132593, May 2014.

[10] Buyong Ma and Ruth Nussinov. 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 U S A*, 99 (22):14126–14131, Oct 2002. doi: 10.1073/pnas.212206899. URL `http://dx.doi.org/10.1073/pnas.212206899`.

[11] Jie Zheng, Hyunbum Jang, Buyong Ma, Chung-Jun Tsai, and Ruth Nussinov. Modeling the alzheimer abeta17-42 fibril architecture: tight intermolecular sheet-sheet association and intramolecular hydrated cavities. *Biophys J*, 93(9):3046–3057, Nov 2007. doi: 10.1529/biophysj.107.110700. URL `http://dx.doi.org/10.1529/biophysj.107.110700`.

[12] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[13] Ivana Stanković, Michael B Hall, and Snežana D Zarić. Construction of amyloid PDB files database. *The IPSI BgD Transactions on Internet Research*, 13(1):47–51, 2017. ISSN ISSN 1820-4503).

[14] Kristóf Takács, Bálint Varga, and Vince Grolmusz. PDB _Amyloid: an extended live amyloid structure list from the PDB. *FEBS Open Bio*, 9(1): 185–190, 2019.

[15] Susana G Santos, Simon J Powis, and Fernando A Arosa. Misfolding of major histocompatibility complex class i molecules in activated t cells allows cis-interactions with receptors and signaling molecules and is associated with tyrosine phosphorylation. *The Journal of biological chemistry*, 279:53062–53070, December 2004. ISSN 0021-9258. doi: 10.1074/jbc.M408794200.

[16] B Vogelstein, D Lane, and A J Levine. Surfing the p53 network. *Nature*, 408:307–310, November 2000. ISSN 0028-0836. doi: 10.1038/35042675.

[17] Ana I Robles and Curtis C Harris. Clinical outcomes and correlates of tp53 mutations and cancer. *Cold Spring Harbor perspectives in biology*, 2: a001016, March 2010. ISSN 1943-0264. doi: 10.1101/cshperspect.a001016.

[18] Magali Olivier, Monica Hollstein, and Pierre Hainaut. Tp53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology*, 2:a001008, January 2010. ISSN 1943-0264. doi: 10.1101/cshperspect.a001008.

[19] D Malkin, F P Li, L C Strong, J F Fraumeni, C E Nelson, D H Kim, J Kassel, M A Gryka, F Z Bischoff, and M A Tainsky. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science (New York, N.Y.)*, 250:1233–1238, November 1990. ISSN 0036-8075. doi: 10.1126/science.1978757.

[20] L A Donehower, M Harvey, B L Slagle, M J McArthur, C A Montgomery, J S Butel, and A Bradley. Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. *Nature*, 356:215–221, March 1992. ISSN 0028-0836. doi: 10.1038/356215a0.

[21] Shinjinee Sengupta, Samir K Maji, and Santanu K Ghosh. Evidence of a prion-like transmission of p53 amyloid in saccharomyces cerevisiae. *Molecular and cellular biology*, 37, September 2017. ISSN 1098-5549. doi: 10.1128/MCB.00118-17.

[22] Jerson L Silva, Claudia V De Moura Gallo, Danielly C F Costa, and Luciana P Rangel. Prion-like aggregation of mutant p53 in cancer. *Trends in biochemical sciences*, 39:260–267, June 2014. ISSN 0968-0004. doi: 10.1016/j.tibs.2014.04.001.

[23] Jerson L Silva, Elio A Cino, Iaci N Soares, Vitor F Ferreira, and Guilherme A P de Oliveira. Targeting the prion-like aggregation of mutant p53 to combat cancer. *Accounts of chemical research*, 51:181–190, January 2018. ISSN 1520-4898. doi: 10.1021/acs.accounts.7b00473.

[24] Gundeep Kaur, Srajan Kapoor, and Krishan G Thakur. Bacillus subtilis held, an rna polymerase interacting helicase, forms amyloid-like fibrils. *Frontiers in Microbiology*, 9:1934, 2018. ISSN 1664-302X. doi: 10.3389/fmicb.2018.01934.

[25] Gundeep Kaur, Soni Kaundal, Srajan Kapoor, Jonathan M Grimes, Juha T Huiskonen, and Krishan Gopal Thakur. Mycobacterium tuberculosis card, an essential global transcriptional regulator forms amyloid-like fibrils. *Scientific Reports*, 8:10124, July 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-28290-4.

[26] Meng-Shan Tan, Jin-Tai Yu, and Lan Tan. Bridging integrator 1 (BIN1): form, function, and Alzheimer's disease. *Trends in molecular medicine*, 19: 594–603, October 2013. ISSN 1471-499X. doi: 10.1016/j.molmed.2013.06. 004.