European Funds
Knowledge Education Development

European Union
European Social Fund

„BioTechNan - the programme of interdisciplinary cross-institutional post gradual studies KNOW in the field of Biotechnology and Nanotechnology"

University of Wrocław
Faculty of Biotechnology

MSc Jaroslaw Chilimoniuk

Bioinformatic and experimental analyses of bacterial functional amyloids CsgA and CsgB

Dissertation performed
in the Department of Genomics and Bioinformatics

Dissertation supervisor: Prof. dr hab. Pawel Mackiewicz
Co-supervisor: Dr. rer. nat. Vytautas Smirnovas

Wrocław, 2023

Wrocław University of Science and Technology
Uniwersytet Wrocławski
Wrocław University of Environmental and Life Sciences

The project is co-financed by the European Union from the European Social Fund

I would like to thank my supervisor, Prof. Paweł Mackiewicz, for his unwavering support and guidance throughout the entirety of my dissertation. His vast knowledge, experience, and expertise in the field of bioinformatics and molecular biology have been instrumental in shaping my research and helping me develop as a scientist. I am grateful for his constant encouragement, valuable feedback, and dedication to helping me achieve my academic and professional goals.

Additionally, I would like to express my gratitude to my second supervisor, Dr. Vytautas Smirnovas, for his insightful comments and suggestions that have been an important part of my research. His expertise in the field of protein structure and dynamics has been valuable in shaping my work and broadening my knowledge and understanding in this area. I appreciate his commitment and dedication to supporting my research and for his availability whenever I needed his assistance.

I would also like to express my sincere gratitude to my friend Dr. Michał Burdukiewicz for his valuable guidance, support, and encouragement throughout my dissertation. His expertise in the field of bioinformatics and computational biology has been instrumental in shaping the direction and focus of my research. I am especially grateful for his patience and understanding during the most challenging times of my project. His insights and feedback have been invaluable, and I will always be grateful for the mentorship that he has provided.

I would like to express my deepest gratitude to my colleagues and friends from the Department of Bioinformatics and Genomics at Wrocław University, who have supported me throughout my doctoral studies. Their guidance, insightful feedback, and constructive criticism have been invaluable in shaping my research and helping me grow as a scientist.

I am also grateful to the Multiparameter Diagnostic Group at BTU for providing me with the opportunity to work with their cutting-edge technologies and to collaborate with some of the most talented researchers in the field. Their expertise in molecular diagnostics and bioinformatics has been instrumental in advancing my work and expanding my knowledge.

Furthermore, I would like to thank the Amyloid Research Group at Vilnius University for their collaboration and shared resources, which have been essential in advancing my understanding of amyloid diseases and their molecular mechanisms.

Finally, I would like to extend my thanks to all the individuals who have supported me in various ways, including family, friends, and fellow researchers. Your encouragement, advice,

and friendship have been crucial in helping me navigate the challenges of a doctoral program and in keeping me motivated to pursue my goals.

Thank you all for your contributions to my dissertation and for making my research journey a rewarding and enriching experience.

Contents

Streszczenie

Analizy bioinformatyczne i eksperymentalne bakteryjnych amyloidów funkcjonalnych CsgA i CsgB

Amyloidy to białka związane z wieloma zaburzeniami klinicznymi, takimi jak choroba Alzheimera, Creutzfeldta-Jakoba czy Huntingtona. Mimo że białka te są mają zróżnicowaną budowę, ich cechą wspólną jest to, że posiadają strukturę β-kartki i wykazują tendencję do agregacji. Oprócz amyloidów niefunkcjonalnych, które mogą przyjmować różne struktury i są błędnie złożonymi wersjami normalnych białek, istnieją również amyloidy funkcjonalne, które spełniają ważne funkcje komórkowe. Jednymi z nich są białka curli, CsgA i CsgB, które stały się przedmiotem niniejszej rozprawy.

W ramach tej rozprawy dokonaliśmy eksperymentalnej walidacji naszego oprogramowania AmyloGram do przewidywania białek amyloidowych z wykorzystaniem tioflawiny T (ThT) oraz mikroskopii sił atomowych (AFM). Algorytm skutecznie rozpoznał eksperymentalnie potwierdzone peptydy i był odporny na przeuczenie. Spośród 24 testowanych peptydów, znaleźliśmy 16, które miały niepoprawne adnotacje w bazie AmyLoad.

Stosując bardziej obiektywne wyszukiwanie motywów, znaleźliśmy pięć powtarzających się regionów w sekwencjach CsgA i CsgB. Regiony w CsgA są rozdzielone, mają 21 reszt i 9 miejsc konserwatywnych, podczas gdy te w CsgB są przyległe do siebie, mają 22 reszty i 7 miejsc konserwatywnych. Regiony te charakteryzują się specyficznym rozmieszczeniem reszt polarnych i hydrofobowych, a także posiadają centralną glicynę, która rozbija dwie wstęgi β w danym regionie. Stosując dokładniejsze porównania sekwencji, odkryliśmy dodatkowy region, który jest umieszczony przed pozostałymi. Wykazuje on istotne podobieństwo do nich na poziomie sekwencji i potencjalnie może przyjmować struktury β.

Aby odpowiedzieć na pytanie, jak ewoluowały białka CsgA i CsgB, zebraliśmy ponad 15,000 ich homologów z konserwatywnymi domenami curli. Większość z nich zawiera również typowy N-terminalny peptyd sygnałowy. Rozległe analizy filogenetyczne wykazały, że białka te ewoluowały głównie u *Bacteroidota*, α-*Proteobacteria* i γ-*Proteobacteria*. CsgA i CsgB okazały się odległymi homologami i pojawiły się w wyniku duplikacji, gdy γ-*Proteobacteria* oddzieliły się od α- i β-*Proteobacteria*. Homologi te prawdopodobnie doświadczały poziomego transferu genów

pomiędzy różnymi grupami bakterii, a także do grzybów.

Zbadaliśmy również szczegółowo białka CsgA i CsgB u *Enterobacterales*. Ich pięć zduplikowanych regionów wykazuje odpowiednio siedem i sześć konserwatywnych miejsc, w tym reszty glicynowe, hydrofobowe i polarne, które są kluczowe do formowania struktur β.
Poszczególne regiony ujawniły różne tempo substytucji. Regiony CsgA ewoluowały szybciej niż CsgB. Region 5 wykazał najniższą dywergencję, co jest prawdopodobnie wynikiem selekcji na interakcje z regionem 1 innych cząsteczek białek curli. Zauważyliśmy ponadto duże korelacje w dystansach ewolucyjnych regionów, co sugeruje ich skoordynowaną ewolucję. Silniejsze korelacje w substytucjach zaobserwowaliśmy w regionach CsgA, co oznaczałoby, że interakcje między tymi regionami w tym białku powinny być bardziej konserwatywne niż w CsgB.

Ponadto oczyszczaliśmy wybrane warianty CsgA i CsgB, które miały usunięte regiony i badaliśmy wpływ tych regionów na szybkość agregacji przy użyciu eksperymentów ThT i AFM. Stwierdziliśmy, że proces ten może być spowalniany przez region 1, podczas gdy region 5 jest niezbędny do polimeryzacji fibryli amyloidowych ze względu na interakcje z regionem 1 innych cząsteczek CsgA.

Poszukiwaliśmy także cech charakterystycznych dla sekwencji amyloidów funkcjonalnych i niefunkcjonalnych. Mimo że amyloidy te są zróżnicowane, odkryliśmy specyficzne cechy, które mogą być wykorzystane do ich rozpoznania. Małe hydroksylowane aminokwasy, seryna i treonina, współwystępujące z innymi małymi aminokwasami, jak glicyną i alaniną, a także z polarną asparaginą i kwasem asparaginowym, są bardzo rozpowszechnione w amyloidach funkcjonalnych. Natomiast bardziej zasadowe aminokwasy, hydrofobowa leucyna, metionina i cysteina oraz polarna tyrozyna dominują w amyloidach niefunkcjonalnych. Na podstawie składu dipeptydów i wskaźników aminokwasowych opracowaliśmy model lasów losowych, który z powodzeniem przewiduje amyloidy funkcjonalne i niefunkcjonalne.

Wreszcie, opracowaliśmy bazę danych interakcji amyloidowych AmyloGraph, która gromadzi wiedzę dotyczącą tego, jak dany amyloid wpływa na inny. Znaleźliśmy interakcje pomiędzy białkami curli a innymi, 48 dla CsgA i 14 dla CsgB. Dzięki tej bazie danych możemy dowiedzieć się, jak białka amyloidowe oddziałują ze sobą i wpływają na proces agregacji.

Abstract

Amyloids are proteins associated with many clinical disorders, such as Alzheimer's, Creutzfeldt-Jakob and Huntington's diseases. Although these proteins have diverse structures, they share β-sheet structure and have a tendency to aggregate. Besides the non-functional amyloids, which can adopt various structures and are misfolding versions of normal proteins, there are also functional amyloids, which fulfill important cellular functions. Ones of them are curly proteins, CsgA and CsgB, which became the subject of this thesis.

In the framework of this dissertation, we experimentally validated using Thioflavin T (ThT) assay and Atomic Force Microscopy (AFM) our software AmyloGram for predicting amyloid proteins. The algorithm recognized experimentally confirmed peptides effectively and was resistant to overfitting. Out of 24 tested peptides, we found 16 that had inaccurate annotations in the AmyLoad database.

Using more objective motif searching, we have found five repetitive regions in CsgA and CsgB sequences. The repeating motifs in CsgA are separated, and have 21 residues and nine conserved sites, whereas those in CsgB are adjacent, and have 22 residues and seven conserved sites. The regions are characterized by a specific distribution of polar and hydrophobic residues as well as the central glycine, which breaks two β-strands in a given region. By using a more accurate sequence comparison, we discovered an additional region that is positioned before the others and shows significant sequence similarity to them and may potentially fold into β-strands.

To answer the question of how evolved CsgA and CsgB proteins, we collected more than 15,000 their homologs with conserved curlin domains. The majority of them also include the typical N-terminal signal peptide. Broad phylogenetic analyses showed that these proteins evolved predominantly in *Bacteroidota*, α-*Proteobacteria*, and γ-*Proteobacteria*. CsgA and CsgB turned out remote homologs and emerged by duplication when *gamma-Proteobacteria* diverged from α- and β-*Proteobacteria*. The homologs probably experienced horizontal gene transfer between various bacterial groups and also to fungi.

We also studied in detail CsgA and CsgB in *Enterobacterales*. Their five duplicated regions show seven and six conserved sites, respectively, including glycine, hydrophobic, and polar residues, which are crucial for folding into β-strands. The individual regions revealed various

substitution rates. CsgA regions evolved more quickly than in CsgB. Region 5 has the lowest rate of divergence, which is likely a result of the selection on interactions with region 1 of other curly molecule. We noticed high correlations in evolutionary distances in the regions, which suggests their coordinated evolution. Stronger correlations in substitutions were seen in CsgA regions, which would mean that interactions between these regions in this protein should be more conserved than in CsgB.

In addition, we purified selected CsgA and CsgB variants that had deleted regions and studied the influence of these regions on the rate of aggregation using ThT assay and AFM. We found that the process can be slowed down by region 1, whereas region 5 is essential for the polymerization of amyloid fibrils due to interactions with region 1 of other CsgA molecules.

Moreover, we searched for sequence features characteristic of functional and non-functional amyloids. Despite that these amyloids are diverse, we discovered specific traits that may be used to recognize them. Small hydroxylated amino acids, serine, and threonine co-occurring with other tiny glycine and alanine, as well as polar asparagine and aspartic acid, are highly prevalent in the functional amyloids. But more basic amino acids, hydrophobic leucine, methionine, and cysteine, and polar tyrosine dominate in the non-functional amyloids. Based on dipeptide composition and amino acid indices, we elaborated a random forest model, which successfully predicts the functional and non-functional amyloids.

Finally, we developed an amyloid interaction database AmyloGraph, which gathers knowledge regarding how a particular amyloid affects another. We found interactions between curly amyloid proteins and others, 48 for CsgB and 14 for CsgB. With the help of the database, we can find out how amyloid proteins interact with each other and influence the aggregation process.

1 Introduction

1.1 General characteristics of amyloids

Amyloid proteins are a peculiar group of proteins that demonstrate a unique ability to assembly into supramolecular filamentous aggregates (fibrils) characterized by the presence of characteristic cross-β sheets. Amyloid fibrils are highly ordered, long, straight, and unbranching, as shown by microscopy, X-ray diffraction, and crystallography studies. These fibrils are extremely resistant to degradation by proteolysis, sodium dodecyl sulfate (SDS), and other detergents [Knauer et al., 1992, Chapman et al., 2002, Toyama and Weissman, 2011].

Amyloid fibrils are usually made of subunits named protofilaments, which in most cases curl up around each other to form the mature fibril [Goldsbury et al., 1999, Jiménez et al., 2002]. Both natural and synthetic amyloid fibrils share a particular core structure. It consists of a β-sheet conformation, in which the direction of β-strand hydrogen bonds runs along the β-strand length and parallel to the fibril axis [Inouye et al., 1993, Makin and Serpell, 2005]. The β-sheet ribbons are associated via side-chain interactions that stabilize the structure [Makin et al., 2005]. The cross-β-sheet structure can be parallel or antiparallel within the protofilaments. The arrangement of such a fibril depends on the properties of the protein from which it originates [Gordon et al., 2004, Petkova et al., 2005].

The propensity of a protein to form amyloid fibrils depends on several factors, such as amino acid sequence, electric charge, and hydrophobicity. Although the vast majority of amyloid fibrils display a similarity at the secondary structural level, they show little similarity in their amino acid composition [Eisenberg and Sawaya, 2017, Iadanza et al., 2018]. To describe in detail the exact mechanism of amyloid fibril formation, we need to first define the differences between various assembly states of amyloid proteins.

1.2 States of amyloid protein assembly

1.2.1 Monomers and oligomers

A monomer is a molecule that is a basic building block of proteins. One monomer can react with other monomers in a polymerization process to create a much bigger macromolecule, an

oligomer, or a polymer. Both show a regular repeating structure, but the oligomer has a lower molecular weight than the polymer, although the boundary between their weight is arbitrary. Monomers can be divided in various ways: synthetic and natural (biopolymers), polar and non-polar, or cyclic and linear [Naka, 2014].

Amyloid oligomers are formed in the polymerization of monomers in a regular and repeating fashion. They are supramolecular structures that are often named as amorphous or soluble aggregates. They represent an intermediate form between the monomer and the amyloid fibril, which is extremely important in the formation of new filamentous aggregates by a wide variety of amyloidogenic proteins [Narayan et al., 2012, Shammas et al., 2015]. They are highly heterogeneous in size, structure, and stability. Unlike amyloid fibrils, oligomers are soluble in solutions and have different structural and functional properties [Grundke-Iqbal et al., 1986].

The stability of oligomers is supported by a wide range of interactions with each other. Typically, the interface between the units is formed by a central, adjacent hydrophobic patch surrounded by hydrophilic residues and water molecules at its periphery. In addition, many hydrogen bonds stabilize the structure. The oligomers typically contain specific structural motifs such as coiled coils, leucine zippers, and helix-loop-helix, which are responsible for the formation of α-helices. In the case of β-sheet, coiled-coil motifs dominate [Garratt et al., 2013].

1.2.2 Polymers and amyloid fibrils

Polymer is a class of very large complex compounds. They are created by the polymerization of multiple monomers (Fig. 1). When they are built of monomers of the same chemical composition, molecular weight, or structure, they are called homopolymers, whereas those derived from more than one species of monomer are named copolymers [Naka, 2014].

Amyloid fibrils have a unique structure characterized by the cross-β-sheet, where β-strands run crosswise to the main fibril axis. Each fibril consists of several protofilaments that are laterally coupled, and each protofilament consists of multiple oligomers. Distinctive features of protofilaments are cross-β structures with β-strands, which are stacked perpendicular to the long axis of the fibril. They usually have a width of 5–20 nm, polar topology, a left-handed supertwist, and a twofold helical symmetry [Riek and Eisenberg, 2016, Schmidt et al., 2016, Fitzpatrick et al., 2013, Annamalai et al., 2016, Ke et al., 2020, Iadanza et al., 2018].

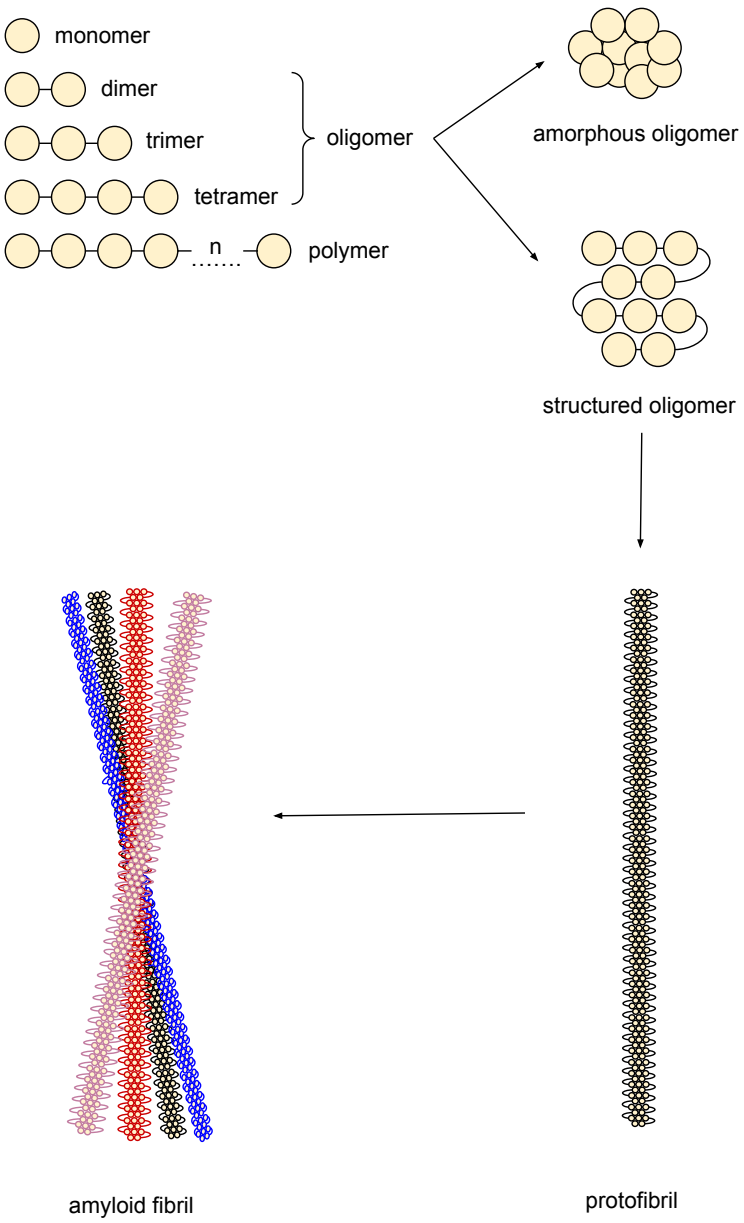

Figure 1: **Amyloid fibril organization**

Amyloid fibrils are highly polymorphic. The structural differences between them depend on the particular polypeptide chain from which they are assembled. Polymorphism of amyloid fibrils can be observed in both *in vitro* and *in vivo*. Changing environmental conditions during fibril formation result in fibril morphologies that are quite different from native ones. This makes the prediction of amyloid fibril structure much more complicated than protein folding [Meinhardt et al., 2009, Fitzpatrick et al., 2013, Annamalai et al., 2016, Kodali et al., 2010, Anfinsen, 1973].

1.3 Stages of amyloid self-assembly

Recently, the theory of amyloid fibrils self-assembly has been formulated [Hellstrand et al., 2009]. It postulates that the formation of a mature amyloid fibril has to initiate the process called nucleation, which proceeds in two stages, primary and secondary nucleation. In primary nucleation, monomers, of which the protofilaments are composed, associate into small aggregates (Fig. 2A). It is done without involving already formed aggregates [Törnquist et al., 2018]. The formation of the first nuclei from the monomers requires very high energy. It is needed to transition from their native to the amyloid state. Kinetics of the amyloid formation, for amyloid-prone proteins, are characterized by a slow, rate-limiting nucleation step [Arosio et al., 2015]. The secondary nucleation (Fig. 2B) occurs when the first amyloid-competent nuclei are already formed. This process requires at least three molecular events: I) the arrival of monomers to the surface of the fibril, II) the formation of monomers prone to aggregation, and III) the release of aggregation-prone monomers. The secondary nucleation, in contrary to the primary one, saturates at higher concentrations of monomers [Meisl et al., 2014, Törnquist et al., 2018].

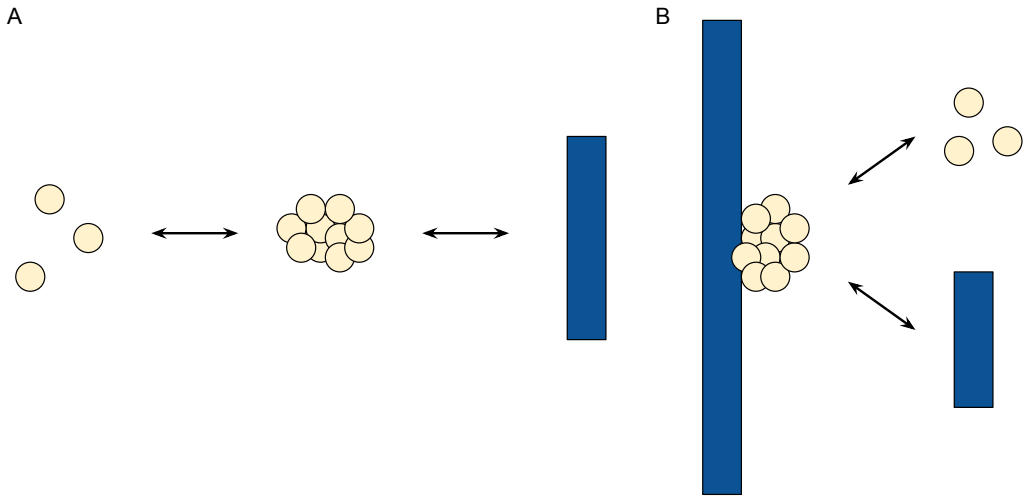

Figure 2: **The primary and secondary nucleation.** A. In the primary nucleation, monomers of one protofibril nucleate in a solution. B. The secondary nucleation involves the nucleation of monomers on the surface with an already existing amyloid aggregate. Light yellow circles symbolize monomers and oligomers, and the dark blue rectangles protofibrils.

1.4 Determinants of amyloid aggregation

One of the amyloid feature is a tendency to their aggregation. This process can compete with protein folding [Tartaglia et al., 2008, Jahn and Radford, 2008, Chiti and Dobson, 2017] because these two processes depend on the physicochemical properties of amino acid side chains. Although, we can also call this process polymerization. Aggregation is described as a non-specific process, while polymerization is described as a specific one. However, during the formation of amyloid aggregates, both processes can occur simultaneously [Frieden, 2007].

In the amyloid aggregation, their polypeptide chains might be also responsible for the propensity of molecules to aggregate. The ability of a protein to adopt a functionally specific and thermodynamically stable three-dimensional structure, as well as the transition from the unfolded state to the native conformation, is encoded in the protein's primary structure, i.e. its amino acid sequence.

This encoding of the protein structure in the amino acid sequence of the protein suggests that aggregation determinants in polypeptide chains are found not only in proteins responsible for various diseases but also in non-toxic ones, which are also able to form oligomers due to abnormal wrapping or so-called functional amyloids, which have some function in organisms.

Studies of the de Groot et al. [2005] indicated that the presence of so-called "hot spots" or protective residues are responsible for such sequence properties as hydrophobicity, a tendency to adopt a β-sheet structure. The hot-spot theory has been confirmed by analyzing aggregation-prone sequences that were devoid of defined three-dimensional conformations [Santos et al., 2020].

Hydrophobicity is one of the main forces responsible for binding both internally and externally the amino acid chains. At the same time, this force influences the formation of oligomers [Riek and Eisenberg, 2016, Durell and Ben-Naim, 2017]. This is confirmed by studies conducted by Jahn and Radford [2008] during which polar residues were swapped for non-polar ones, resulting in a higher tendency of proteins to aggregate. The secondary structure is also important in amyloid aggregation. A larger amount of β-sheet structures was detected in proteins capable of aggregation, which increases their stability by forming hydrogen bonds between the main polypeptide chains [Ventura, 2005, Kulandaisamy et al., 2017].

Hot spots are short amino acid sequences that have a high tendency to aggregate and are

responsible for protein oligomerization. They are characterized by the presence of numerous hydrophobic residues, both aliphatic (Val, Leu, Ile) and aromatic (Phe, Tyr, Trp) ones [Ventura et al., 2004]. The hot spots can be generated by just one inappropriate mutation in a protein, which can start its aggregating. One mutation will not significantly increase the overall hydrophobicity of the protein, but will significantly affect the rate of aggregation [Carija et al., 2017].

1.5 Functional and non-functional amyloids

Amyloids can be divided into two groups. One includes functional amyloids, containing proteins that can be utilized in many organisms to fulfill a variety of functions [Blanco et al., 2012, Schwartz and Boles, 2013, Balistreri et al., 2020]. The second group, non-functional amyloids are associated with various neurodegenerative diseases, caused by protein misfolding [Cooper et al., 1987, Prusiner, 1996]. Both groups share similar structural and biochemical properties. The functional amyloids are assembled by highly regulated biosynthetic pathways [Blanco et al., 2012], whereas the non-functional ones can change their conformation which leads to loss of function and is associated with many diseases.

1.5.1 Functional amyloids

The functional amyloids were detected in many bacteria, fungi, insects, plants and mammals (Tab. 1), where they fulfill crucial molecular and cellular functions[Romero and Kolter, 2014, Santos and Ventura, 2021]. The amyloid proteins produced by bacteria, in most cases, perform physiological tasks on the cell surface. They are involved in biofilm formation, adhesion, host-pathogen interactions and host cells invasion. We can distinguish here, e.g., curli proteins, which are produced by *Escherichia coli* and *Salmonella* spp. [Chapman et al., 2002, Wang and Chapman, 2008], *Pseudomonas fluorescens* FapC protein [Dueholm et al., 2010], chaplins formed by *Streptomyces* spp. [Elliot et al., 2003] and *Xanthomonas axonopodia* harpins [Oh et al., 2007]. The mechanism of biofilm formation by Fap is highly similar to that of *E. coli* curli proteins [Dueholm et al., 2013]. All of these proteins are fully functional and help the bacteria to promote multiple interactions between other them and other microbes.

HET-s protein present in fungus *Podospora anserina* is responsible for the fusion of compatible heterokaryons, i.e. multinucleate cell that contains genetically different nuclei [Wasmer et al., 2008, Turcq et al., 1991]. The process is called heterokaryon incompatibility and ensures that during spontaneous, vegetative cell fusion only compatible cells from the same colony survive.

Functional amyloids can also be found in various multicellular organisms. Hevb1 is a Rubber Elongation Factor in *Hevea brasiliensis* and takes a part in the biosynthesis of natural rubber but could be also involved in defense/stress mechanisms[Berthelot et al., 2012]. Cn-AMP2 is an antimicrobial peptide found in *Cocos nucifera* [Gour et al., 2016] and RsAFP-19, an antifungal peptide present in *Raphanus sativus* [Garvey et al., 2013]. Vicilin from *Pisum sativum L.* takes a part in detergent resistance and also displays antifungal activity [Santos and Ventura, 2021]. Spider's spidroin and silkworm fibroin are known to form insoluble silk. Fibroin is a structural element of silk, which had been successfully applied for various biomedical purposes [Zhang et al., 2012].

Functional amyloids were also identified in mammals. Pmel17 helps the maturation of melanosomes, leading to the synthesis of melanin, which protects cells against UV radiation and oxidative damage [Fowler et al., 2006]. There is also a shred of evidence that peptides and protein hormones, found in secretory granules of the endocrine system, are stored in the cross-β-sheet conformation as typical amyloids [Maji et al., 2009]. Interestingly, Rip1 and Rip3 kinases can be also considered as functional amyloids. They are involved in necroptosis, a type of programmed cell death with necrotic morphology. Motifs found in these proteins mediate the assembly of heterodimeric filamentous structures [Li et al., 2012, Liu et al., 2019].

1.5.2 Curli proteins

Curli proteins are typical functional amyloids produced by gram-negative bacteria, mostly *Enterobacteriales*, are CsgA and CsgB [Dueholm et al., 2012]. These proteins are exported outside the cell into the extracellular matrix, where they participate in biofilm formation. The biofilm also includes other proteins and polysaccharides, which protect multicellular communities from chemical and physical stresses. Living in biofilms provides benefits because biofilmic bacteria are more resistant to antibiotics and the host's immune system [Simm et al., 2014,

Jamal et al., 2015].

The main curli protein, involved in the development of biofilm scaffolds, is CsgA. It is a 151-residue amyloid protein, encoded by *csgBAC* operon. This protein consists of the 22-residue signal peptide targeting this protein outside the cell, and the amyloid core domain, whose sequence includes five non-identical repeats (R1-R5). Each of these regions contains a common, highly conserved motif (Ser-X5-Gln-X-Gly-X-Gly-Asn-X-Ala-X3-Gln) (Fig. 3) [Barnhart and Chapman, 2006, Evans and Chapman, 2014, Chapman et al., 2002]. When CsgA is incorporated into an amyloid fibril, a strand–loop–strand motif in each repeat stacks between the neighboring repeats and is stabilized by hydrogen bonds in a β-sheet. CsgA fibrils are resistant to chemical and proteolytic degradation. These fibrils can be identified by dyes, e.g., thioflavin T (ThT) and congo red (CR), binding to the amyloid. The amyloid fibrils can also be visualized under transmission electron microscopy (TEM) or atomic force microscopy (AFM) [Malmos et al., 2017b, Erskine et al., 2018].

In order to CsgA could start creating the fibrils, an initiator is necessary. Such a role is fulfilled by a nucleator CsgB, another curly protein encoded in the *csgBAC* operon. Similar to CsgA, it also contains a signal peptide, 23 amino acid residues long, and an amyloid core domain, which also includes five repeating units (Fig. 3). However, only the regions R1–R4 contain a common conserved motif (Ala-X3-Gln-X-Gly-X2-Asn-X-Ala-X3-Gln). The R5 instead contains four positively charged amino acids (one lysine and three arginines), which are absent from the other repeating units [Barnhart and Chapman, 2006, Dueholm et al., 2012, Evans and Chapman, 2014, Chapman et al., 2002, Dunbar et al., 2019].

Although each repeat shows a similarity with others, they are not functional equivalents. The repeats R1 and R5 in CsgA form amyloid fibrils and are critical to the CsgA seeding ability and the nucleation by CsgB. The internal repeats R2-R4 contain 'gatekeeper' residues that modulate the amyloid formation by softening the amyloidogenicity of CsgA. The exposure of R1 or R5 on the growing tip of a curli fibril contributes to its efficient elongation. They provide a recognition site for subsequently secreted CsgA monomers [Wang et al., 2008].

The interchangeability of CsgA regions was studied by Wang et al. [2010]. These authors used bacteria with the deletion of *csg* operon genes and complemented them with constructed plasmids. The replacement of the R1 region by R5 and *vice versa* had no major impact on the

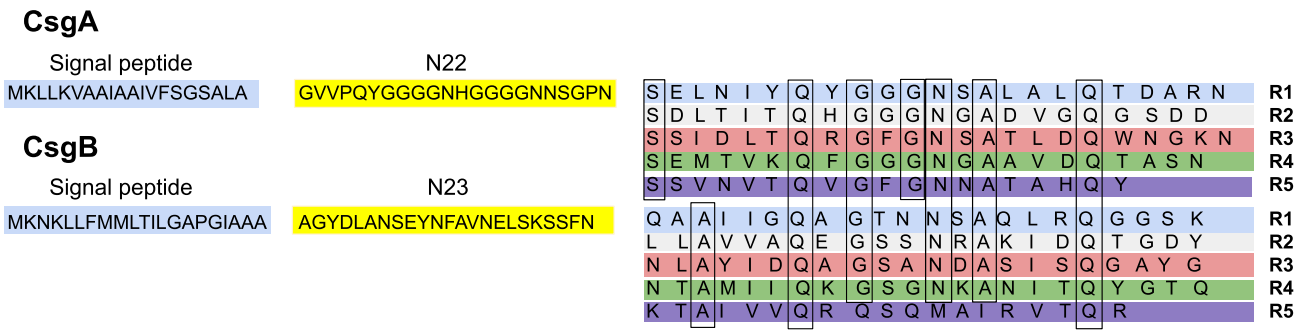

Figure 3: **Sequence organization of CsgA and CsgB proteins.** They consist of a signal peptide, a separating sequence (N22 and N23) and the non-identical repeating units (R1–R5), which were aligned. Boxed columns represent amino acids conserved throughout the repeating units in one or both proteins. Modified from Hammer et al. [2007].

formation of amyloid fibrils. In contrast, the replacement of the above-mentioned regions with R3 resulted in a decrease in this formation. Furthermore, the plasmid including gene CsgA with interchanged regions R1 and R5 could not compensate the lack of CsgA and CsgB. However, the plasmid with only R1 and R5, deprived of the gatekeeping regions (R2, R3, and R4), was able to form fibrils [Wang et al., 2008, 2010].

In the case of CsgB, the repeats R4 and R5 are responsible for the CsgA nucleation *in vivo*. The mutation in these regions caused that this protein was not localized in the outer membrane, instead, it was secreted into the extracellular matrix, but the deletion of regions R1, R2 or R3 had no impact on the nucleation [Hammer et al., 2012].

Although these experiments revealed the importance of the individual repeats in amyloidogenicity, they were conducted on selected curly proteins and bacteria, which likely do not represent their whole variation. It is also not known how their unique sequence organization originated and evolved. Current analyses of curli proteins included only the closest homologues and did not explore this subject [Dueholm et al., 2012]. The curli proteins and their distant homologues may be more widespread in the bacteria world than it is commonly assumed. Moreover, the mechanism of interaction between CsgA and CsgB is still not known in detail. Finding evolutionary conserved and variable sites in their sequences can help to determine their functional and structural significance. Therefore, we decided to study these issues in this project.

1.5.3 Non-functional amyloids

Many different proteins or peptides are known to form non-functional amyloids (Tab. 1). In a native state, they fulfill normal functions but after misfolding, they aggregate and can cause many diseases.

Some of them are involved in human disorders called amyloidoses [Baker and Rice, 2012]. These proteins are amyloid-β (Aβ) present in Alzheimer's disease, α-synuclein in Parkinson's disease, huntingtin in Huntington's disease, and β_2-microglobulin (β2M) in dialysis-related amyloidosis [Vidal and Ghetti, 2011, Sipe et al., 2016a, Knowles et al., 2014, Chiti and Dobson, 2017]. They can also occur in other types of disorders like type II diabetes [Hull et al., 2004]. Prions (PrP) cause Creutzfeldt-Jakob disease and other transmissible spongiform encephalopathies [Gibbs et al., 1968].

Fibrils of previously mentioned amyloids can accumulate in extracellular plaques, which might disrupt cellular physiology by blocking the transport of proteins and other non-protein components to the cell [Sipe et al., 2016b, Drummond et al., 2017]. Almost all proteins, which can turn into amyloids, have known functions but in their native state. When the proteins adopt the cross-β structure, it transforms the molecules into solid fibrils, causing the loss of function.

It was found that several lipoproteins, antibodies and IAPP (Islet Amyloid Polypeptide), but in the amyloid form, are not able to fulfill their primary functions, which leads to Apo-AI amyloidosis, light-chain amyloidosis and diabetes [Malmberg et al., 2020].

The most studied non-functional amyloids, which cause neurodegenerative disorders, are Aβ, PrP and α-syn. In the native state, Aβ is important for synaptic plasticity and memory [Puzzo et al., 2011]. PrP is involved in myelin maintenance and cellular proliferation processes [Legname, 2017], whereas α-syn takes a part in the regulation of neurotransmission and response to cellular stress [Benskey et al., 2016].

Table 1: **List of various proteins forming amyloid fibrils.**

Protein name	**Functional amyloid**	**Function**
AIMP2	no	takes a part in the assembly of the aminoacyl-tRNA synthase complex
Albumin	no	functions as a transporter for a diverse range of molecules, including hormones, vitamins, and enzymes
α-crystallin	no	major lens protein
α-lactalbumin	no	regulates the lactose production in milk
α-S2-casein	no	unknown function
α-synuclein	no	plays multiple roles in synaptic activity
Amyloid β	no	plays an important role in neural growth and repair
Apolipoprotein A-I	no	takes a part in the transport of cholesterol to the liver
Apolipoprotein E	no	involved in fat metabolism
β-casein	no	phosphoprotein found in milk
β-crystallin	no	structural protein of unknown function
β-lactoglobulin	yes	transport protein
β-parvalbumin	yes	involved in muscle relaxation
β2-microglobulin	no	lymphocyte surface modulator and potential regulator of the immune system
Bri2	yes	potential regulator of amyloid-β protein precursor processing
CRES	yes	involved in sperm development and maturation
CRES3	yes	might be involved in spermatogenesis
CsgA	yes	involved in biofilm formation
CsgB	yes	involved in biofilm formation
Cystatin C	no	inhibits cysteine proteases
Cytochrome C	no	involved in the electron transport chain in mitochondria and apoptosis
delta-toxin	no	lyses erythrocytes and other mammalian cells
DJ-1	no	regulates transcription and signal transduction pathways
FapC	yes	involved in biofilm formation
Fibroin	yes	core component of silk filament
FUS	no	RNA-binding proteins regulating transcription
γ-crystallin	no	major lens protein
GroES	no	inhibitor of ATP hydrolysis
HET-s	yes	involved in heterokaryon incompatibility process
IAPP	no	maintain glucose levels
Insulin	no	carbohydrates and fat metabolism regulator

Table 1: **List of various proteins forming amyloid fibrils. (Continued).**

Protein name	**Functional amyloid**	**Function**
κ-casein	no	phosphoprotein found in milk
Lysozyme	no	antimicrobial agent
Medin (AMed)	no	induces endothelial dysfunction and vascular inflammation
Myoglobin	no	serves as a reserve supply of oxygen for muscles
New1 (NU+)	yes	involved in translation, termination, and recycling
p53	no	plays an essential role in tumor suppression
p73	no	plays an essential role in tumor suppression
Pmel17	yes	mediates formation of melanosomes
Polyglutamine (polyQ)	no	stabilizes protein interactions
proSP-C	no	stabilization of the protein structure
PrP	no	receptor of β-amyloid peptide oligomers
PSMα1	yes	might be involved in biofilm structuring
PSMα2	yes	might be involved in biofilm structuring
PSMα3	yes	might be involved in biofilm structuring
PSMα4	yes	might be involved in biofilm structuring
PSMβ1	yes	might be involved in biofilm structuring
PSMβ2	yes	might be involved in biofilm structuring
Rnq1 (PIN+)	no	unknown function
S100A9	no	calcium binding proteins
Sericin	yes	joins two fibroin filaments forming a silk yarn
Serum amyloid A	no	acute-phase protein
SEVI (PAP 248-286)	yes	increase the infectivity of HIV
Sup35 (Psi+)	yes	factor of translation termination
Tau	no	plays a role in a broad range of biological processes
TDP-43	no	performs several mRNA-related processes in the nucleus
Transthyretin	no	thyroxin transport and retinol binding
Tubulin	no	forms microtubules
Vicilin	yes	involved in detergent resistance and antifungal activity

1.5.4 Prions

Prions, i.e. proteinaceous infectious particles, are a special group of non-functional amyloids. Unlike normal amyloids, their aggregation becomes self-perpetuating and infectious. Prions

can infect not only individuals among the same species but also between different species. In mammals, prions cause various progressive neurodegenerative brain disorders, known as transmissible spongiform encephalopathies [Aguzzi and Calella, 2009]. They include scrapie in sheep [Wood et al., 1992], bovine spongiform encephalopathy (BSE) in cattle [Wells et al., 1987], and Creutzfeldt-Jakob disease in humans [Gibbs et al., 1968]. However, in fungi, they play an important role in epigenetic inheritance [Chien et al., 2004]. Prions consist of PrP^{Sc}, which are abnormally folded and protease-resistant forms of the physiological cellular versions of the PrP^{C} [Bolton et al., 1982]. Most of the PrP^{Sc}, as other amyloids, are highly resistant to proteases, heat, and decontamination methods. However, some evidence shows that these properties do not correlate with infectivity because the majority of infections are associated with oligomers that are proteinase-sensitive [Aguzzi and Lakkaraju, 2016].

The aggregation of prions reassembles those in other amyloids. Highly ordered PrP^{Sc} oligomers incorporate soluble PrP^{C}. Large PrP^{Sc} fibrils can break into smaller fragments, each of which can initiate a new aggregation cycle [Cox et al., 2003].

1.6 Experimental confirmation of amyloid-like assembly

We can divide amyloid examination methods into two types, direct and indirect. The direct method is when protein content is calculated based on the analysis of amino acid residues to get the results. The indirect technique is when we infer results from other compounds or reactions, e.g. nitrogen determination or chemical reactions with functional groups within a protein. These methods are not always accurate, as they usually require protein extraction and purification for further analysis. In addition, various mathematical calculations are required to obtain a result.

1.6.1 Thioflavin T assay

One of the most commonly used methods to detect amyloid fibrils is based on the benzothiazole dye — thioflavin T [Giehm and Otzen, 2010]. It binds to cross-β-sheet structures commonly present in amyloids. The interaction of fibrils with ThT is highly specific within proteins, which makes it an excellent fluorescent probe for all known amyloidogenic proteins and peptides, regardless of their origin [Malmos et al., 2017a].

The mechanism of ThT action and the enhancement of its fluorescence upon binding to

amyloid is related to the rotational immobilization of the central C-C bond, which connects the benzothiazole and aniline rings [Srivastava et al., 2010, Voropai et al., 2003]. It is widely recognized that ThT binds to side chains along the long axis of amyloid fibrils. In addition, the binding site of ThT on the fibril surface is believed to involve at least four β-sheet subunits [Krebs et al., 2005, Wu et al., 2009, Biancalana et al., 2009].

ThT displays an enhanced fluorescence and dramatic shift of the excitation maximum (from 385 nm to 450 nm) and emission maximum (from 448 nm to 482 nm) [Biancalana and Koide, 2010, LeVine, 1993]. In 1989 Naiki et al. [1989] proved that the fluorescence emission of ThT shows a linear relationship between the amyloid fibril concentration and the emission intensity. ThT concentrations at 20-50 µM have been shown to give the highest fluorescence intensity. However, a higher concentration can affect the amyloid formation, although this is protein-dependent. The concentration of 10-20 µM is recommended for studying the kinetics of amyloid aggregation, whereas 50 µM ThT is recommended for quantifying pre-formed amyloid fibrils [Xue et al., 2017].

So far, recent studies contradict the linear relationship of ThT in the substoichiometric concentration range. This may be due to the sensitivity of ThT to self-quenching during binding to amyloid. To prevent this, excess or equimolar concentrations of ThT, whose self-quenching ceases at higher ratios, should be used. This can also result in the saturation of ThT binding to amyloid fibrils, which is less variable over time [Sulatskaya et al., 2014, Lindberg et al., 2017].

The use of ThT also has disadvantages. Thioflavin is capable of binding to DNA, cyclodextrin or SDS micelles. Moreover, in the case of amyloids, it can bind to the surface of fibrils. For this reason, this method should be used with caution, and we should be sure which compounds are in the solution to eliminate those that can co-react. The ThT emission is also affected by pH, ionic strength, buffer viscosity and type of amyloid fibrils. In addition, some small molecules may have a similar structure to ThT and compete with the binding site. ThT assay is also unable to fully detect early amyloid aggregates [Malmos et al., 2017a].

1.6.2 Atomic Force Microscopy

Atomic Force Microscope (AFM) was originally developed as a technique for surface characterization in solid material science. Nowadays, it is used in vast areas of research and became

one of the most powerful tools in biology, materials science, and nanotechnology [Adamcik and Mezzenga, 2012].

AFM is a scanning probe method that relies on the piezo-driven movement of a sharp probe tip across a sample surface, generating deflections in the cantilever attached to the probe. A topographical map is created from deflections from each scanned pixel. AFM does not rely on light or electron beams for imaging, which makes that the resolution is not limited by diffraction. Since the AFM does not need vacuum conditions to operate effectively, it has become an invaluable tool for scientists interested in studying surfaces at the nanoscale. This can provide a view on the structural and morphological characteristics of amyloid fibrils, which includes their contour, length, width, height, periodicity or high-order assembly of single protofilaments into mature fibrils [Round and Miles, 2004, Adamcik and Mezzenga, 2012, G. Creasey et al., 2012].

AFM operates using, for example, an optical detection system, which is the most common. This is due to the fact that it has a simple and robust principle of laser detection with a photodiode. It points the laser at the end of the cantilever, where the probe tip is attached and on which the photodiode is located. The laser reflects off it, depending on the movement of the cantilever, which is monitored by appropriate detectors. Thus, based on changes in the voltages on the photodiode, the direction of the cantilever can be determined [Santos and Castanho, 2004, Morris et al., 2009].

This technique can be operated in two imaging modes: contact mode (static) and tapping mode (dynamic). In the contact mode, the probe is brought into contact with the surface and then "dragged" laterally across the surface. The force between the cantilever and the surface is maintained by keeping the deflection of the cantilever constant. This causes three values to be obtained when scanning the sample, height, deflection, and friction [Ascoli et al., 1994, Santos and Castanho, 2004]. In the tapping mode, the cantilever oscillates near its resonant frequency, which is monitored for changes caused by interaction with the surface. Intermittent contact between the probe and the surface reduces the chance of damage to the probe or surface. This mode allows the detection of the values of height, amplitude error, and phase. In addition, this mode is commonly used for preliminary studies of biological surfaces due to its mild and robust operational characteristics, as it does not lead to damage or shape change of the material under study [Ascoli et al., 1994, Silva, 2005, García and Pérez, 2002, Morris et al., 2009].

AFM can also be used to determine Young's modulus. This is a factor that determines the elasticity of the material under test in tension and compression. Each material has its own characteristic value between the linear deformation and the stress that occurs in it. Amyloid fibrils are one of the stiffest biological materials known today, with Young's modulus of 3-20 GPa. In addition, amyloid fibrils show high resistance to fracture. Their ultimate strength has been shown to be on the order of 0.6 GPa, which is comparable to steel [J. Roa et al., 2011, Liu et al., 2020, Lamour et al., 2017, Smith et al., 2006, Scheibel et al., 2003].

1.6.3 Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS)

Hydrogen exchange mass spectrometry is one of the most robust analytical methods for studying protein conformations and dynamics. It monitors a hydrogen isotope exchange in the amides of the protein backbone, making this approach highly sensitive for studying protein conformation and dynamics along the entire protein backbone, except proline [Jensen and Rand, 2016]. This method was first used to study the protein structure by Zhang and Smith [1993] in 1993. It is based on an isotopic exchange of the protein under study in the excess of deuterium, followed by its fragmentation by pepsin under quenching conditions [Rosa and Richards, 1979, Englander et al., 1985]. Then, it measures changes in the content of deuterium, which is located in labile side chain groups or in the N-terminal amine group and exchanges much faster than hydrogen ions contained in the main amide chain [Jensen and Rand, 2016].

The advantage of the HDX-MS method is that it does not require covalent labeling of the protein under study, large amounts of sample, and tolerates its heterogeneity and complexity [Martens et al., 2018, Jia et al., 2020]. The disadvantage, on the other hand, is the lack of strict information about distance changes associated with conformational transitions. One can only learn about the H-bond stability of the amide backbone, determined mainly by the parameter of local structural dynamics and solvent availability [Martens et al., 2018, Vadas and Burke, 2015].

Two types of HDX can be distinguished, continuous and pulsed. The continuous HDX is the most widely used. In it, the protein under study is diluted in D_2O buffer at different times, and then deuterium uptake is measured. The increased deuterium uptake as a function of exchange time provides information about the protein's conformation [Engen, 2009, Konermann et al.,

2011]. The pulsed HDX, on the other hand, examines short time windows between incubation of the protein under study for different lengths of time. Changes in the deuterium uptake help to determine which populations of molecules are in solution [Pan et al., 2005, Zhang et al., 2013, Wang et al., 2015].

1.7 Prediction of amyloids

All predictors of the protein 3D structure are based on Anfinsen's thermodynamic hypothesis [Anfinsen, 1973]. It states that all information about the protein folding is encoded by its amino acid sequence and the protein's native state is characterized by the lowest free energy. This postulate formed the basis for the development of various computer simulations, which try to predict protein conformations using an energy-driven scoring method identifying the lowest energy state [Anfinsen, 1973, Anfinsen and Scheraga, 1975, Levitt and Warshel, 1975]. However, this assumption can be violated by many proteins that undergo aggregation and misfolding. One of them are amyloid proteins.

Over the years, a good deal of various software for amyloid prediction have been developed. Different methods or combinations have also been used to obtain good results. Currently, we can distinguish, among others, methods based on the physicochemical properties of amino acids, their order in the protein secondary structure and thermodynamic interactions between them or methods based on machine learning.

1.7.1 Structure-based methods

Structure-based approaches to predict protein amyloidogenicity are based on the protein structure as input data, taking into consideration their folding and native state. This involves the use of solvent accessibility of protein residues to estimate surface hydrophobicity. Moreover, short simulations of molecular dynamics are performed to calculate protein retention (proteins that are retained in the endoplasmic reticulum after folding) over time. However, this method might not be applied to highly dynamic proteins.

An example of this approach is Aggrescan3D 2.0 [Pujols et al., 2018, Kuriata et al., 2019], which in addition to the features listed above, also simulates changes in the protein solubility and stability upon mutation and conformational fluctuations in the amyloid aggregation. Thanks

to the extended dynamics calculations, it allows for studying larger proteins and screening for functional protein variants with improved solubility. In addition, Aggrescan3D Database is available. It contains the analyzed human proteome in terms of their aggregation properties [Badaczewska-Dawid et al., 2021].

Other programs, AggScore [Sankar et al., 2018] use the distribution of hydrophobic and electrostatic patches on the protein surface, including the intensity and relative orientation of surface patches. PASTA 2.0 is based on the data derived from the protein secondary structure to make a residue-residue contact map [Walsh et al., 2014]. On top of that, it also uses the residue-residue energy potential and scoring functions for β-sheet structure formation. It is worth noting that Aggrescan3D 2.0 and AggScore can be used to predict amyloid aggregation and globular proteins, whereas PASTA 2.0 is specific for amyloids.

1.7.2 Machine learning methods

Machine learning (ML) is a part of artificial intelligence (AI) and computer science. It focuses on the use of data and algorithms to imitate the way in which humans learn, gradually improving its accuracy. Nowadays, ML is an important part of the growing field of data science and biology.

Predictions are based on the use of one of three main classification methods:

- supervised machine learning, in which labeled datasets are used to classify data or predict outcomes;
- unsupervised machine learning, which is used to analyze and cluster unlabeled datasets;
- semi-supervised learning, which shows properties between the supervised and unsupervised ML. It uses a smaller labeled data set to guide classification and feature extraction from a larger unlabeled data set.

Almost all existing tools for the prediction of amyloid-like proteins utilize supervised learning. However, to train a non-deep supervised model, they need to transform information hidden in an amino acid sequence to a tabular or matrix format of various features because the protein sequences are not structured data. One of the possibilities is to encode the sequence using

the physiochemical and biochemical properties of amino acids. This approach is used in, e.g., AGGRESCAN [Conchillo-Solé et al., 2007] and APPNN [Família et al., 2015], which also includes the frequency of β-sheet structures. WALTZ [Lopez de la Paz and Serrano, 2004] and Zyggregator [Tartaglia and Vendruscolo, 2008] are based on identifying sequence patterns of peptides, which form amyloid-like fibrils *in vitro*. Software like TANGO [Belli et al., 2011] and BetaSerpentine [Bondarev et al., 2018] estimate the probability of amino acid sequence segment to form β-sheet structures, which mediate the protein aggregation. Another method is based on identifying amyloidogenic sequence patterns called n-grams. They are continuous or discontinuous sequences of n elements used in, e.g., Budapest Amyloid Predictor [Keresztes et al., 2021] and software developed by our team AmyloGram 1.0 [Burdukiewicz et al., 2017]. The only tool that utilizes unsupervised ML is Cordax [Louros et al., 2020b]. It clusters sequences using t-Distributed Stochastic Neighbor Embedding (t-SNE).

Table 2: **List of example software for prediction of amyloid properties of proteins and peptides.**

Software	**Size of training dataset**	**Sequence encoding**	**Model**
AGGRESCAN	57	polypeptide sequence	amino acid properties
APPNN	296	orthogonal encodings	artificial neural networks
WALTZ	213	polypeptide sequence	computational
Zyggregator	no data	polypeptide sequence	amino acid properties
Budapest Amyloid Predictor	948 and 553	n-grams	SVM
AmyloGram 1.0	1088 (6) and 1887 (6-10) and 2373 (6-15)	n-gram	random forest
TANGO	no data	polypeptide sequence	computational
BetaSerpentine	no data	polypeptide sequence	structural
Cordax	1402	hexapeptides	t-SNE clusters

1.7.3 AmyloGram 1.0

AmyloGram 1.0 [Burdukiewicz et al., 2017] is an amyloidogenicity prediction software, which uses n-grams, i.e., continuous or discontinuous sequences of n elements. They are widely used in

the analysis of biological sequences, thanks to their highly interpretable nature. These features help to identify motifs responsible for amyloidogenic properties of peptides, because the shortest motif can have a length of only six residues [de Groot et al., 2005].

The data set used to train AmyloGram was extracted from AmyLoad [Wozniak and Kotulska, 2015]. It contained 421 amyloid peptides and 1044 non-amyloid peptides. Peptides with a sequence length shorter than six and longer than 25 amino acids were removed. The former were too short and the latter too rare and diverse. The final dataset contained 397 amyloid and 1033 non-amyloid peptides.

Overlapping hexapeptides were extracted from the dataset and labeled as amyloid or non-amyloid based on the annotation in the database. Since hexamers from longer peptides may not always have amyloidogenic properties as the peptide from which they were extracted, false positive or false negative amyloid motifs could be used to train AmyloGram. To diminish this problem, the maximum length of peptides in the training set was restricted to 15 amino acids. Finally, the training set consisted of 3 groups that differed in length (6, 6-10 and 6-15 amino acid residues).

The algorithm also utilizes a reduced amino acid alphabet, which represents certain subgroups of amino acids retaining information about protein properties. As several studies show, peptide structures do not depend only on amino acid sequence, but also on their general physiochemical properties [Murphy et al., 2000]. Multiple reduced alphabets based on various combinations of physiochemical properties of amino acids were created. After cross-validation of reduced alphabets, 18,535 unique amino acid encodings, which used 17 peptide physiochemical properties, were extracted. The best-reduced alphabet consisted of six amino acid groups (Tab. 3). The selected hexapeptides were encoded using the reduced alphabet. As classification for the cross-validation, the random forest method was used, and only discriminating n-grams selected by Quick Permutation Test were considered [Burdukiewicz et al., 2017].

Table 3: **Best performing amino acid encoding used in amyloidogenicity prediction of peptides by AmyloGram 1.0.**

	Subgroup ID	Amino acids
1	I	G
2	II	K, P, R
3	III	I, L, V
4	IV	F, W, Y
5	V	A, C, H, M
6	VI	D, E, N, Q, S, T

1.7.4 AlphaFold

Another ML predictor, which has recently become extremely popular, is AlphaFold developed by Google DeepMind [Callaway, 2020]. For two consecutive iterations of CASP [Callaway, 2020], a worldwide benchmark focused on the prediction of protein structure, AlphaFold has consecutively outperformed other methods. Structures modeled by AlphaFold had more accurate domains and side chains. Moreover, it can provide estimates of predictions [Jumper et al., 2021, Senior et al., 2020].

To achieve such outcomes, AlphaFold utilizes several methods. It includes novel neural network architectures, evolutionary-based training procedures, as well as physical and geometric constraints of protein structures. The use of these techniques allows the development of a new way to jointly embed multiple sequence alignments, a new output representation enabling accurate end-to-end structure prediction, iterative improvement of predictions by the use of intermediate losses, and learning from unlabeled protein sequences. This predictor is able to predict 3D coordinates of all heavy atoms of a protein from primary amino acid sequence and aligned sequences of homologs [Jumper et al., 2021, Senior et al., 2020].

1.8 Difficulties in modeling of amyloids

While AlphaFold is a breakthrough model for the 3D structure prediction of globular proteins, it is not effective for a significant fraction of disease-associated amyloids or other aggregating proteins. They contradict Anfinsen's postulate, on which AlphaFold is based. In their

case, the assumption that the primary amino acid sequence determines the 3D structure can be invalid. The main reason for it is that these proteins contain the chameleon sequences or intrinsically disordered regions (IDRs) [Pinheiro et al., 2021]. The former are able to adopt different secondary structures despite having the same sequence. This is related to the transition between the α-helix and β-sheet structures. This mechanism is found in the prion protein (PrP) [Bahramali et al., 2016, Gendoo and Harrison, 2011, Guo et al., 2007]. The latter, on the other hand, lack a specific 3D structure while retaining their functions. IDRs exist as dynamic conformation assemblies. They have the ability to swiftly change from one conformation to another, from extended one to compact one. This property allows them to bind to many other proteins and ligands, performing diverse functions [Das and Pappu, 2013, Van Roey et al., 2014, Tompa, 2012, García-Jacas et al., 2022].

Liquid–liquid phase separation is one more thing that affects the correct prediction of amyloids. The process involves the formation of membrane-less compartments in the cell that have important physiological but also pathological functions [Banani et al., 2017, Lyon et al., 2021]. In this case, proteins undergo an aggregation process to become amyloids through a condensation pathway instead of a deposition one. Research is currently being conducted on whether the amino acid sequence affects the condensation pathway and how to predict the amyloid aggregation within condensates [Vernon and Forman-Kay, 2019, Vendruscolo and Fuxreiter, 2022].

One of the methods to overcome the difficulties in amyloid prediction was introduced by Koliński et al. [2021]. They have tested a multiscale procedure using the CABS-dock algorithm to model a highly amyloidogenic peptide arising from insulin A-chain [Kurcinski et al., 2019]. The first step of this procedure is to make multiple docking simulations using the CABS-dock algorithm. Then the models were recreated to atom representations, improved by molecular dynamics simulations, and the best models were assembled into fibrils. The obtained fibril models have been compared with experimental data from atomic force microscopy (AFM) proving that the multiscale modeling procedure is highly accurate in the prediction of amyloid protofilaments and fibrils [Koliński et al., 2021].

2 The aims of the dissertation

Since many aspects of amyloids are still unknown, the goal of this thesis was to conduct various bioinformatic and experimental analyses of amyloid proteins including functional amyloids CsgA and CsgB:

- experimental testing amyloidogenicity of peptides that were accurately predicted and incorrectly recognized by the software AmyloGram. The aim of this approach was to validate the predictor and receive a verified set of peptides for improving this algorithm in the future. Moreover, learning these experimental methods based on different amyloid peptides was necessary for further research on much longer functional amyloids, CsgA and CsgB.

- detailed bioinformatic research of CsgA and CsgB sequences showing a unique arrangement of five repeating regions. The purpose of this study was to assess how similar the duplicated units are to one another and to identify a common consensus for them.

- extensive phylogenetic analyses of CsgA and CsgB homologs. The goal of these investigations was to reconstruct the evolutionary history of these proteins and verify if the similar structural organization evolved convergently or was inherited from a common ancestor.

- studying the role of repeating regions of CsgA and CsgB in the aggregation process. The target of this investigation was to purify selected CsgA and CsgB variants with deleted regions and determine the influence of these regions on the rate of aggregation.

- comparison of functional and non-functional amyloids. The objective of these analyses was to find specific sequence features that can distinguish these types of amyloids and can be used in their prediction based on a machine learning model.

- building a database of amyloid interactions. The intention of this subject was to gather information about the interaction of various amyloid proteins including functional ones, e.g. CsgA and CsgB, as well as designed definitions and descriptors of these interactions.

The thesis was divided into several parts relevant to these subjects including in each of them the section of Research objectives, Material and Methods as well as Results. The subjects were jointly discussed in the section Discussion at the end.

3 Amyloid peptide validation

3.1 Research objectives

The aim of this research objective was to evaluate experimentally the performance of AmyloGram, a software for amyloidogenicity prediction. To achieve that, we have selected 10 amyloids correctly predicted by this algorithm and 24 peptides that were incorrectly predicted according to initial assumptions. Then, the peptides were experimentally verified using Thioflavin T assay and Atomic Force Microscopy. The goal of this research was also to learn these experimental techniques and develop appropriate protocols on the base of various amyloid peptides. It was necessary for further studies of much longer functional amyloid CsgA and CsgB proteins described in the next sections.

3.2 Materials and Methods

3.2.1 Peptide selection

In order to validate AmyloGram 1.0 prediction algorithm, we have chosen 3 sets of data (Fig. 4). The first set consisted of 10 peptides, which were predicted in accordance with the annotations in the AmyLoad database. The other two sets included 12 false positive and 12 false negative peptides. To select these peptides, we downloaded all hexapeptide sequences from AmyLoad. After splitting them into two separate sets, amyloidogenic and non-amyloidogenic, we cross-checked our sets to eliminate sequences that occurred in both groups. The peptide sequences were encoded using the reduced alphabet with 6 amino acid groups (Tab. 3). The encoding resulted in the occurrence of identical sequences, which were removed from the final set. Hexapeptides that were also present in AmylHex database [Fernandez-Escamilla et al., 2004] were removed because AmylHex includes experimentally validated peptides that were used to check if AmyloGram is working correctly after the learning phase. The amyloidogenicity of the selected peptides were predicted by AmyloGram, which provided probability values. Based on these values and AmyLoad annotations, we selected correctly (Tab. 4) and incorrectly (Tab. 5), that were predicted opposite to the annotations in the AmyLoad database.

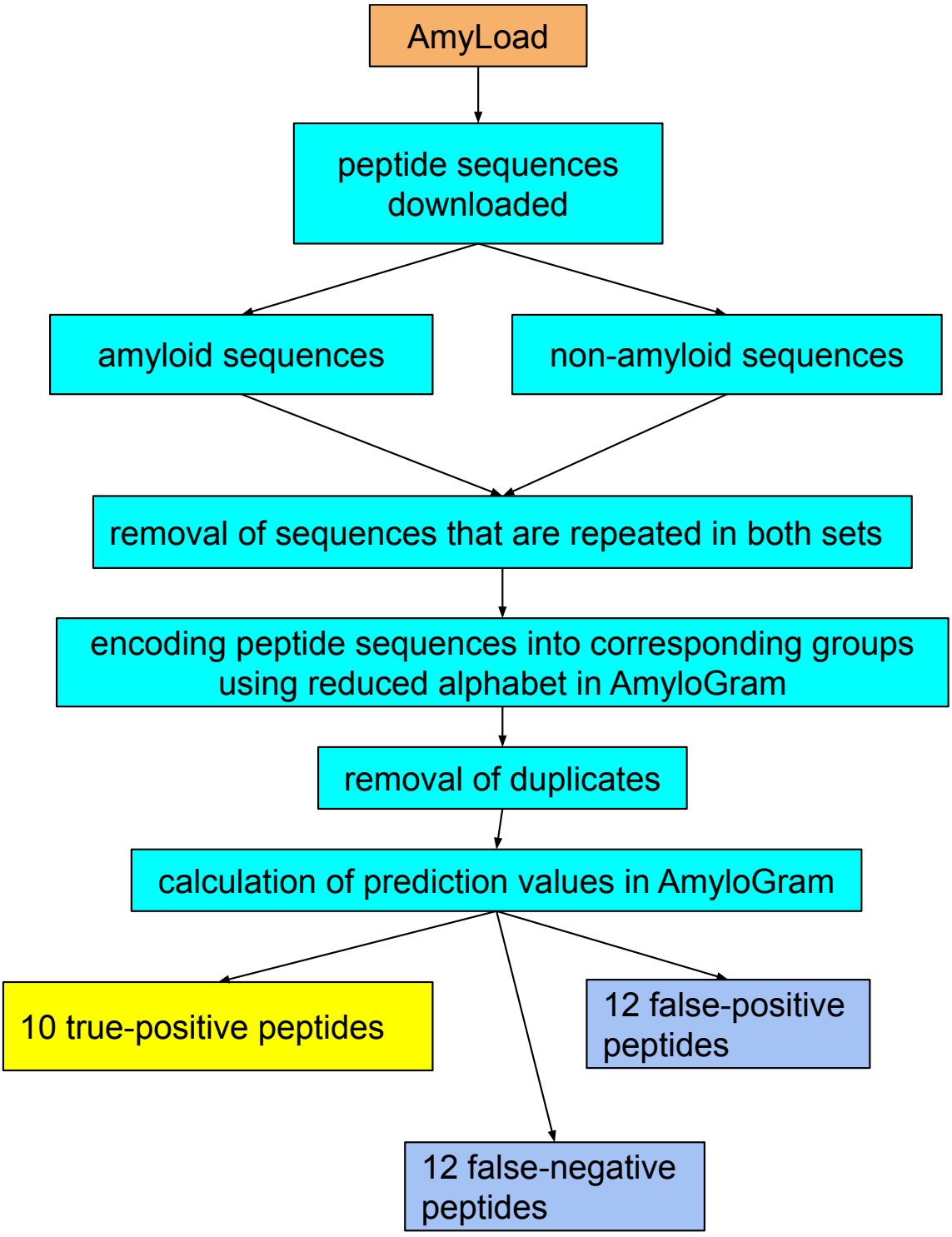

Figure 4: **The scheme of peptide selection for AmyloGram improvement.**

Table 4: **Reference peptides that were correctly predicted by AmyloGram algorithm according to annotations in AmyLoad database.**

	Peptide	AmyLoad	AmyloGram	Probability
1	SFLIFL	amyloid	amyloid	0.9157
2	ISFLIF	amyloid	amyloid	0.9132
3	YLLYYT	amyloid	amyloid	0.9124
4	LVFYQQ	amyloid	amyloid	0.8876
5	YTVIIE	amyloid	amyloid	0.9176
6	KPAESD	non-amyloid	non-amyloid	0.0005
7	FNPQGG	non-amyloid	non-amyloid	0.0023
8	NPQGGY	non-amyloid	non-amyloid	0.0023
9	TKPAES	non-amyloid	non-amyloid	0.0024
10	SWVIIE	non-amyloid	non-amyloid	0.6161

The selected peptides for experimental verification were synthesized *de novo* by an external company. To be certain that they do not aggregate during this process, we dissolved them using NaOH because an increase in alkalinity disturbs the tertiary structure of proteins. After few seconds, we neutralized pH and measured fluorescence intensity using ThT assay.

Table 5: **Peptides that were incorrectly predicted by AmyloGram algorithm according to annotations in AmyLoad database.**

No.	Peptide	AmyLoad	AmyloGram	Probability
1	NNSGPN	amyloid	non-amyloid	0.0138
2	QANKHI	amyloid	non-amyloid	0.0434
3	QEMRHF	amyloid	non-amyloid	0.0519
4	MMHFGN	amyloid	non-amyloid	0.0528
5	ALEEYT	amyloid	non-amyloid	0.0687
6	HGFNQQ	amyloid	non-amyloid	0.0790
7	ASSSNY	amyloid	non-amyloid	0.0880
8	HSSNNF	amyloid	non-amyloid	0.0880
9	MIENIQ	amyloid	non-amyloid	0.0984
10	NIFNIT	amyloid	non-amyloid	0.1244
11	MIHFGN	amyloid	non-amyloid	0.1375
12	HLFNLT	amyloid	non-amyloid	0.1441
13	STVVIE	non-amyloid	amyloid	0.8627
14	ELNIYQ	non-amyloid	amyloid	0.8216
15	FTFIQF	non-amyloid	amyloid	0.8093
16	WSFYLL	non-amyloid	amyloid	0.7741
17	YYTEFT	non-amyloid	amyloid	0.7184
18	NTIFVQ	non-amyloid	amyloid	0.7013
19	DETVIV	non-amyloid	amyloid	0.6726
20	FTPTEK	non-amyloid	amyloid	0.6655
21	FQKQQK	non-amyloid	amyloid	0.6655
22	FGELFE	non-amyloid	amyloid	0.6547
23	SHVIIE	non-amyloid	amyloid	0.6449
24	STTIIE	non-amyloid	amyloid	0.6366

3.2.2 Thioflavin T (ThT) assay

ThT stock (Sigma, product no. T3516) was dissolved in MilliQ water and filtered through 0.22 µm filter to make the stock of 10 µM solutions. The 250 µl of prepared ThT solution was added to 50 ml of 50 mM phosphate buffer with pH = 7. The final concentration of the ThT buffer was approximately 50 mM. For the measurement of ThT fluorescence in the presence of amyloid fibrils, 90 µl of ThT buffer was mixed with 10 µl of the protein solution.

The protein samples were measured just after the preparation and each day for 14 days in the case of reference peptides and 4 days in the case of doubtful peptides. Depending on the primary results, we measured them many times. The ThT fluorescence emission spectrum was measured at room temperature at 480 nm using the 440 nm excitation wavelength on Cary Eclipse Fluorescence Spectrophotometer (Agilent Technologies). Each sample was measured at least 3 times. Intensity curves were normalized using the protein median value and the standard deviation for the peak at 480 nm. For the ThT control, we used 90 µl of ThT buffer mixed with 10 µl MilliQ water. We assumed that the peptide is amyloidogenic if its fluorescence intensity is twice that of the ThT control.

3.2.3 Atomic Force Microscopy (AFM)

For AFM experiments, the peptide electric charge was checked using ProtParam [Walker, 2005]. Mica, the surface on which the samples are investigated, is charged negatively. In order to increase the adhesion to the mica, some acid has been added to the negatively charged peptides. It should change its electric charge to positive and improve the interaction with the surface. Peptide solution with the concentration of 20 µl was pipetted onto the freshly etched mica surface and incubated for 10 min, rinsed with 1 ml of MilliQ water, and dried under gentle airflow. AFM images were recorded in the Tapping-in-Air mode at the drive frequency of approximately 300 kHz using a Dimension Icon (Bruker) scanning probe microscope system. Aluminum reflective coated tips Tap300Al-G (BudgetSensors) were used as a probe [Šneideris et al., 2015]. Although AFM is the most time-consuming procedure, it is the most reliable in concluding whether a peptide forms amyloid fibrils or not.

3.3 Results

Part of the results obtained under this dissertation were published in Szulc et al. [2021].

3.3.1 ThT assay

First, ThT assay was performed on 10 reference peptides to validate the proper functioning of AmyloGram 1.0. Five of them were amyloidogenic and five non-amyloidogenic according to

their annotations in the AmyLoad database [Wozniak and Kotulska, 2015]. The fluorescence spectra intensity of 10 peptides in the respective days of incubation are presented in Fig. 5.

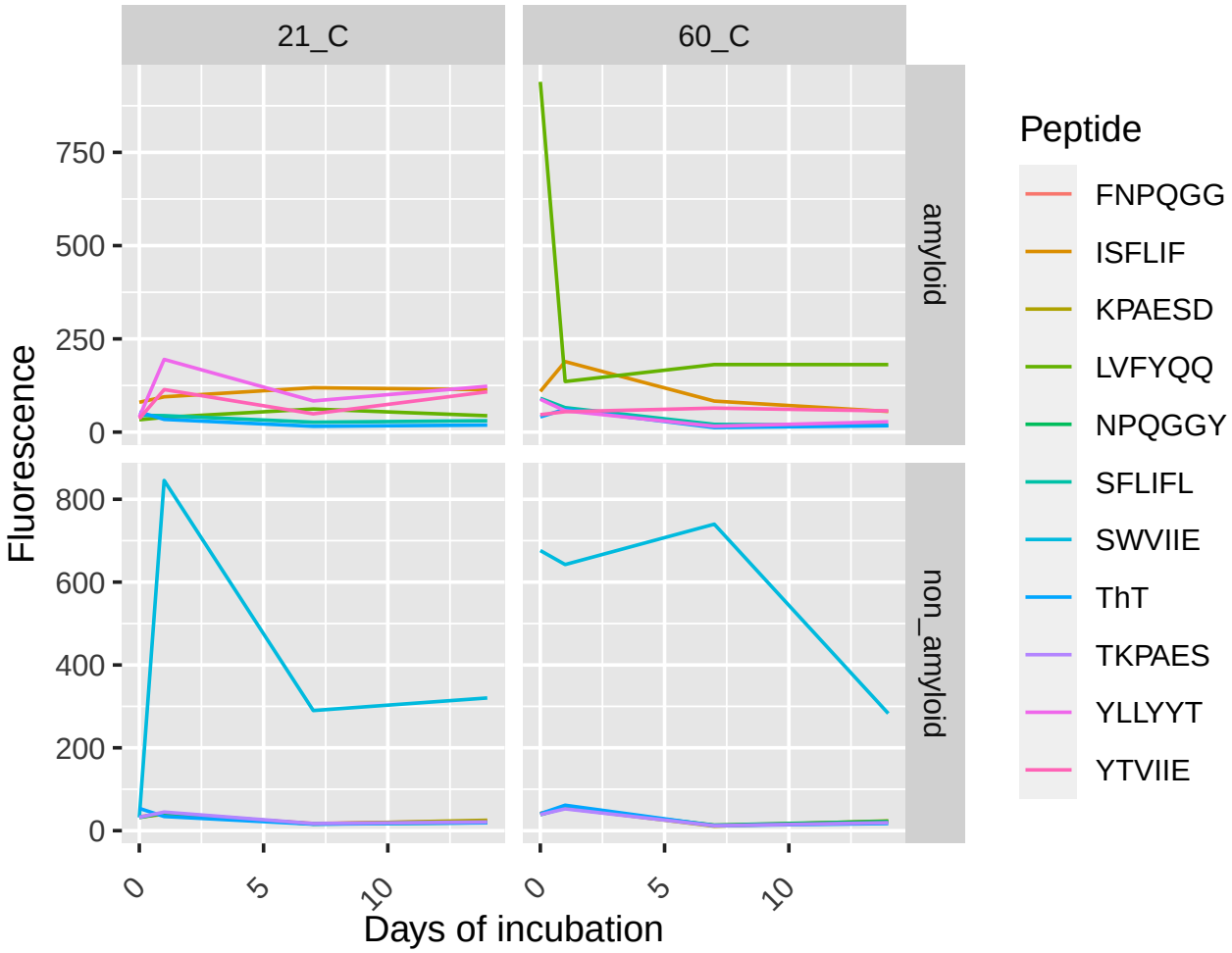

Figure 5: **Aggregation kinetics of reference amyloid and non-amyloid peptides in the presence of ThT dye in time.** Day 0 represents the measurement right after the sample preparation. Samples were incubated in 21°C and 60°C. Plots were divided according to peptide amyloidogenicity annotations given by AmyLoad and temperature.

Values for amyloids were mostly higher than for the ThT control, whereas those for non-amyloids were very close or below the control values. Generally, the fluorescence intensity decreased with time. The temperature has no important influence on fluorescence. The highest intensity indicating effective binding to ThT was shown by peptide LVFYQQ. Interestingly, among the non-amyloid peptides, SWVIIE revealed a very high binding property to ThT. It can be considered a false positive amyloid because it does not form fibrils but only oligomers when studied under AFM. The oligomerization leads to higher binding of ThT dye.

The results for 24 incorrectly predicted peptides are presented in Fig. 6. Peptide samples were measured right after preparation and for 4 consecutive days. Several peptides showed the highest fluorescence values just after preparation and next the values decreased.

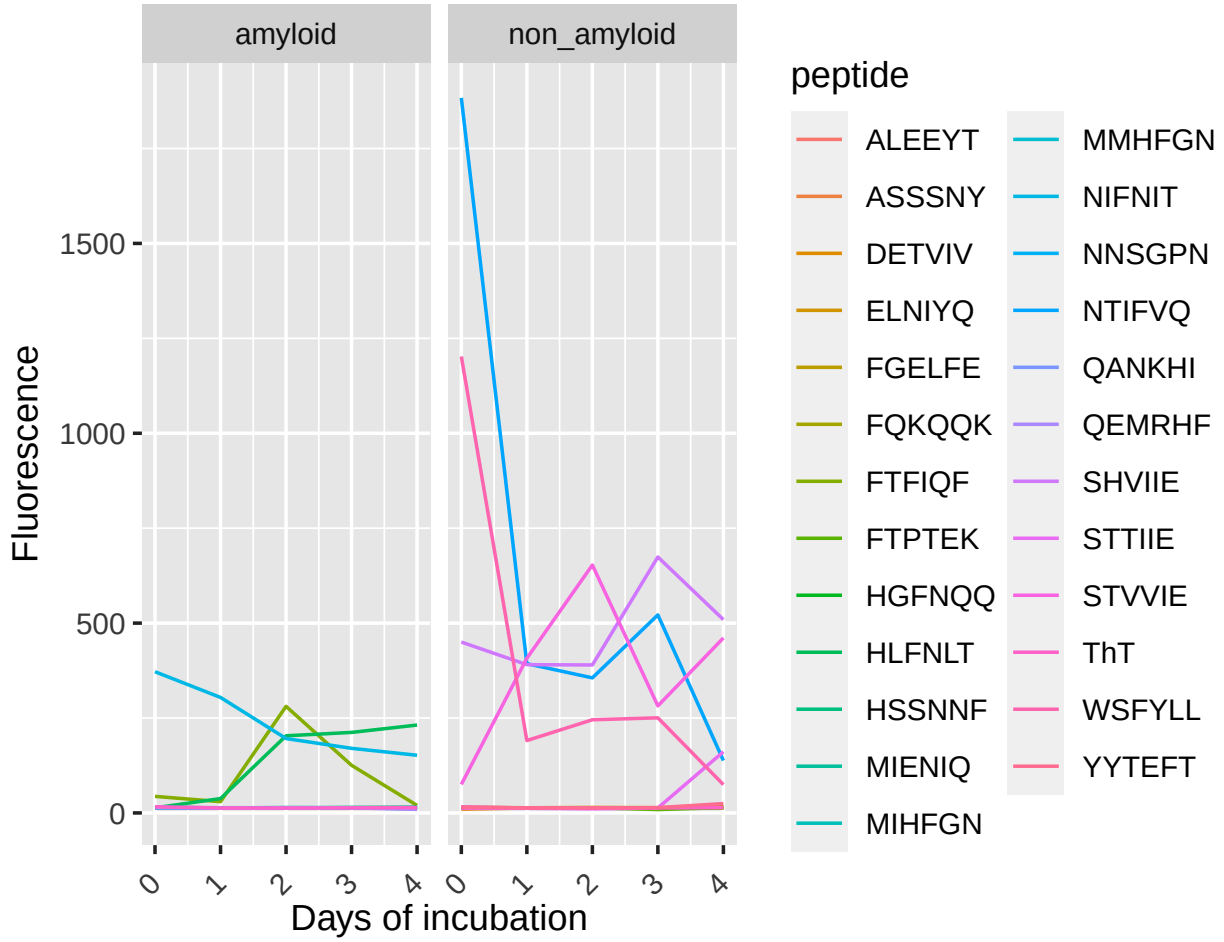

Figure 6: **Aggregation kinetics of incorrectly predicted peptides in the presence of ThT dye in time.** Plots were divided according to peptide amyloidogenicity annotations given by AmyLoad.

We have found 4 peptides annotated in the AmyLoad database as non-amyloids and predicted by AmyloGram as amyloids, which showed very high fluorescence values. Three amyloids from the AmyLoad database and predicted by AmyloGram as non-amyloids revealed also very high values (Fig. 7).

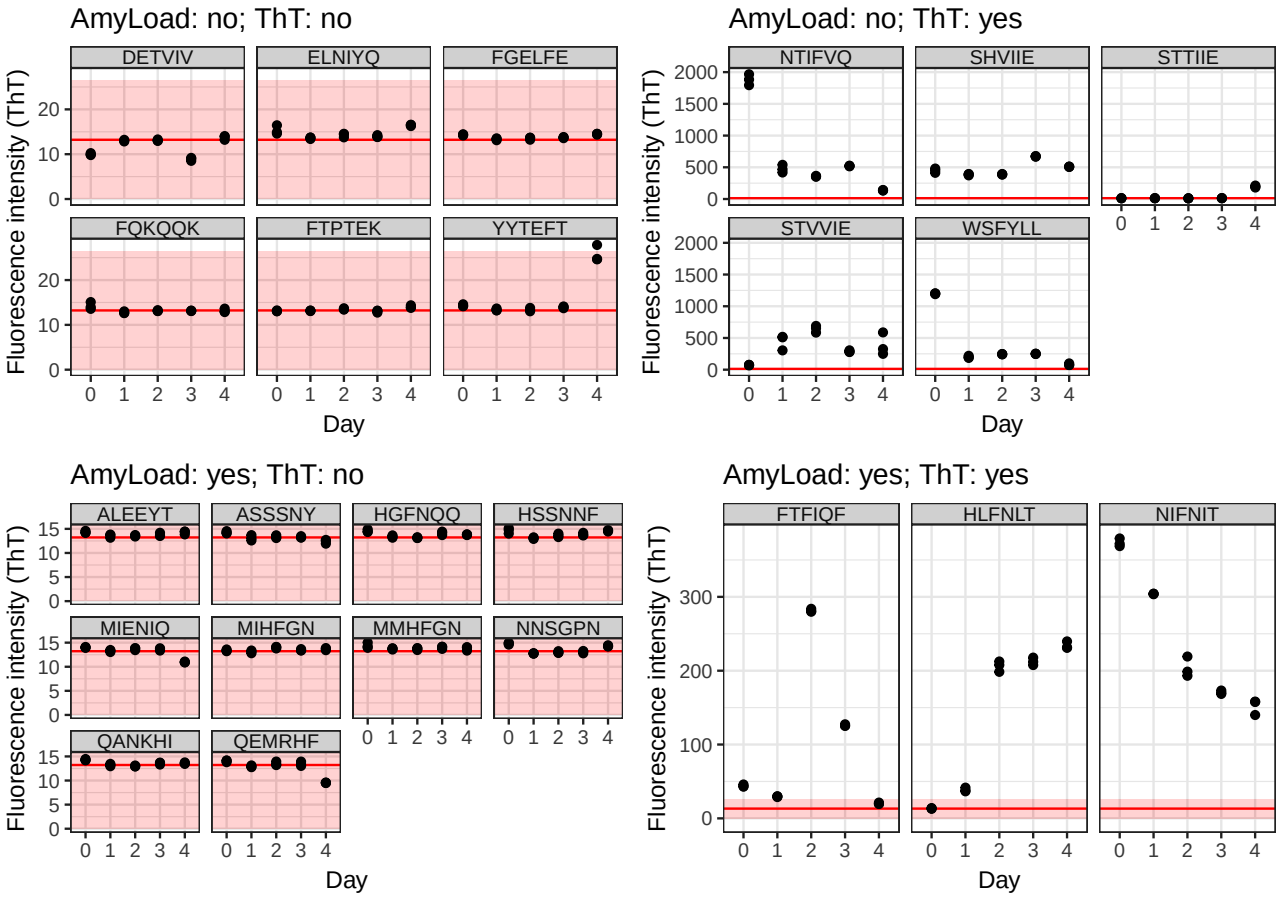

Figure 7: **Fluorescence intensity of 24 peptides incorrectly predicted by AmyloGram.** Plots were divided according to peptide amyloidogenicity annotations given by the AmyLoad database and the result of the ThT assay. The red line indicates ThT control fluorescence intensity. The light red area indicates values that exclude ThT binding to the peptide. AmyloGram prediction is opposite to that of AmyLoad.

Due to very high-intensity values for some peptides in the presented plots (Fig. 6), the values for other peptides are not easy to compare with the control. Therefore, we demonstrated fluorescence intensities for each peptide separately and summarized the results in Fig. 7. The study indicates that six peptides predicted as amyloids by AmyloGram did not bind ThT in agreement with the annotation in AmyLoad as non-amyloid. On the other hand, ten peptides predicted as non-amyloids by AmyloGram did not show the amyloiodgenicity in the ThT assay in contrast to the AmyLoad annotation. Five peptides bounded ThT were also computationally predicted as amyloids, whereas three other peptides also interacting with ThT were not predicted as amyloids by AmyloGram in opposition to AmyLoad annotations.

3.3.2 AFM

The results of Atomic Force Microscopy for reference peptides are presented in Fig. 8-14. All typical amyloid peptides should form long, thin amyloid fibrils such as that in Fig. 8 and 9. Non-amyloid peptides should not form any aggregates, as shown for example in Fig. 10. The exception is peptide SWVIIE, which formed oligomeric aggregates (Fig. 11) binding also to ThT.

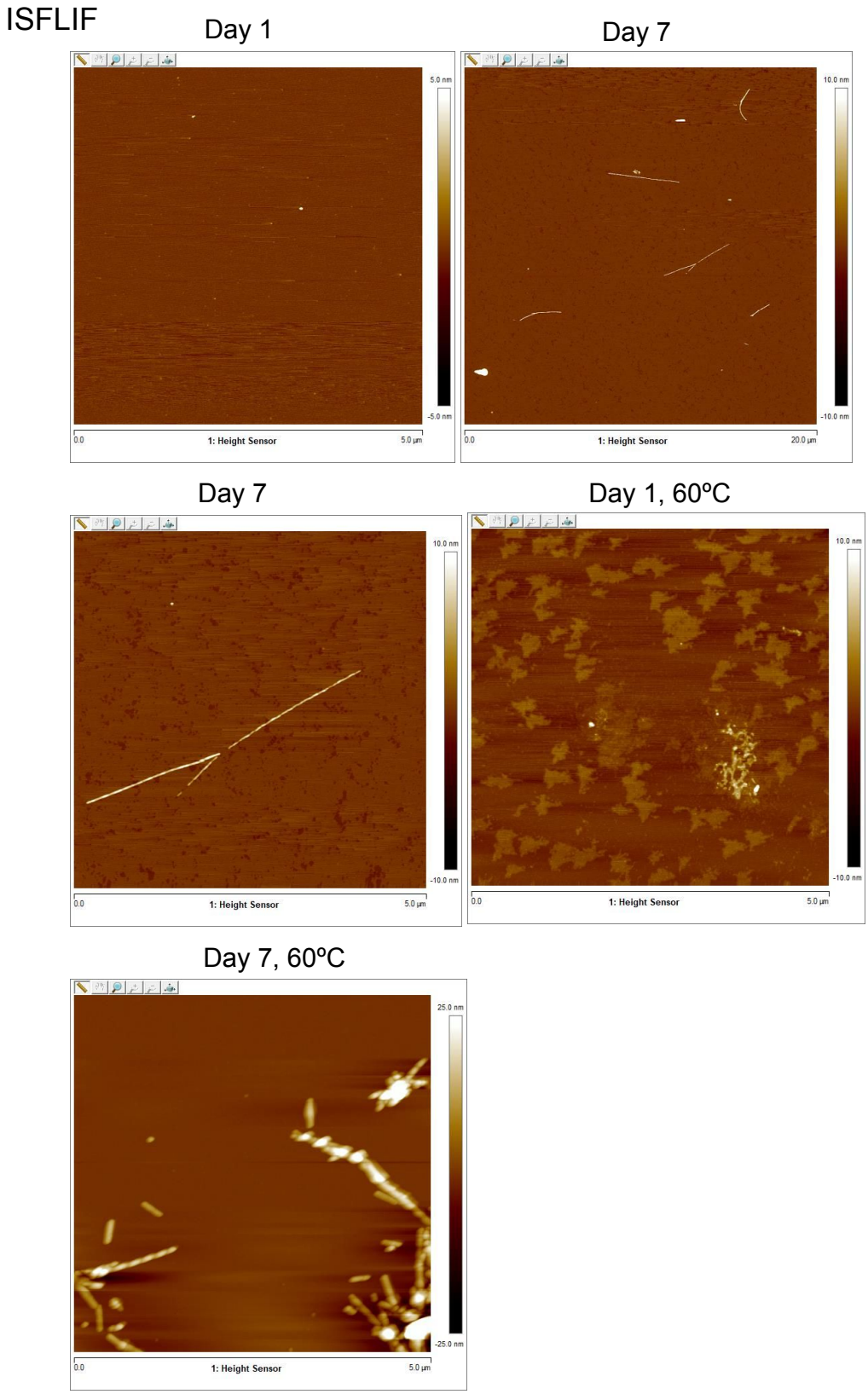

Figure 8: **ISFLIF peptide under AFM.** AmyLoad: amyloid, AmyloGram: amyloid.

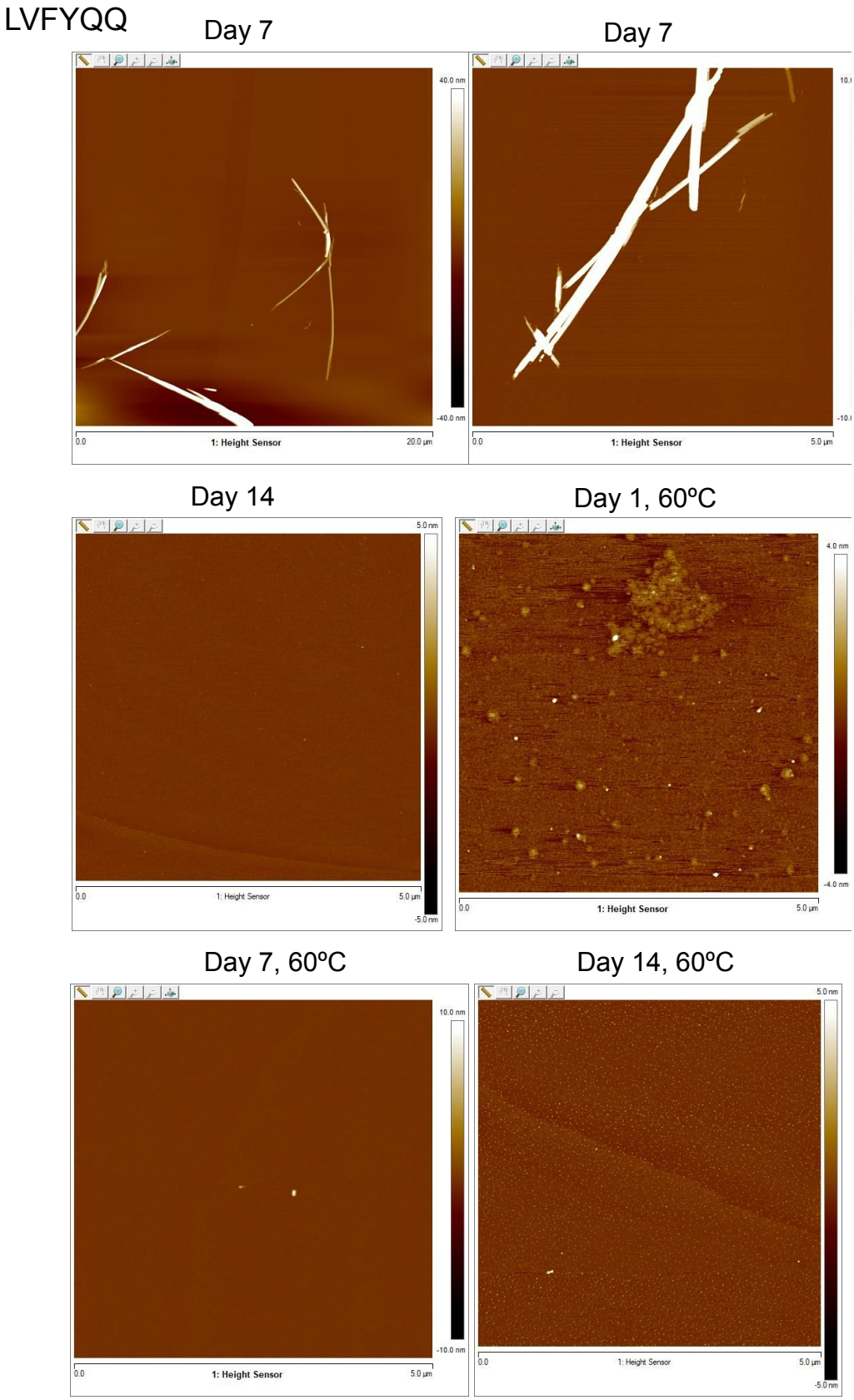

Figure 9: **LVFYQQ peptide under AFM.** AmyLoad: amyloid, AmyloGram: amyloid.

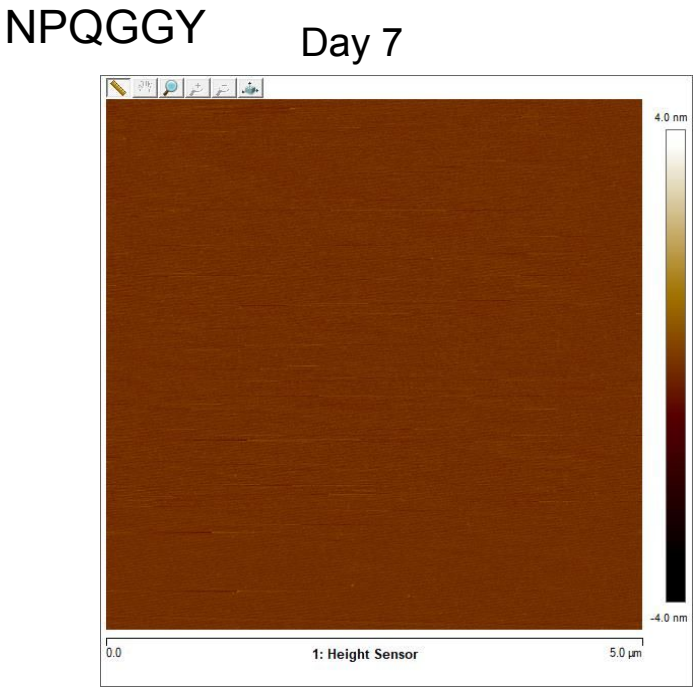

Figure 10: **NPQGGY peptide under AFM.** AmyLoad: non-amyloid, AmyloGram: non-amyloid.

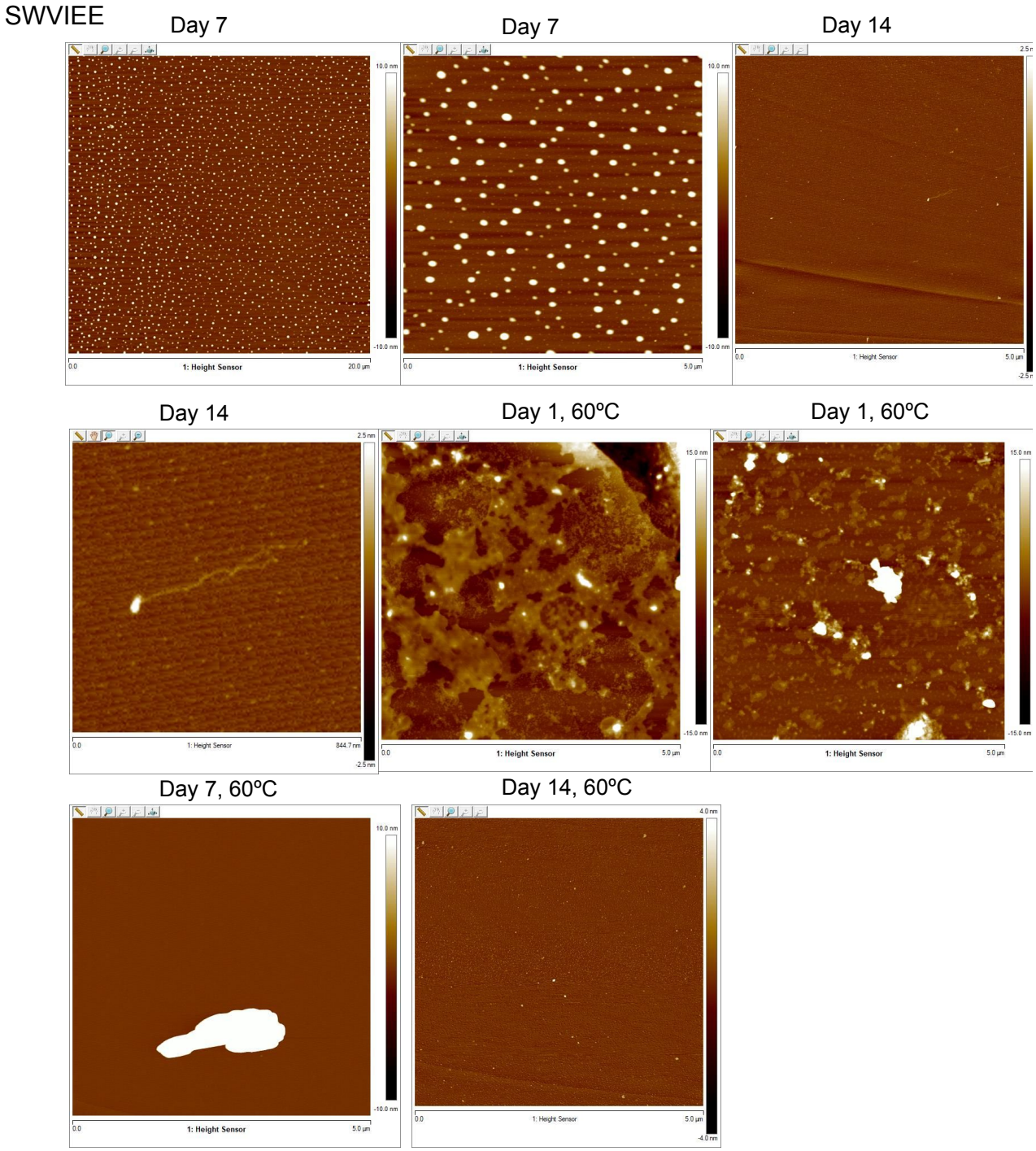

Figure 11: **SWVIEE peptide under AFM.** AmyLoad: non-amyloid, AmyloGram: non-amyloid.

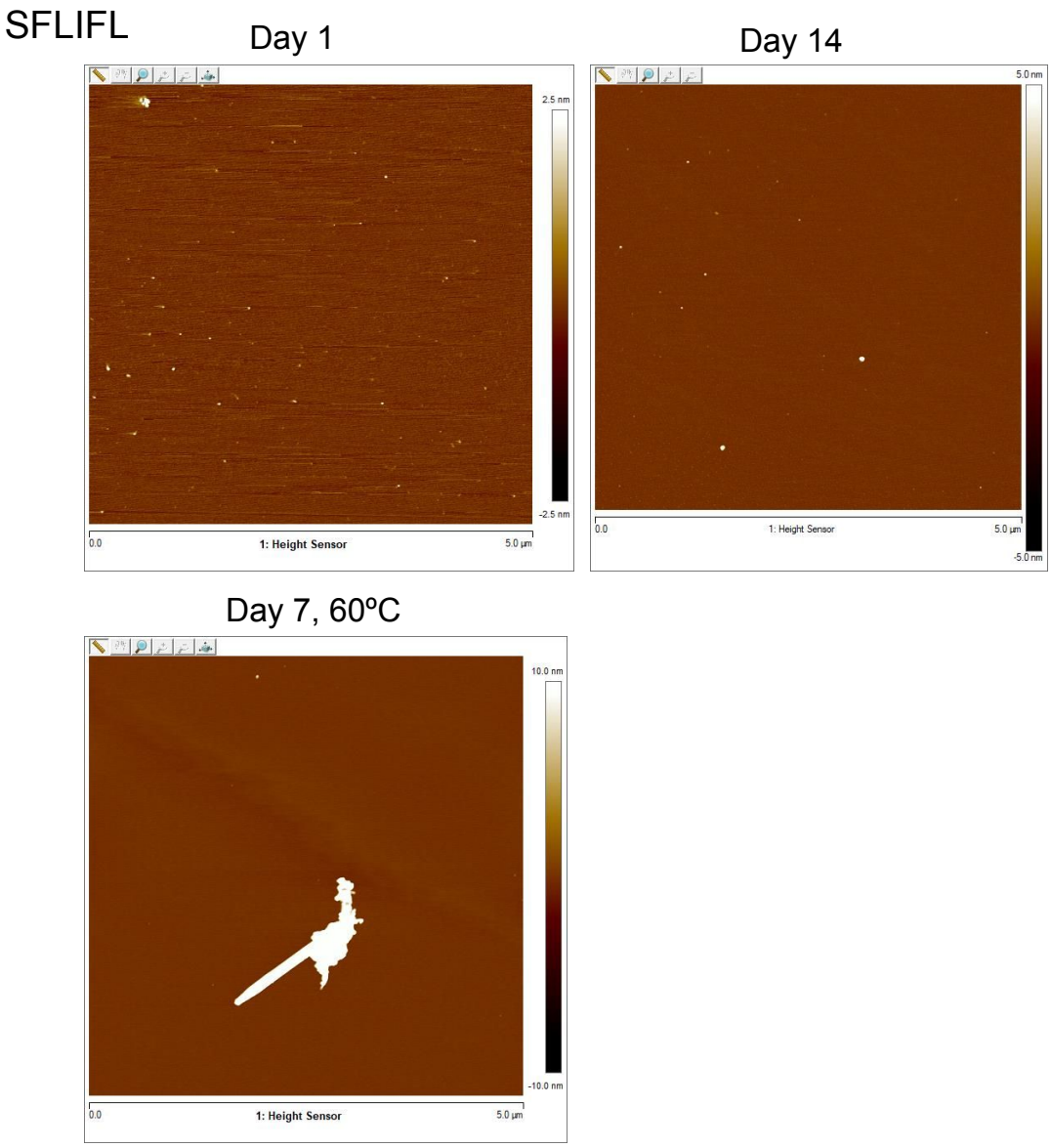

Figure 12: **SFLIFL peptide under AFM.** AmyLoad: amyloid, AmyloGram: amyloid.

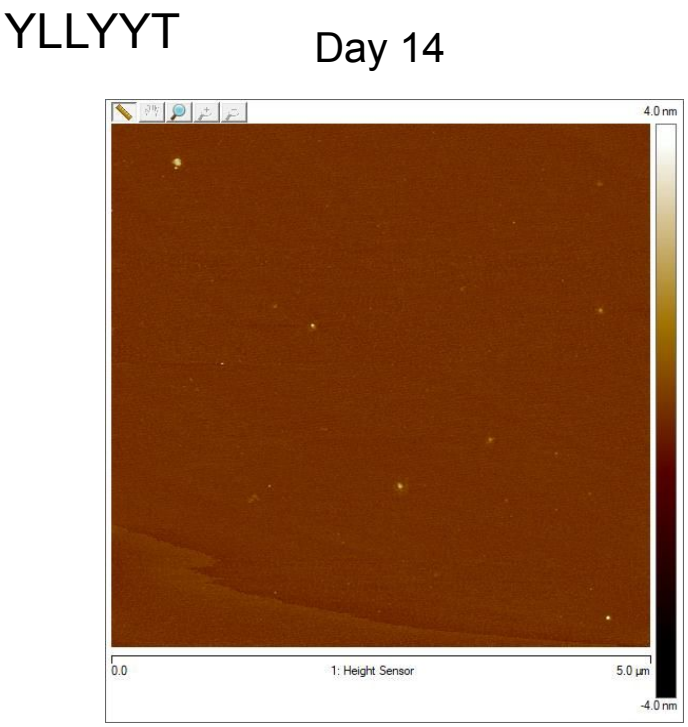

Figure 13: **YLLYYT peptide under AFM.** AmyLoad: amyloid, AmyloGram: amyloid.

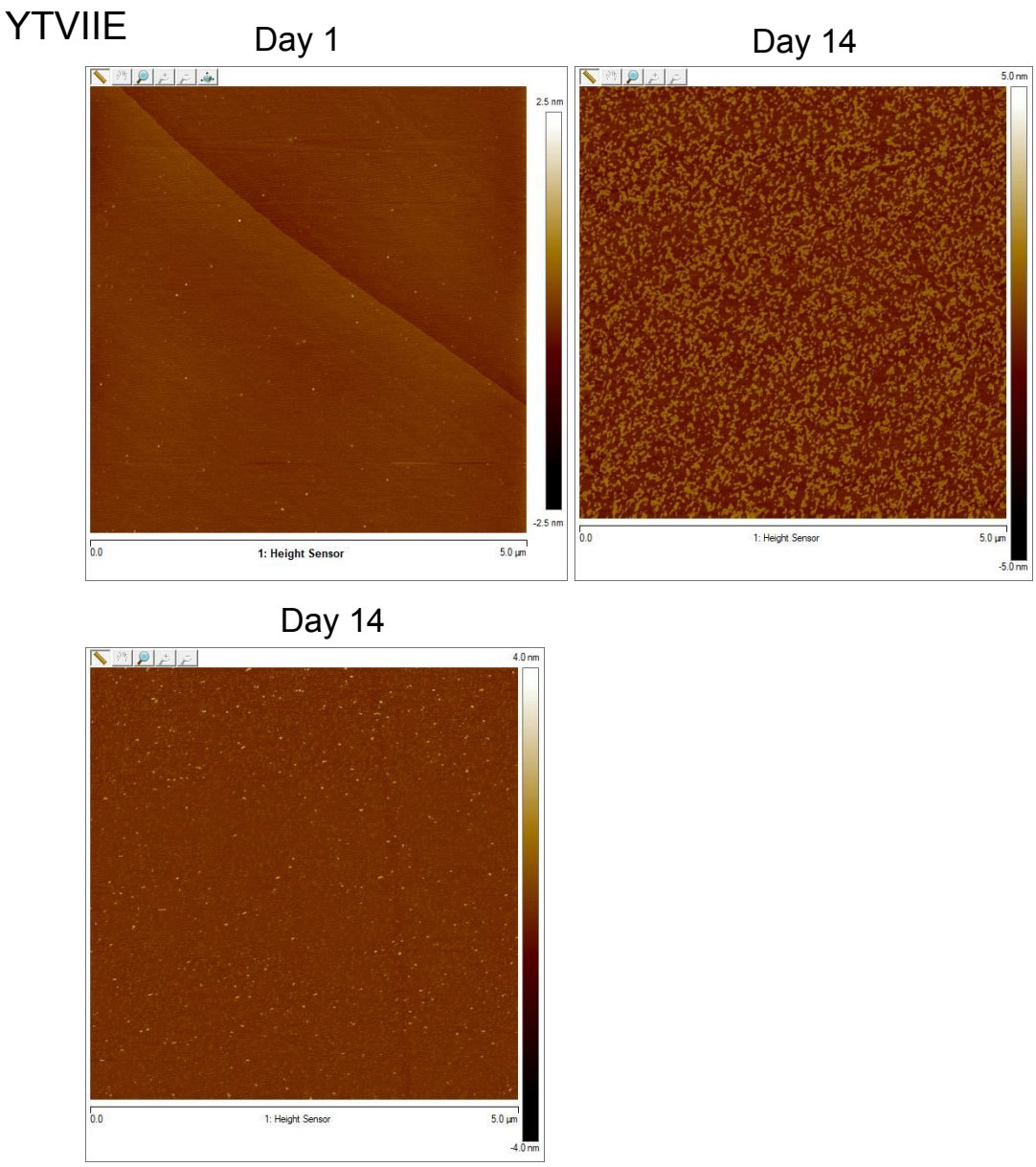

Figure 14: **YTVIIE peptide under AFM.** AmyLoad: amyloid, AmyloGram: amyloid.

3.3.3 Validation of results

The final results from the experimental validation of AmyloGram are collected in Tab. 6 and 7. In the case of reference peptides, we can conclude that AmyloGram made correct predictions (Tab. 6). All predictions were confirmed in most experiments.

Table 6: **Validation of amyloid propensities of reference dataset by various methods.**

Sequence	**AmyLoad**	**ThT**	**AFM**	**AmyloGram**
SFLIFL	amyloid	amyloid	non-amyloid	amyloid
ISFLIF	amyloid	amyloid	amyloid	amyloid
YLLYYT	amyloid	amyloid	non-amyloid	amyloid
LVFYQQ	amyloid	amyloid	amyloid	amyloid
YTVIIE	amyloid	amyloid	non-amyloid	amyloid
KPAESD	non-amyloid	non-amyloid	-	non-amyloid
FNPQGG	non-amyloid	non-amyloid	-	non-amyloid
NPQGGY	non-amyloid	non-amyloid	non-amyloid	non-amyloid
TKPAES	non-amyloid	non-amyloid	non-amyloid	non-amyloid
SWVIIE	non-amyloid	amyloid	non-amyloid	non-amyloid

At least one experimental method confirmed the computational results. SWVIIE peptide predicted as non-amyloid and showing a high-intensity peak in ThT assay did not form fibrils under AFM. Unfortunately, in the case of SFLIFL, YLLYYT, YTVIIE peptides (Fig. 12, 13 and 14), we did not find fibrils on mica using AFM. It indicates that this method does not always provide clear findings in the case of amyloids. In the case of amyloids, which were predicted by AmyloGram contrary to the annotations in the AmyLoad database, ThT assay confirmed the predictions in 16 out of 24. The results of the peptide verification can be used to modify the learning stage and initial peptide classification of the updated version of AmyloGram and the software for the prediction of functional amyloids to which belong CsgA and CsgB.

Table 7: **Validation of amyloid propensities of incorrectly predicted peptides by various methods.**

Sequence	**AmyLoad**	**ThT**	**AmyloGram**
NNSGPN	amyloid	non-amyloid	non-amyloid
QANKHI	amyloid	non-amyloid	non-amyloid
QEMRHF	amyloid	non-amyloid	non-amyloid
MMHFGN	amyloid	non-amyloid	non-amyloid
ALEEYT	amyloid	non-amyloid	non-amyloid
HGFNQQ	amyloid	non-amyloid	non-amyloid
ASSSNY	amyloid	non-amyloid	non-amyloid
HSSNNF	amyloid	non-amyloid	non-amyloid
MIENIQ	amyloid	non-amyloid	non-amyloid
NIFNIT	amyloid	amyloid	non-amyloid
MIHFGN	amyloid	non-amyloid	non-amyloid
HLFNLT	amyloid	amyloid	non-amyloid
STVVIE	non-amyloid	amyloid	amyloid
ELNIYQ	non-amyloid	non-amyloid	amyloid
FTFIQF	non-amyloid	amyloid	amyloid
WSFYLL	non-amyloid	amyloid	amyloid
YYTEFT	non-amyloid	non-amyloid	amyloid
NTIFVQ	non-amyloid	amyloid	amyloid
DETVIV	non-amyloid	non-amyloid	amyloid
FTPTEK	non-amyloid	non-amyloid	amyloid
FQKQQK	non-amyloid	non-amyloid	amyloid
FGELFE	non-amyloid	non-amyloid	amyloid
SHVIIE	non-amyloid	amyloid	amyloid
STTIIE	non-amyloid	amyloid	amyloid

4 Computational analyses of CsgA and CsgB sequences

4.1 Research objective

Sequences of CsgA and CsgB show an interesting organization, characterized by the presence of peculiar five repeating units. Therefore, we decided to investigate in detail the sequence features and organization of CsgA and CsgB proteins. We planned to evaluate the similarity between the duplicated regions and find common consensus sequences for them. To achieve that, we conducted a relevant bioinformatic analysis based on motif finding and aligning sequences (Fig. 15).

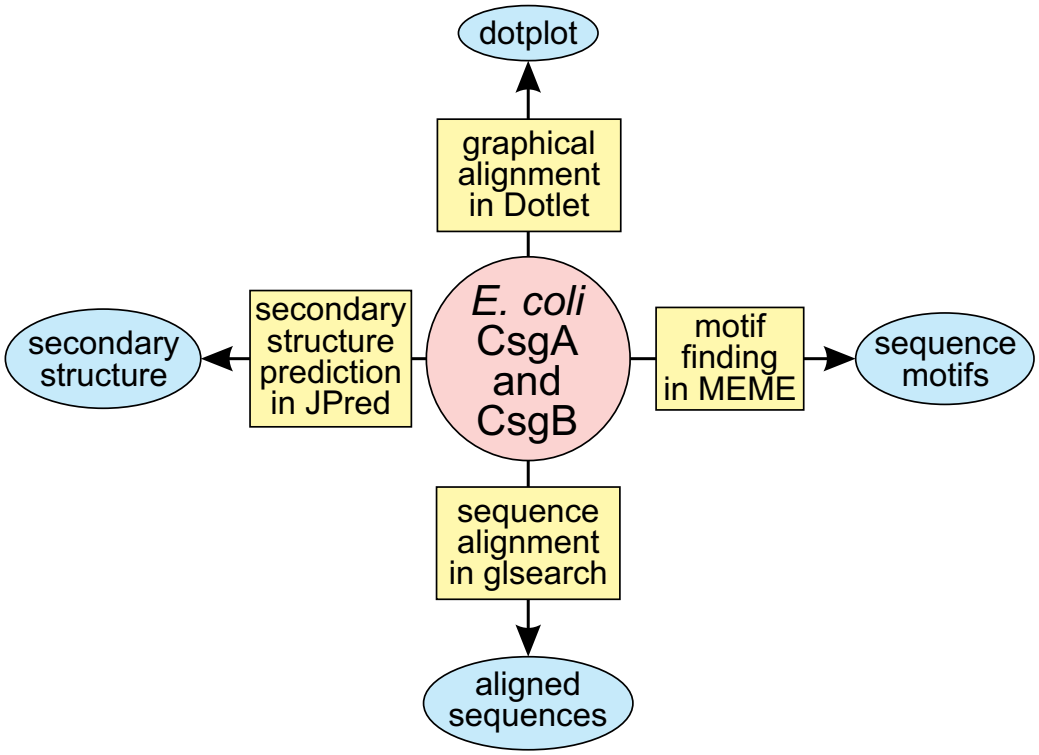

Figure 15: **Flowchart of computational analyses of CsgA and CsgB sequences.**

4.2 Materials and Methods

4.2.1 Motif finding

Motifs in CsgA and CsgB sequences were searched using MEME (Multiple Expectation maximizations for Motif Elicitation) Suite 5.5.1 [Bailey et al., 2015] using default settings. This

algorithm is dedicated for ungapped motif search and is based on several methods, which include the expectation-maximization (EM) algorithm, EM-based heuristics, maximum likelihood ratio-based heuristics as well as multi-start and greedy search [Bailey and Elkan, 1994]. It applies statistical modeling to automatically choose the best width, number of occurrences, and characteristics for each motif. Using the EM, MEME searches for motifs in provided sequences iteratively modifying the search parameters and found tentative motifs.

4.2.2 Aligning sequences

Graphical pairwise alignments of dotplot type for CsgA and CsgB sequences were performed with Dotlet at the website https://dotlet.vital-it.ch/ assuming the window size of 13 and the scoring matrix Blosum 62. The dot plot is a graphical method for comparing two sequences and identifying regions of close similarity. One sequence is presented on the x-axis and another on the y-axis of the plot. When the residues of the compared sequences match at the same position on the plot, a dot is drawn at the corresponding location. If there are many adjacent dots, they arrange into lines in the plot. In the study, we aligned CsgA or CsgB sequence with itself to demonstrate a potential similarity between the duplicated regions.

Alignments between new potentially duplicated regions and those already determined were conducted using the optimal global:local affine Needleman-Wunsch algorithm (glsearch) from FASTA package version 36.3.8g [Pearson et al., 1997]. This algorithm is more sensitive and provides the statistical significance of alignments. We assumed the number of shuffle 1,000,000 and tested all scoring matrices. Finally, we selected the alignments that showed an E-value smaller than 0.05 and the lowest for the set of matrices.

4.2.3 Secondary structure prediction

The secondary structure was predicted using JPred [Drozdetskiy et al., 2015] via JalView [Waterhouse et al., 2009]. This software uses the Jnet algorithm based on a neural network to make more accurate predictions. In addition to the protein secondary structure JPred also makes predictions on Solvent Accessibility and Coiled-coil regions using Lupas method. For the sequence for which the prediction is made, homologs are searched in UniProt database [UniProt Consortium, 2018] and next a sequence profile is constructed for the prediction.

4.3 Results

4.3.1 Analysis of CsgA and CsgB sequence organization

Both CsgA and CsgB sequences contain five non-identical repeating units participating in amyloid fiber formation [Barnhart and Chapman, 2006, Dueholm et al., 2012, Evans and Chapman, 2014, Chapman et al., 2002]. They were recognized by general sequence similarity and the identification of only identical residues. The motifs were not statistically evaluated, either. This approach could be subjective, so we decided to apply a more objective motif search using a MEME algorithm [Bailey et al., 2015] dedicated to this purpose.

The analysis revealed the presence of a common motif with E-value 2.7e-047, which is repeated five times in CsgA sequence (Fig. 16). The individual motifs are separated by one or two amino acid residues and are also significant with p-value $\leq$ 3.17e-17. The consensus motif is 21 residues in length and is characterized by at least nine conserved sites. In the middle, there are three glycine residues, whereas on both sites there are polar asparagine, glutamine, and serine. In the right part of the consensus, there is also a conserved alanine. The 5th, 7th, and 18th positions in the motif are also conserved and occupied by only hydrophobic amino acids.

In the case of the CsgB sequence, the MEME algorithm also discovered a motif with E-value 4.2e-027 repeating five times but with a length of 22 residues (Fig. 17). The individual motifs are adjacent and separated by no residue. They are significant with p-value $\leq$ 2.19e-13. The consensus motif includes at least seven conserved sites. In the middle, there is also a dominant glycine but in contrast to CsgA only one. To the left of it is conserved polar glutamine and to the right polar serine, asparagine and glutamine. Only hydrophobic residues are present in the 8th, 17th and 19th positions in the motif. It can add that two regions (2 and 5) start with glycine and two others (1 and 3) end with this amino acid, which is visualized also in the consensus motif.

The sequence consensus of CsgA and CsgB motifs share common features. Both have the central glycine surrounded by polar and hydrophobic residues, occurring alternatively. They also contain two glutamine residues in similar positions, as well as conserved asparagine and alanine separated by only one less conserved site. It may indicate a common structural organization

and common evolutionary history of these regions. The same analysis aimed to find common motifs between CsgA and CsgB sequences but did not produce statistically significant results.

Start	*p*-value	A motif site with the 10 flanking letters on either side
41	9.11e-20	NHGGGGNNSG PNSELNIYQYGGGNSALALQT DARNSDLTIT
64	3.17e-17	NSALALQTDA RNSDLTITQHGGGNGADVGQG SDDSSIDLTQ
86	7.92e-19	GNGADVGQGS DDSSIDLTQRGFGNSATLDQW NGKNSEMTVK
109	1.92e-19	NSATLDQWNG KNSEMTVKQFGGGNGAAVDQT ASNSSVNVTQ
131	2.22e-17	GNGAAVDQTA SNSSVNVTQVGFGNNATAHQY

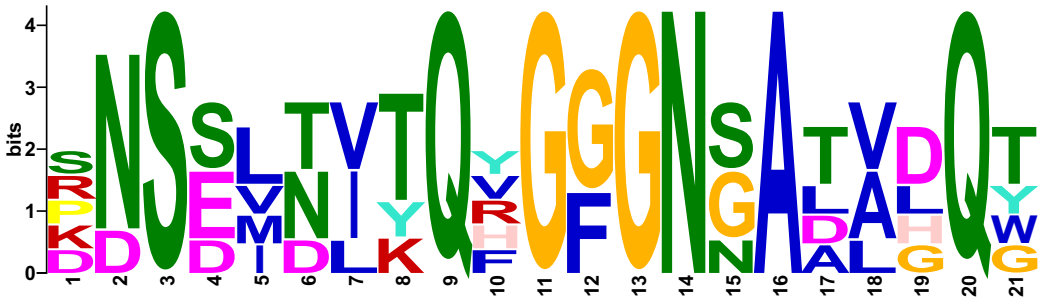

Figure 16: **Motifs discovered by MEME algorithm in CsgA sequence.** The location of individual motifs and the logo sequence of the consensus motif were presented.

Start	*p*-value	A motif site with the 10 flanking letters on either side
42	2.19e-13	NFAVNELSKS SFNQAAIIGQAGTNNSAQLRQG GSKLLAVVAQ
64	1.79e-17	TNNSAQLRQG GSKLLAVVAQEGSSNRAKIDQT GDYNLAYIDQ
86	9.20e-17	SSNRAKIDQT GDYNLAYIDQAGSANDASISQG AYGNTAMIIQ
108	1.23e-16	SANDASISQG AYGNTAMIIQKGSGNKANITQY GTQKTAIVVQ
130	4.42e-15	SGNKANITQY GTQKTAIVVQRQSQMAIRVTQR

Figure 17: **Motifs discovered by MEME algorithm in CsgB sequence.** The location of individual motifs and the logo sequence of the consensus motif were presented.

4.3.2 Searching for a new duplicated region in CsgA and CsgB sequences

To visualize the similarity between the five repeating regions, we aligned separately the same sequence of CsgA or CsgB with itself (Fig. 18). In this case, we should expect a symmetric plot with four series of short lines above and under the diagonal. These lines should correspond to matches between the appropriate duplicated regions arranged in columns and rows of the plot. The number of these lines should decrease from the diagonal to the vertices of the square plot. In fact, such a pattern can be recognized for the CsgA sequence and also for CsgB, but not all matching lines are clearly visible due to poor sequence similarity between some regions (Fig. 18). Interestingly, after the detailed investigation of the CsgA alignment, we identified some additional lines that suggest a similarity between a fragment located before the determined regions and the known duplicates. These lines are indicated by arrows in Fig. 18 and suggest the presence of an additional duplicated region. In fact, a potential new region includes stretches of glycines and polar residues present in the determined motifs. Thus, we performed a more sensitive analysis to verify the similarity between the regions.

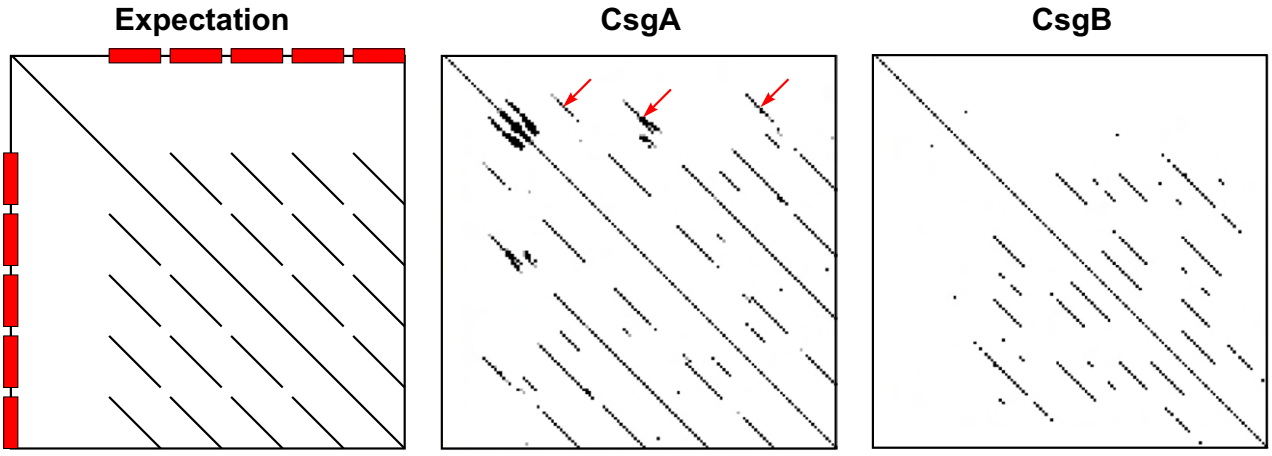

Figure 18: **Dotplot (graphical pairwise alignment) of CsgA and CsgB sequences, as well as the expected result for five duplicated regions.** Red arrows indicate a similarity between a sequence located before the determined regions and some already approved regions.

In agreement with the dot plot results, we found a statistically significant similarity between the sequence named N1 located between 17 and 31 residues (SALAGVVPQYGGGGN) and the five known repeating regions. Searches of N2 sequence with the location 32-40 (HGGGG NNSG) occurred also statistically significant. E-value for the produced alignments was from

0.038 to 0.00001 and identity between 30.8% to 83.3% (Tab. 8).

Table 8: **Results of glsearch between CsgA N1 (17-31) and N2 (32-40) as well as CsgB N (15-41) sequences against the five repeating regions (R1-R5).** Statistical significance (E-value) and percent (%) of identity, the used scoring matrices were included.

Protein	**Query**	**Subject**	**% identity**	**E-value**	**Matrix**
CsgA	N1	R1	53.9	0.00003	MD20
CsgA	N1	R2	58.3	0.0005	MD20
CsgA	N1	R3	30.8	0.034	OPT5
CsgA	N1	R4	46.2	0.00001	VT80
CsgA	N1	R5	40.0	0.0068	BL50
CsgA	N2	R1	71.4	0.000041	MD10
CsgA	N2	R2	83.3	0.015	VT10
CsgA	N2	R3	33.3	0.0089	MD40
CsgA	N2	R4	44.4	0.0047	BL80
CsgA	N2	R5	44.4	0.038	BL80
CsgB	N	R3	30.0	0.001	P120

In the alignments, we can recognize identical matches between asparagine, glutamine, glycine, histidine, leucine, serine, and tyrosine (Fig. 19). In the case of CsgB we found only one significant match (E-value = 0.001, 30% identity) between the sequence AGYDLANSEYNFAVNELSKS (placed between 15 and 41 residue) and the region R3. These sequences shared homologous positions of tyrosine, leucine, alanine, asparagine and serine (Fig. 19).

These findings suggest that initially, at least six regions could exist in the curli proteins. The five stayed more conserved and the one degenerated. It is not inconceivable that the amyloid fibrils were created by the six stretches of β-sheet in some ancestral forms.

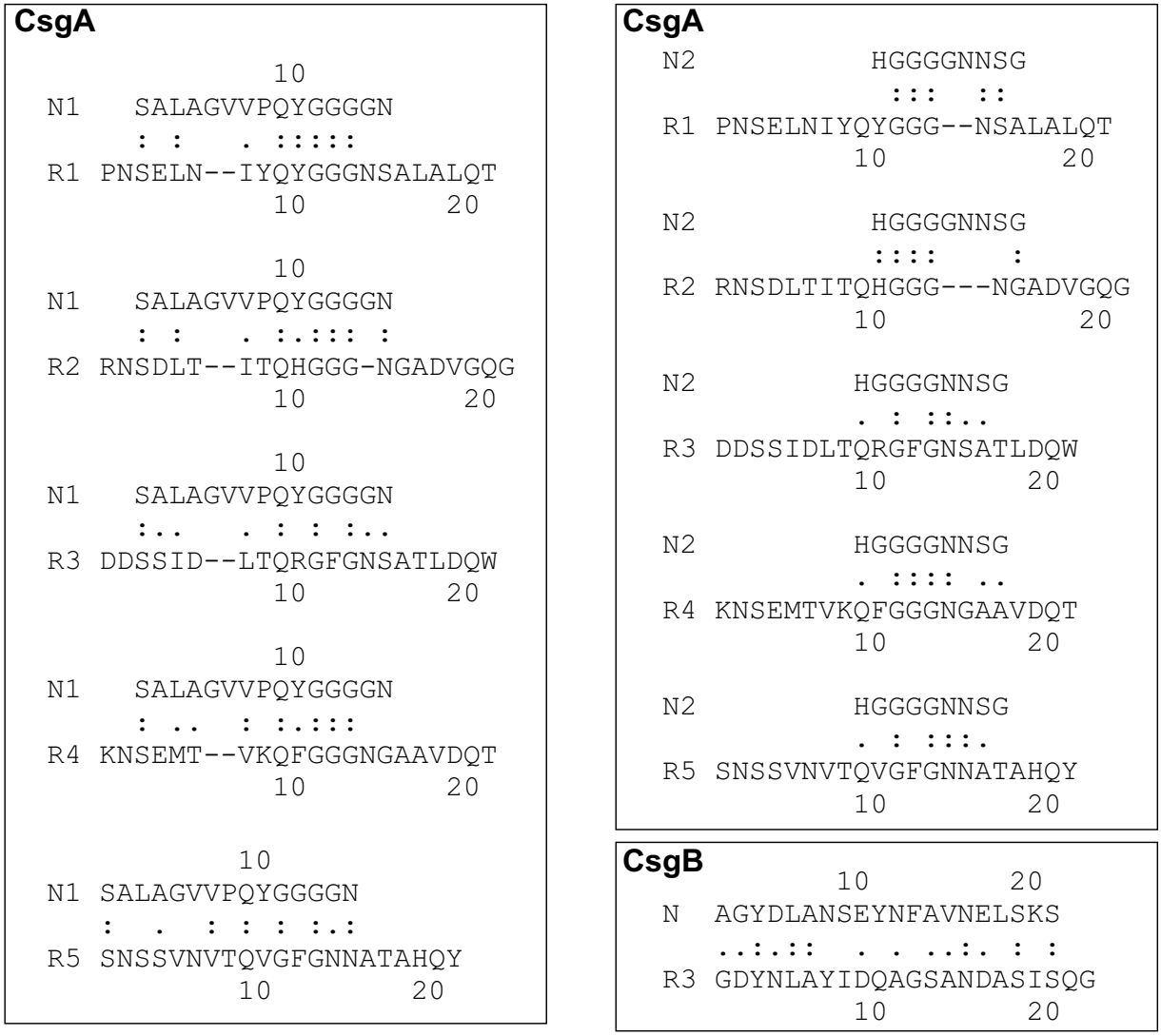

Figure 19: **Statistically significant alignments found for N1 (17-31) and N2 (32-40) CsgA sequences, as well as N (15-41) CsgB sequence with the five known repeating regions (R1-R5).** The alignments were produced in glsearch from FASTA package.

4.3.3 Prediction of the secondary structure of CsgA and CsgB

Using JPred, we predicted the secondary structure in CsgA and CsgB sequences (Fig. 20). The applied algorithms made concordant predictions. The analyses showed that each of the repeated regions consists of two β-strand, which are interrupted in the middle of the given region. In most cases, the second strand in the regions ends with their boundaries, whereas the first strand begins several residues later. Interestingly, in the fragment before the region R1 and between a signal peptide, β-strands were also predicted. This fragment includes sequences

that showed similarity to already identified duplicated regions and can represent an additional duplication. In the case of the CsgB protein, the sequence of the signal peptide was predicted as α-helix, whereas in CsgA the prediction is ambiguous. One part of it can form an α-helix and the other β-strand.

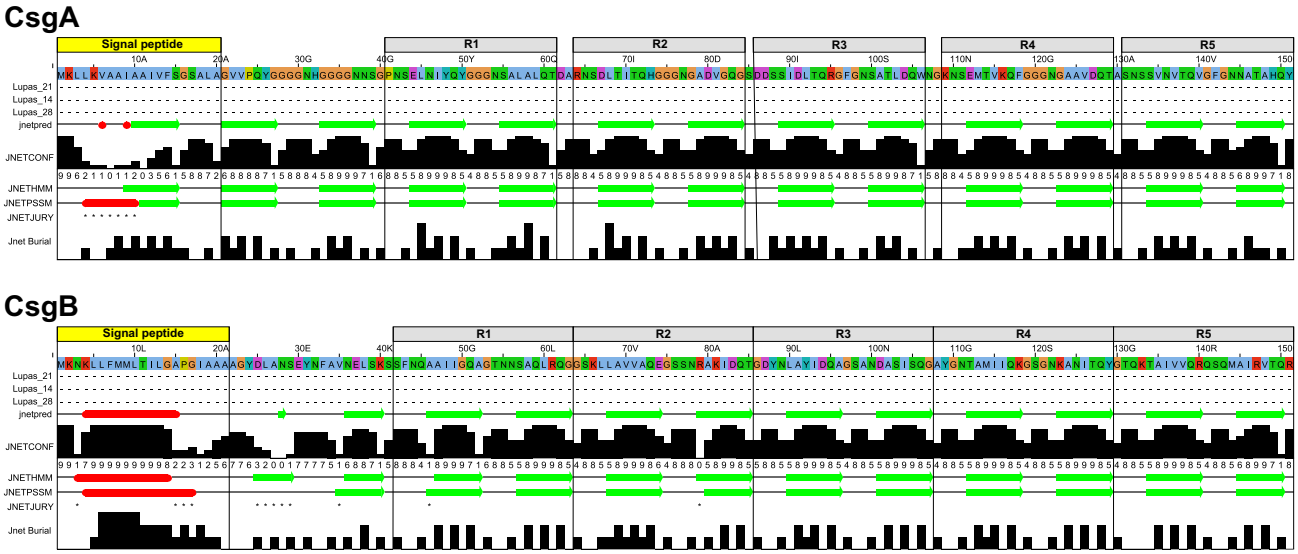

Figure 20: **Secondary prediction for CsgA and CsgB sequences.** Lupas_21, Lupas_14 and Lupas_28 are coiled-coil predictions for the sequence; Jnet Burial is the prediction of solvent accessibility; JNetPRED is the consensus prediction; JNetCONF is the confidence estimate for the prediction; JNetHMM is the HMM profile based prediction; JNETPSSM is the PSSM based prediction; JNETJURY - '*' in this annotation indicates that the JNETJURY was invoked to rationalize significantly different primary predictions. Helices are marked as red tubes and sheets as dark green arrows.

5 Phylogenetic analyses of CsgA and CsgB homologs

5.1 Research objectives

CsgA and CsgB sequences from *E. coli* have the same length, and demonstrate a similar organization, consisting of signal peptide and five repeating regions. In spatial structure, they are also very similar. Moreover, they interact with each other, and CsgB is a specific nucleator protein of the extracellular self-assembly of CsgA [Dunbar et al., 2019]. These would suggest that these proteins are phylogenetically related and should share a common ancestry. However, their sequence similarity is quite weak, so their evolution should have been more complex. It would be also possible that the organization and structure of these proteins evolved independently and convergently. Therefore, we collected distant homologs to these proteins and conducted extensive phylogenetic analyses to reconstruct their evolutionary history.

5.2 Materials and methods

5.2.1 Alignment of CsgA and CsgB sequences

Global (using the Needleman-Wunsch algorithm) and local (using the Smith-Waterman algorithm) pairwise alignments of CsgA (P28307) and CsgB (P0ABK7) sequences from *Escherichia coli* were conducted with needle and water applications from EMBOSS package [Rice et al., 2000] at EMBL-EBI web site, respectively (https://www.ebi.ac.uk/Tools/psa/) (Fig. 21). The parameters of the alignments were matrix: BLOSUM62, gap penalty: 10.0 and extend penalty: 0.5.

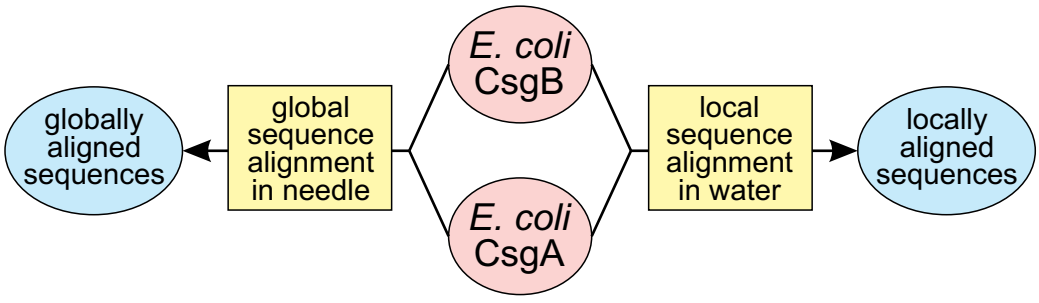

Figure 21: **Flowchart of alignment of CsgA and CsgB sequences.**

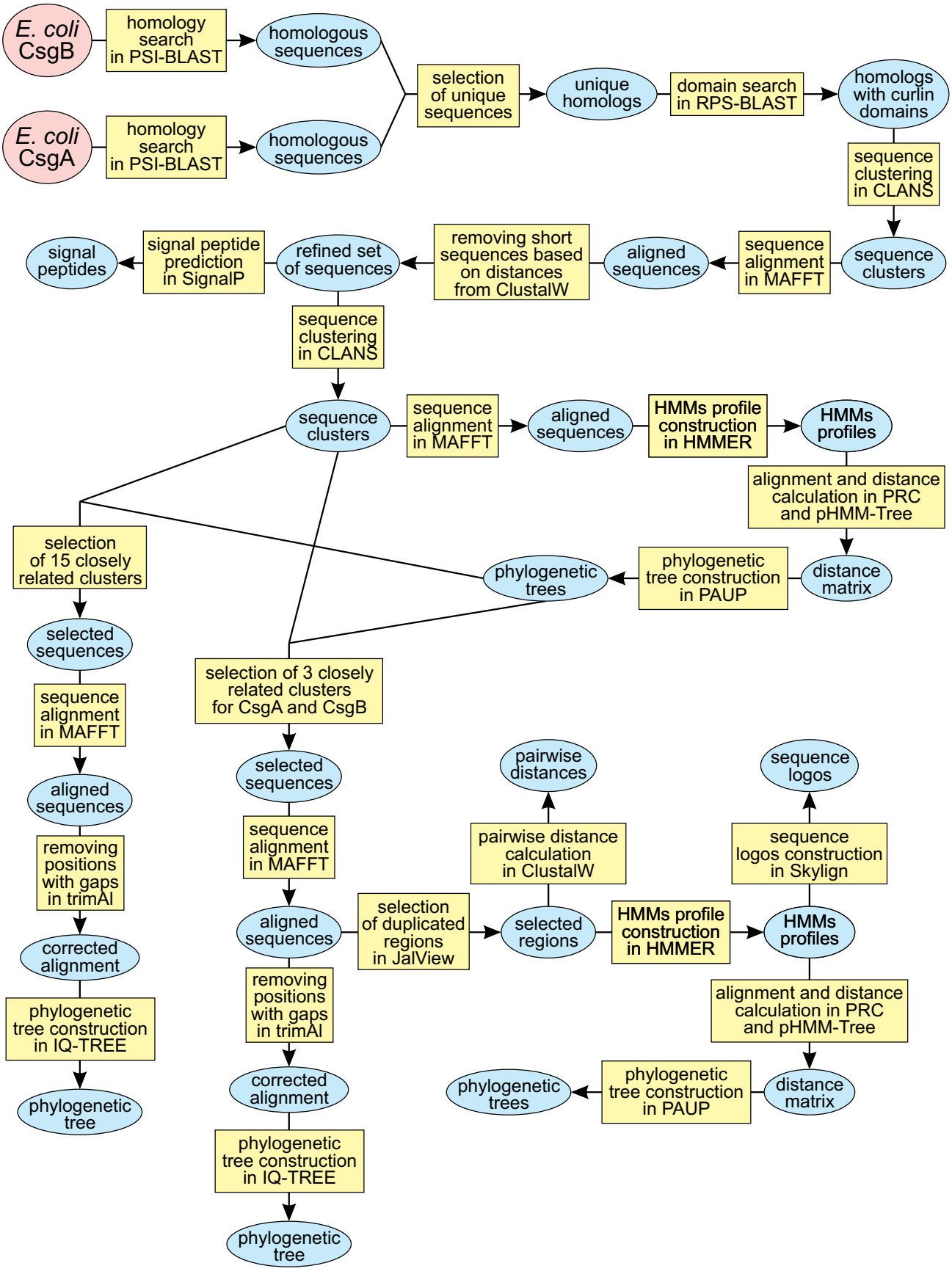

Figure 22: **Flowchart of phylogenetic and analyses of CsgA and CsgB sequences.**

5.2.2 Searching for homologs

In order to collect a comprehensive, non-redundant and reliable set of homologs to CsgA and CsgB proteins, we conducted extensive bioinformatic analyses, which consisted of many steps (Fig. 22). At first, we searched, separately for the CsgA and CsgB sequences, the NCBI non-redundant amino acid sequence database, consisting of 514,690,806 records, using PSI-BLAST 2.13.0+ [Altschul et al., 1997]. We assumed the word size = 2, E-value < 0.001 for saving hits, and E-value < 0.0001 for the inclusion of sequences in the construction of the position-specific scoring matrix (PSSM). We applied three iterations in these searches. The unique hits from these two searches were selected to one set of 15,705 sequences, from which two with the annotation "synthetic" were removed.

PSI-BLAST, i.e. Position-Specific Iterated Basic Local Alignment Search Tool, is used to detect distant relationships between proteins. After finding homologous sequences to a query in the first step, it calculates a profile or a position-specific score matrix (PSSM) from the multiple alignment of the homologs. The PSSM captures the conservation pattern in the alignment and stores it as a matrix of scores for each position in this alignment. Then, this profile is used to search again the database to find sequences that match the pattern described by the matrix. The newly selected sequences from this second round of the search are again added to the alignment and the profile is refined. This process is iteratively run until a sufficient number of homologs are collected or no new sequences can be detected above the assumed threshold. Thereby, PSI-BLAST is capable of detecting more distant than a single search done in BLASTP.

5.2.3 Identification of conserved domains

We used RPS-BLAST with Conserved Domain Database 3.20 [Marchler-Bauer et al., 2013] and each of the found sequences as a query to identify in them three domains that are characteristic of CsgA and CsgB proteins: CDD:182211, PRK10051, csgA, major curlin subunit CsgA; CDD:182242, PRK10101, csgB, curlin minor subunit CsgB; Provisional; CDD:429248, pfam07012, CurlinS_rpt, Curlin associated repeat. We applied E-value < 0.01 in these searches. Due to this approach, we identified 15,180 sequences that contained at least one of these domains.

RPS-BLAST, i.e. Reverse Position-Specific BLAST, is a variant of the BLAST that searches

a query sequence against a pre-calculated position-specific scoring matrix (PSSM) derived from a set of related protein sequences. This allows for more sensitive and specific detection of distant homologs and conserved domains specific for a protein family. Using RPS-BLAST, we can search the Conserved Domain Database (CDD), which is a collection of pre-calculated PSSMs for conserved protein domains, motifs, and functional sites.

5.2.4 Clustering and aligning of sequences

The obtained set of sequences was subjected to a clustering analysis made in CLANS (Cluster Analysis of Sequences) software [Frickey and Lupas, 2004], which visualizes BLAST pairwise sequence similarities in either two-dimensional or three-dimensional space. Analyzed sequences are represented in the graph by vertices, which are connected by edges reflecting attractive forces proportional to the negative logarithm of the P-value. In these BLASTP searches, we used the word size 2 and E-value threshold 1. To automatically detect clusters of sequences, we applied the network approach setting 2 as the minimum number of sequences per cluster. This method assumes that each sequence forms a node of the input layer for a network. These nodes emit the number of the cluster to which the sequence belongs.

The clustering resulted in 40 groups. Sequences in each cluster were aligned in MAFFT 7.505 [Katoh and Standley, 2013] using the slow and accurate algorithm E-INS-i with 1,000 cycles of iterative refinement except for the sequences of the most numerous cluster containing as many as 5200 sequences, which were aligned by a much faster algorithm FFT-NS-I assuming also 1,000 cycles. For each set of the aligned sequences, we calculated pairwise differences using ClustalW 2.1 [Thompson et al., 1994] and removed shorter sequences that were identical in the aligned positions with longer ones. Thereby, the final set was reduced to 13,652 sequences and was subjected to further studies. We searched in these sequences a potential signal peptide using SignalP 6.0 [Teufel et al., 2022] assuming a slow model mode.

The sequences were also again clustered using the network algorithm in CLANS, which produced 17 groups. The most abundant cluster including 9642 sequences was subjected to additional separation into a further 17 clusters. Thus, the total number of clusters was 33. The sequences in each cluster were once more aligned using the algorithm E-INS-i with 1,000 cycles in MAFFT.

5.2.5 Analyses of profile Hidden Markov models

For the results from multiple alignments of individual clusters, we constructed profile hidden Markov models (HMMs) using HMMER 3.3.2 package [Eddy, 1998, 2011]. The profiles are probabilistic models that capture position-specific information about conservation of each residue in each column of multiple sequence alignment. In other words, the profiles transform the alignment into a position-specific scoring system.

Distance matrices between the produced profiles were generated with software pHMM-Tree [Huo et al., 2017] after aligning them with the Profile Comparer PRC [Madera, 2008]. To reconstruct the evolutionary relationships between the profiles, phylogenetic trees were created based on the matrices in PAUP* 4.0a [Swofford, 1998]. We applied three algorithms: minimum evolution, balanced minimum evolution and Fitch-Margoliash criterion, i.e., weighted least squares with power 2. In each tree construction, starting trees were obtained via random stepwise addition with 10 replicates followed by the branch-swapping algorithm tree-bisection-reconnection (TBR) with a reconnection limit of 8. Using these three trees, a majority rule consensus was produced.

5.2.6 Phylogenetic analyses

Based on the consensus tree of HMM profiles, 6097 sequences from the 15 most closely related clusters including those grouping CsgA and CsgB proteins were selected. After the elimination of fragmentary sequences, the set including 5764 sequences was used for further investigations.

The sequences were aligned with the algorithm E-INS-i with 1,000 cycles in MAFFT and all positions in the alignment with gaps in at least 50% of sequences were removed using trimAl 1.4 [Capella-Gutiérrez et al., 2009]. This procedure provided the alignment with 145 most reliable sites.

Based on this alignment, we inferred a maximum likelihood phylogenetic tree with IQ-TREE 2.2.0 [Minh et al., 2020] assuming EXS_EHO+R10 as the best-fit substitution model as found according to BIC by ModelFinder [Kalyaanamoorthy et al., 2017] associated with this software. In the tree inferring, we used the more thorough and slower NNI (nearest-neighbor interchange) branch-swapping algorithm, which takes into account all possible NNIs instead of only similar to the previous ones. Moreover, we assumed 1000 initial parsimony trees and 100 top initial

parsimony trees to optimize with the NNI search to initialize the candidate set. To assess the significance of clades, we applied the Shimodara-Hasegawa-like approximate likelihood ratio test (SH-aLRT) with 10,000 replicates.

Moreover, to study in more detail the evolution of close homologs to CsgA and CsgB, we selected sequences from the CLANS clusters that comprised the reference sequences of these proteins from *E. coli*. Each of these clusters was grouped together with two other clusters according to the tree based on HMM profiles. Sequences from these clusters were also chosen to the sequences from the reference clusters. The set of close homologs to CsgA consisted of 1083 sequences, whereas that of close homologs to CsgB included 1517 sequences. The sequences were aligned with E-INS-i with 1,000 cycles in MAFFT and 151 reliable sites in each of these multiple alignments were selected as described previously using trimAl. Phylogenetic trees were constructed using the same methodology mentioned above assuming LG4M+R5 and Q.plant+R6 substitution models for the CsgA and CsgB sets, respectively.

5.2.7 Analyses of individual duplicated regions

From the multiple sequence alignments, we extracted five regions as identified by MEME motif searches in section 4.3.1. For these regions, we constructed HMM profiles and constructed phylogenetic trees as described previously, but instead of a heuristic search, we applied an exhaustive search. Based on these profiles, logos were generated with the Skylign tool [Wheeler et al., 2014], assuming information content: All. Pairwise differences (p-distance, i.e. fraction of different positions) between sequences for the individual regions were calculated in ClustalW. Differences between these distances were compared in the non-parametric unpaired Wilcoxon test between CsgA and CsgB regions, whereas in the comparison of regions for the given set of homologs CsgA or CsgB, we used the paired version of this test. The Spearman correlation coefficient was also calculated between the distances in all combinations of these regions. The Benjamini-Hochberg method was applied for p-value correction to control the false discovery rate. P-values smaller than 0.05 were considered significant. The statistical analysis was performed in R package [RStudio Team, 2020].

5.2.8 Other software used

Sequence alignments were inspected and studied in JalView [Waterhouse et al., 2009]. Phylogenetic trees were inspected and edited in MEGA 11 [Tamura et al., 2021]. FigTree [Rambaut and Drummond, 2012] and functions from R package ggtree [Yu et al., 2023].

5.3 Results

5.3.1 Comparison of CsgA and CsgB sequences

Both sequences of CsgA and CsgB from *Escherichia coli* have the same length of 151 amino acid residues and demonstrate the same structure consisting of a signal peptide, a separating sequence and five repeating units of similar length (Fig. 3). They also are amyloids that interact with each other [Zhou et al., 2012c]. All of that would suggest that these proteins are close homologs with common and rather recent ancestry. However, the optimal global alignment using the Needleman-Wunsch algorithm produced quite poor alignment, with only 22% identity (Fig. 23).

```
CsgA 1 ---------MKLLKVAAIAAIVFSGSALAGVVPQYGGGGNHGGGGNNSGP 41
              :.:|....|||.  :|..||
CsgB 1 MKNKLLFMMLTILGAPGIAAA--AGYDLA--------------------  27

CsgA 42 NSELNIYQYGGGNSALALQTDARNSDLTITQHGGGNGADVGQGSDDSSID 91
        |||.|.       :...|...:.|....|.|.|..|.|.:.||......
CsgB 28 NSEYNF-------AVNELSKSSFNQAAIIGQAGTNNSAQLRQGGSKLLAV 70

CsgA 92 LTQRGFGNSATLDQWNGKNSEMTVKQFGGGNGAAVDQTASNSSVNVTQVG 141
        :.|.|..|.|.:|| .|.::...:.|.|..|.|::.|.|..::..:.|.|
CsgB 71 VAQEGSSNRAKIDQ-TGDYNLAYIDQAGSANDASISQGAYGNTAMIIQKG 119

CsgA 142 FGNNATAHQY----------------------  151
         .||.|...||
CsgB 120 SGNKANITQYGTQKTAIVVQRQSQMAIRVTQR 151
```

Figure 23: **Global alignment between CsgA (P28307) and CsgB (P0ABK7) sequences.**

The optimal local alignment using the Smith-Waterman algorithm was not better. It increased the identity to only 30% (Fig. 24).

It would suggest that these proteins are not at least close homologs. These analyses imply also that the sequence organization of CsgA and CsgB could evolve in converge and their evolution was more complicated. Therefore, we decided to study this subject in detail using a more advanced approach.

```
CsgA 42 NSELNIYQYGGGNSALALQTDARNSDLTITQHGGGNGADVGQGSDDSSID 91
        |||.|.       :...|...:.|....|.|.|..|.|.:.||.......
CsgB 28 NSEYNF-------AVNELSKSSFNQAAIIGQAGTNNSAQLRQGGSKLLAV 70

CsgA 92 LTQRGFGNSATLDQWNGKNSEMTVKQFGGGNGAAVDQTASNSSVNVTQVG 141
        :.|.|..|.|.:|| .|..:...:.|.|..|.|::.|.|..::..:.|.|
CsgB 71 VAQEGSSNRAKIDQ-TGDYNLAYIDQAGSANDASISQGAYGNTAMIIQKG 119

CsgA 142 FGNNATAHQY    151
          .||.|...||
CsgB 120 SGNKANITQY    129
```

Figure 24: **Local alignment between CsgA (P28307) and CsgB (P0ABK7) sequences.**

5.3.2 Collection of CsgA and CsgB homologs

Since CsgA and CsgB sequences from *E. coli* show a very poor sequence similarity, we applied a sensitive search using PSI-BLAST dedicated to distant homologs. Thanks to this approach, we collected the set of 15,703 potential homologous sequences in separated searches for these curli proteins. It can be added that these sequenced showed no significant similarity at the assumed threshold E-value < 0.001 after the first iteration. The CsgA found the CsgB after the second searching iteration with E-value 9.4E-10, whereas CsgB identified CsgA only after the third iteration with E-value 2.4E-10. It means that they are distant homologs, but the significant similarity can be confirmed after more sensitive searches. From that, we selected 15,180 sequences that contained at least one of three conserved curlin domains identified in the reference CsgA and CsgB proteins (Tab. 9). In the vast majority of cases, more than 95%, the curlin domain was found as the best hit. Among sequences with the curlin domains, 488 contained also other domains. Some of them can represent spurious hits. The most common was CDD:227596, i.e., AAA ATPase containing von Willebrand factor type A (vWA) domain, found in 118 cases. Some of them can represent spurious hits.

Table 9: **Result of searches for curlin domains (CDD:182211, CDD:182242 and CDD:429248) in the set of collected CsgA and CsgB homologs.**

Type of hits	**Number**	**Percent**
Curlin domain found as the best hit	15,029	95.7
Curlin domain found but not as the best hit	151	1.0
Only other domain was found	165	1.1
No domain was found	358	2.3

5.3.3 Taxonomic distribution of CsgA and CsgB homologs

The taxonomic distribution of CsgA and CsgB homologs (Tab. 10) indicates that they are in majority representatives of Bacteria and in this domain of life these proteins mainly evolved. More than 98% of sequences are annotated as bacterial. Their presence in other domains of life can be associated with a horizontal gene transfer, e.g., to viruses and other bacterial groups. However, the contamination of samples cannot be excluded, and these cases should be individually verified. It concerns especially those obtained from metagenomics studies and draft genomic sequencing. Other sequences can represent false positives, especially those in higher eukaryotes, e.g., K^{+-} dependent Na^{-}/Ca^{+} exchanger and AAA ATPase with vWA domain in a higher plant, collagen α-1 and ABC transporter F in crustaceans, a zinc finger protein, dynein and intraflagellar transport protein in fishes as well as histone-lysine N-methyltransferase in birds. Some regions of these sequences due to specific features can resemble curli protein sequences, which is an interesting example of molecular convergence.

Table 10: **Taxonomic distribution of CsgA and CsgB homologs in domains of life and their main groups.**

Domain	Group	Number	Percent
Archaea	Euryarchaeota	11	84.62
Archaea	Candidatus_Pacearchaeota	2	15.38
Eukaryota	Metazoa	123	72.78
Eukaryota	Viridiplantae	37	21.89
Eukaryota	Fungi	7	4.14
Eukaryota	SAR	2	1.18
Bacteria	Proteobacteria	12795	85.31
Bacteria	Bacteroidota	1825	12.17
Bacteria	Firmicutes	116	0.77
Bacteria	Balneolaeota	68	0.45
Bacteria	Actinobacteria	28	0.19
Bacteria	Nitrospinae_Tectomicrobia_group	22	0.15
Bacteria	others	18	0.12
Bacteria	Nitrospirae	18	0.12
Bacteria	Calditrichaeota	16	0.11
Bacteria	Chlorobi	16	0.09
Bacteria	Ignavibacteriae	13	0.08
Bacteria	Rhodothermaeota	12	0.06
Bacteria	Thermodesulfobacteria	9	0.04
Bacteria	Cyanobacteria	6	0.03
Bacteria	Fibrobacteres	5	0.03
Bacteria	Candidatus_Dadabacteria	4	0.02
Bacteria	Aquificae	3	0.02
Bacteria	candidate_division_KSB1	3	0.02
Bacteria	Patescibacteria_group	3	0.01
Bacteria	Acidobacteria	2	0.01
Bacteria	Candidatus_Auribacterota	2	0.01
Bacteria	Candidatus_Poribacteria	2	0.01
Bacteria	Chloroflexi	2	0.01
Bacteria	environmental_samples	2	0.01
Bacteria	Kiritimatiellaeota	2	0.01
Bacteria	Parcubacteria_group	1	0.01
Bacteria	Planctomycetota	1	0.01
Viruses	Duplodnaviria	3	0.02
Viruses	environmental_samples	1	0.01

Table 11: **Taxonomic distribution of CsgA and CsgB homologs in selected bacterial groups.**

Group	Subgroup	Number	Percent
Bacteroidota	Flavobacteriia	1010	55.34
Bacteroidota	Cytophagia	471	25.81
Bacteroidota	Bacteroidia	160	8.77
Bacteroidota	Bacteroidetes	85	4.66
Bacteroidota	other	50	2.74
Bacteroidota	Saprospiria	40	2.19
Bacteroidota	Sphingobacteriia	7	0.38
Bacteroidota	Chitinophagia	1	0.05
Bacteroidota	environmental	1	0.05
Proteobacteria	γ-Proteobacteria	8986	70.23
Proteobacteria	α-Proteobacteria	3076	24.04
Proteobacteria	β-Proteobacteria	648	5.06
Proteobacteria	δ-Proteobacteria	43	0.34
Proteobacteria	other	23	0.18
Proteobacteria	ζ-Proteobacteria	9	0.07
Proteobacteria	Oligoflexia	7	0.05
Proteobacteria	Hydrogenophilalia	2	0.02
Proteobacteria	ϵ-Proteobacteria	1	0.01

The most abundant in curli homologs in Bacteria are *Bacteroidota* and *Proteobacteria* (Tab. 10 and 11. They constitute, 85% and 12% , respectively. Considering their subgroups, the largest number of homologs was detected in *Flavobacteriia* (55% of *Bacteroidota*) α-*Proteobacteria* (24% of *Proteobacteria*) and γ-*Proteobacteria* (70% of *Proteobacteria*) (Tab. 12), which indicates that these proteins evolved mainly in these groups. Among α-*Proteobacteria*, *Hyphomicrobiales* (56%) has most of the homologs, whereas in γ-*Proteobacteria*, *Enterobacterales* (43%) and *Pseudomonadales* (29%) are most abundant (Tab. 12). The sequences are not evenly distributed across subgroups. It can be related with the bias in the number of sequenced genomes associated with a preference of researchers and the ease of culturing and isolation from the environment.

Table 12: **Taxonomic distribution of CsgA and CsgB homologs in domains of life and their main groups.**

Group	Subgroup	Number	Percent
α-Proteobacteria	Hyphomicrobiales	1717	55.82
α-Proteobacteria	Sphingomonadales	592	19.25
α-Proteobacteria	Rhodobacterales	356	11.57
α-Proteobacteria	Hyphomonadales	117	3.80
α-Proteobacteria	other	108	3.51
α-Proteobacteria	Rhodospirillales	82	2.67
α-Proteobacteria	Maricaulales	45	1.46
α-Proteobacteria	Caulobacterales	42	1.37
α-Proteobacteria	Parvularculales	7	0.23
α-Proteobacteria	Rickettsiales	6	0.20
α-Proteobacteria	Emcibacterales	4	0.13
γ-Proteobacteria	Enterobacterales	3833	42.66
γ-Proteobacteria	Pseudomonadales	2626	29.22
γ-Proteobacteria	Alteromonadales	1114	12.40
γ-Proteobacteria	Oceanospirillales	406	4.52
γ-Proteobacteria	Vibrionales	304	3.38
γ-Proteobacteria	Cellvibrionales	261	2.90
γ-Proteobacteria	Aeromonadales	216	2.40
γ-Proteobacteria	other	93	1.03
γ-Proteobacteria	Chromatiales	71	0.79
γ-Proteobacteria	Methylococcales	20	0.22
γ-Proteobacteria	Moraxellales	10	0.11
γ-Proteobacteria	Nevskiales	10	0.11
γ-Proteobacteria	Gallaecimonas	6	0.07
γ-Proteobacteria	Xanthomonadales	6	0.07
γ-Proteobacteria	sulfur-oxidizing	4	0.04
γ-Proteobacteria	Thiotrichales	3	0.03
γ-Proteobacteria	Candidatus	2	0.02
γ-Proteobacteria	Pasteurellales	1	0.01

5.3.4 Initial clustering of CsgA and CsgB homologs

In order to comprehend such a huge set of sequences, we carried out their clustering in CLANS based on the results of pairwise BLASTP searches. The algorithm distributed the studied sequences in two-dimensional space according to their BLAST pairwise sequence similarities (Fig. 25). Each sequence is represented by a point in the plot, and lines correspond to attractive forces proportional to the significance of the similarity.

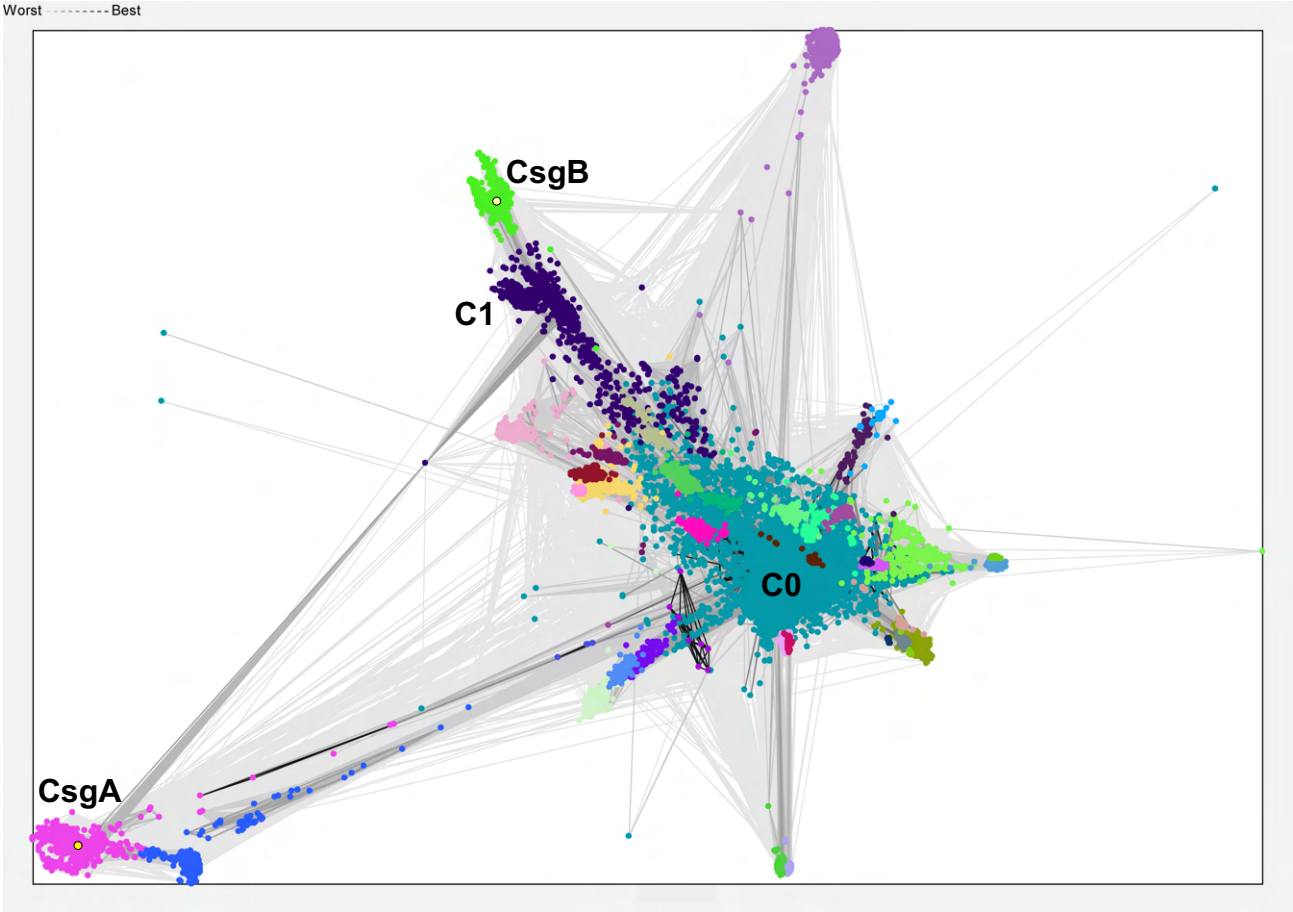

Figure 25: **Analysis of 15,180 CsgA and CsgB homologs in CLANS showing identified clusters.** The analyzed sequences are represented by vertices connected by edges reflecting attractive forces proportional to the negative logarithm of P-value. The grayness intensity of the connections is proportional to these forces. Recognized clusters were marked by different colors. The most numerous are indicated as C0 and C1. Yellow circles represent the reference CsgA and CsgB sequences.

The analysis distinguished 40 clusters including from 4 to 5200 sequences. The clusters are marked by various colors in the plot. The most numerous cluster C0 is located in the center

of the plot and surrounded by smaller ones. Six clusters are clearly separated from the main grouping in the center. Among them are those that contain the reference sequences of CsgA and CsgB, which are separated into two different clusters located at a great distance from each other. The CsgA cluster is the most distantly placed from the middle of the plot. Although members of these two clusters are not directly connected by the lines, other clusters relate them. The CsgA cluster is strongly connected with cluster C1, which is very close to the CsgB cluster. The results indicate that CsgA and CsgB are distant homologs not directly related and CsgB is more similar to other sequences than CsgA.

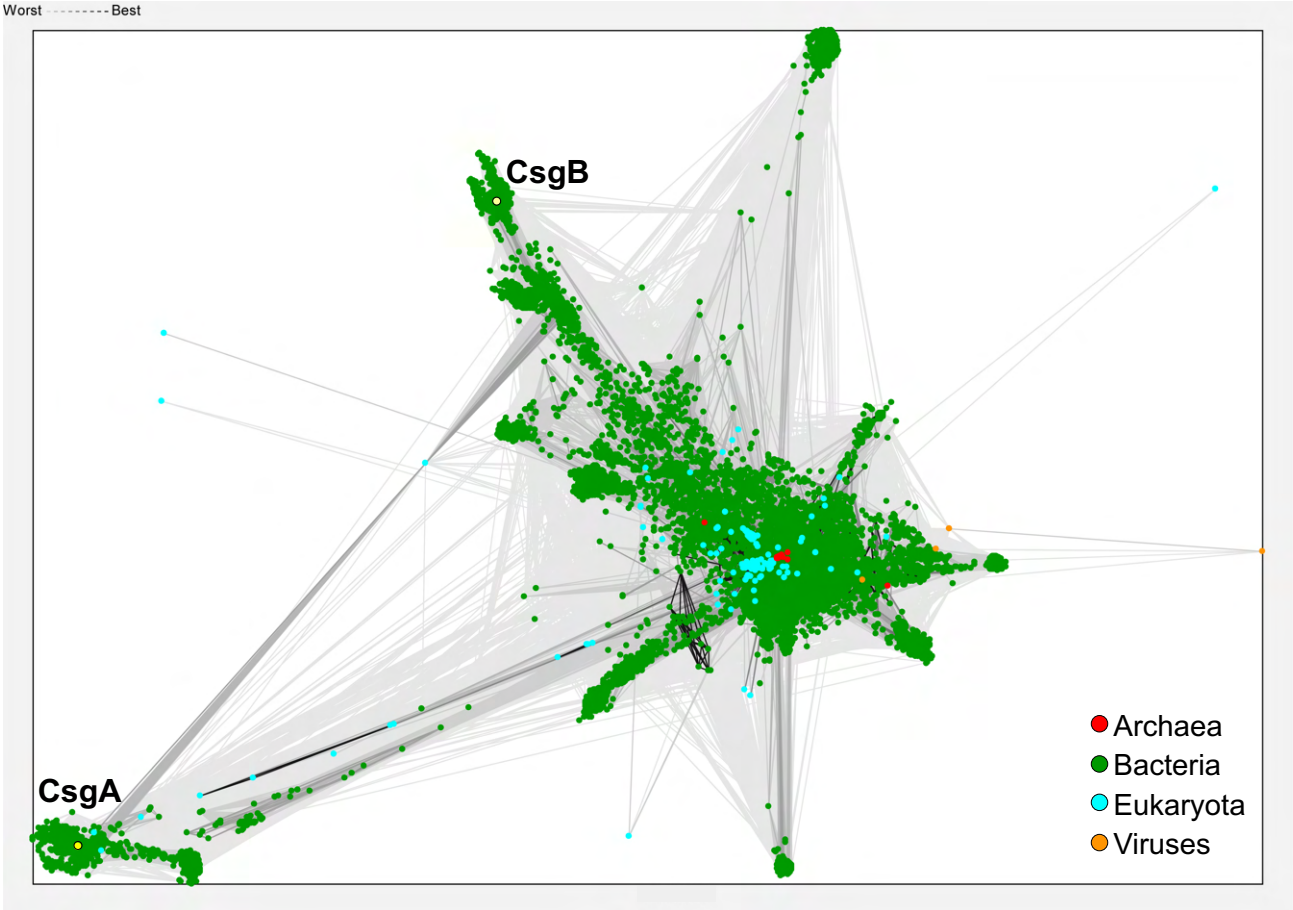

Figure 26: **Analysis of 15,180 CsgA and CsgB homologs in CLANS, showing sequences from domains of life.** The analyzed sequences are represented by vertices connected by edges reflecting attractive forces proportional to the negative logarithm of the P-value. The grayness intensity of the connections is proportional to these forces. Yellow circles represent the reference CsgA and CsgB sequences.

Most homologs (14,993) that were found belong to the Bacteria domain, which are widely distributed in the plot (Fig. 26). Only 4 were found in viruses, 13 in Archaea and 169 in Eukaryota. The archaeal and eukaryotic (169) sequences are mainly grouped in the center, whereas viral sequences are placed at various locations. Some of them are distantly located from others and represented by separate points. Some eukaryotic sequences are placed close to CsgA homologs.

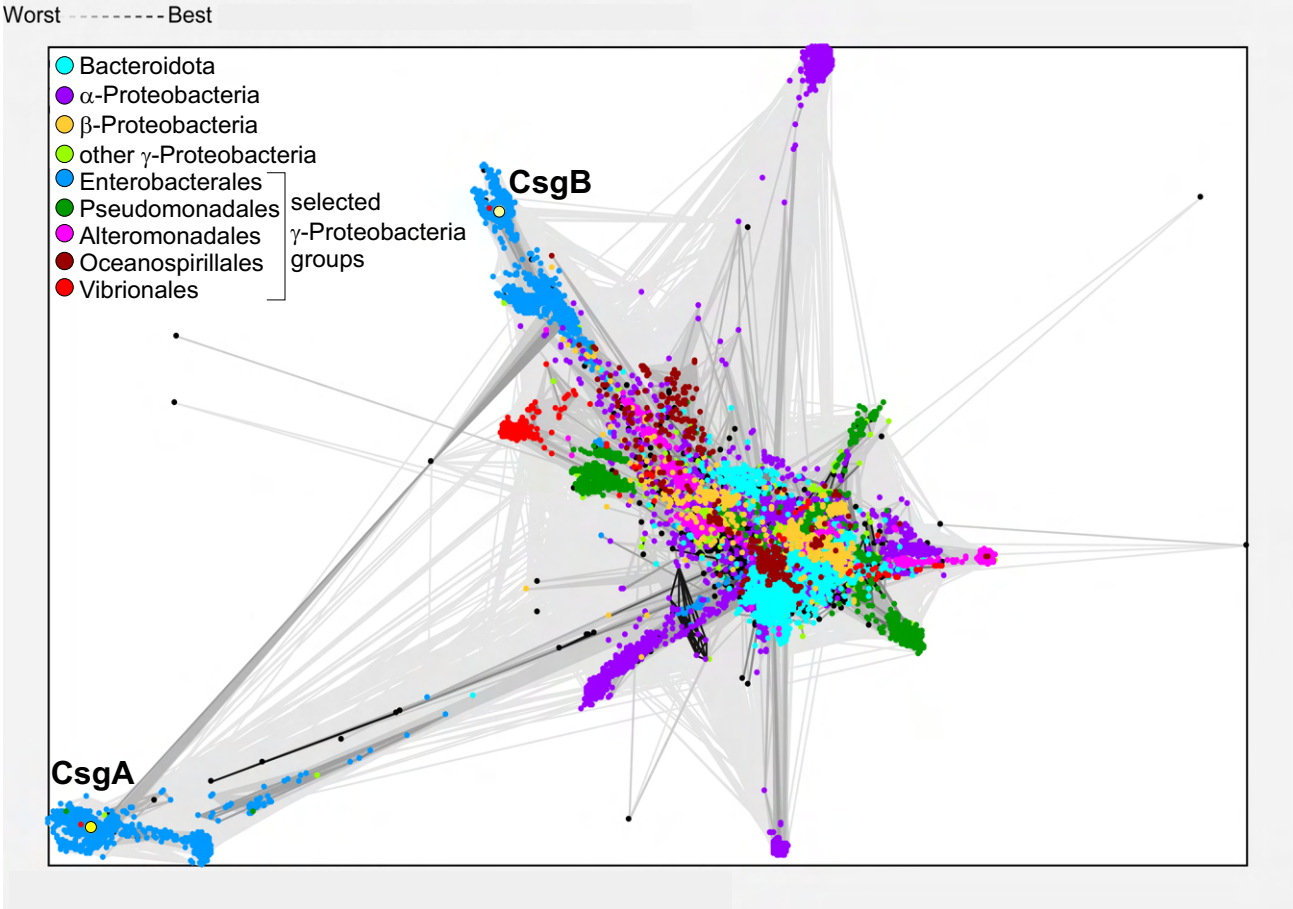

Figure 27: **Analysis of 15,180 CsgA and CsgB homologs in CLANS, showing sequences from selected bacterial groups.** The analyzed sequences are represented by vertices connected by edges reflecting attractive forces proportional to the negative logarithm of P-value. The grayness intensity of the connections is proportional to these forces. Yellow circles represent the reference CsgA and CsgB sequences.

Since the most abundant are bacterial sequences, we marked the most numerous bacterial groups in the CLANS plot (Fig. 27). *Bacteroidota* sequences are separated into two main groups

in the center of the plot, *β-Proteobacteria* are placed in the middle too. Some *α-Proteobacteria* sequences are also located in the center, but others create one clear group at the border of the central groupings and two distantly located. Selected *γ-Proteobacteria* subgroups are distributed into separated sets. *Enterobacterales* are placed in two groups, including the reference CsgA and CsgB sequences, clearly isolated from others. *Pseudomonadales* are present in at least three groups at the boundary of the main set. *Alteromonadales* and *Oceanospirillales* are inside the plot, but at least two groups for each of them can be recognized. *Vibrionales* are clearly separated from the main grouping. The distribution of these sequences suggests that CsgA and CsgB homologs are the most abundant among *Proteobacteria*, but some homologs can be also found in *Bacteroidota*. The presence of many separated clusters with a main grouping indicates a rapid differentiation of curli proteins into various subgroups and further expansion in one taxonomic clade. However, sequences affiliated with one taxonomic clade are often separated, which means that some duplications could occur before this clade evolved and/or the sequences were subjected to rapid differentiation. This applies also to *Enterobacterales* sequences, which are clearly separated.

5.3.5 Signal peptide prediction

The inspection of obtained multiple sequence alignments showed that many sequences are truncated and fragmentary. Therefore, we removed them, leaving their identical but longer homologs. It resulted in a set of 13,652 sequences, which were further studied.

Since the CsgA and CsgB proteins are equipped with an N-terminal signal peptide responsible for their extracellular transport, we searched for this feature (Tab. 13).

Table 13: **Results of signal peptide prediction in 13,652 CsgA and CsgB homologs.**

Feature	**Number**	**Percent**
SP (Sec/SPI)	12760	93.47
OTHER (no SP)	778	5.70
LIPO (Sec/SPII)	113	0.83
TAT (Tat/SPI)	1	0.01

The analyses showed the presence of the Sec signal peptide (Sec/SPI), which is a “standard”

secretory signal peptide, transported by the Sec translocon and cleaved by Signal Peptidase I (Lep). It was identified in more than 93% of sequences. For almost 95% of them, the probability was greater than 0.9. Less than 1% of sequences revealed a lipoprotein signal peptide (Sec/SPII), which is transported by the Sec translocon and cleaved by Signal Peptidase II (Lsp). They can represent false positives because we found no specific taxonomic distribution of these sequences. Only one sequence showed, with a low probability of 0.43, Tat signal peptide (Tat/SPI), which is transported by the Tat translocon and cleaved by Signal Peptidase I (Lep). In almost 6% of sequences, no signal peptide was predicted. Among them, there are also many eukaryotic sequences. In the case of bacterial sequences showing the unquestionable presence of the curlin domains, the negative results can be related to the incompleteness of their N-terminal ends.

The cleavage site of the signal peptide was predicted with a median probability of 0.978. More than 91% of cases showed a probability greater than 0.95 The length of the predicted signal peptide varied from 3 to 67 residues, but almost 63% of cases were in a narrow range from 20 to 22 and 83% from 20 to 26 (Fig. 28).

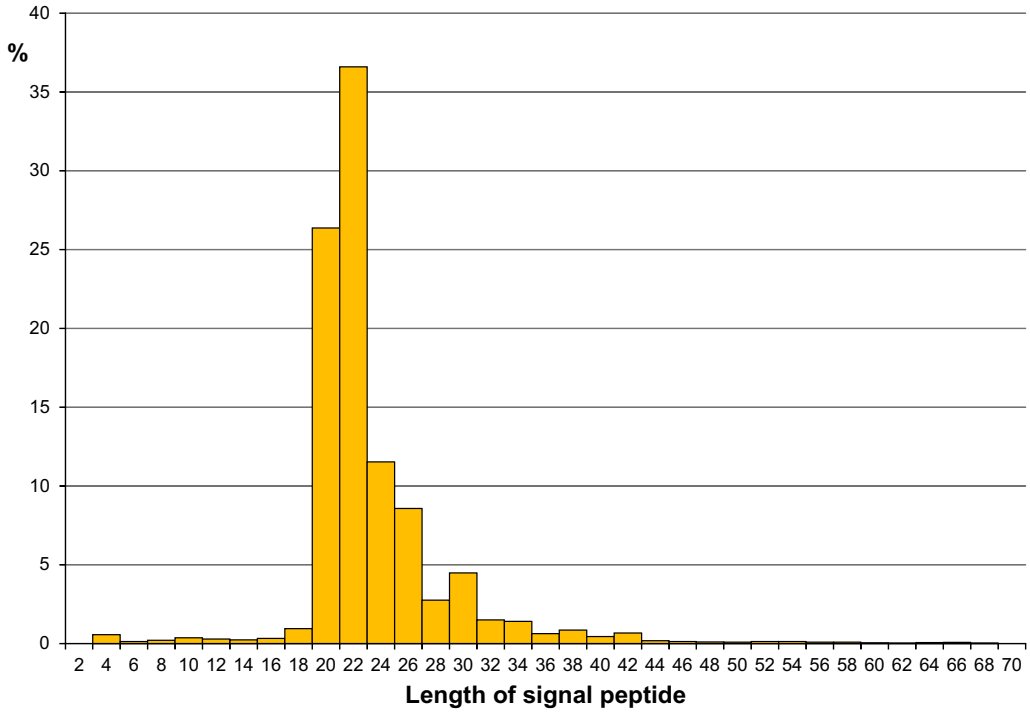

Figure 28: **The distribution of signal peptide length predicted in 12760 CsgA and CsgB homologs.**

The shortest signal peptides were predicted with very low probability. The length of most

peptides corresponds well to those annotated in UniProt [UniProt Consortium, 2018] for *E. coli* CsgA and CsgB, which are 20 and 21 residues long. The peptide for CsgA was verified experimentally [Arnqvist et al., 1992]. The results indicate that the most collected homologs demonstrate a typical CsgA and CsgB structure, including the N-terminal signal peptide.

5.3.6 Clustering of the refined set of CsgA and CsgB homologs

After removing the shorter sequence, we clustered again the sequences in CLANS. The algorithm identified 17 main groups (Tab. 14, Fig. 29). Similarly to the previous clustering, the center of the plot is occupied by one huge cluster, which is surrounded by smaller distinct clusters. Clusters including CsgA and CsgB reference sequences, C6 and C1 respectively, are similarly located as previously, but the algorithm recognized additional clusters in their neighborhood. They are C3 and C5 at C6 and C2 with some sequences from C0 at C1.

Table 14: **Clusters and their count identified for 13652 CsgA and CsgB homologs in CLANS.**

Cluster	Number	Percent
C0	9642	70.63
C1 (CsgB)	572	4.19
C2	570	4.18
C3	442	3.24
C4	412	3.02
C5	373	2.73
C6 (CsgA)	366	2.68
C7	346	2.53
C8	279	2.04
C9	193	1.41
C10	185	1.36
C11	111	0.81
C12	88	0.64
C13	50	0.37
C14	16	0.12
C15	4	0.03
C16	3	0.02

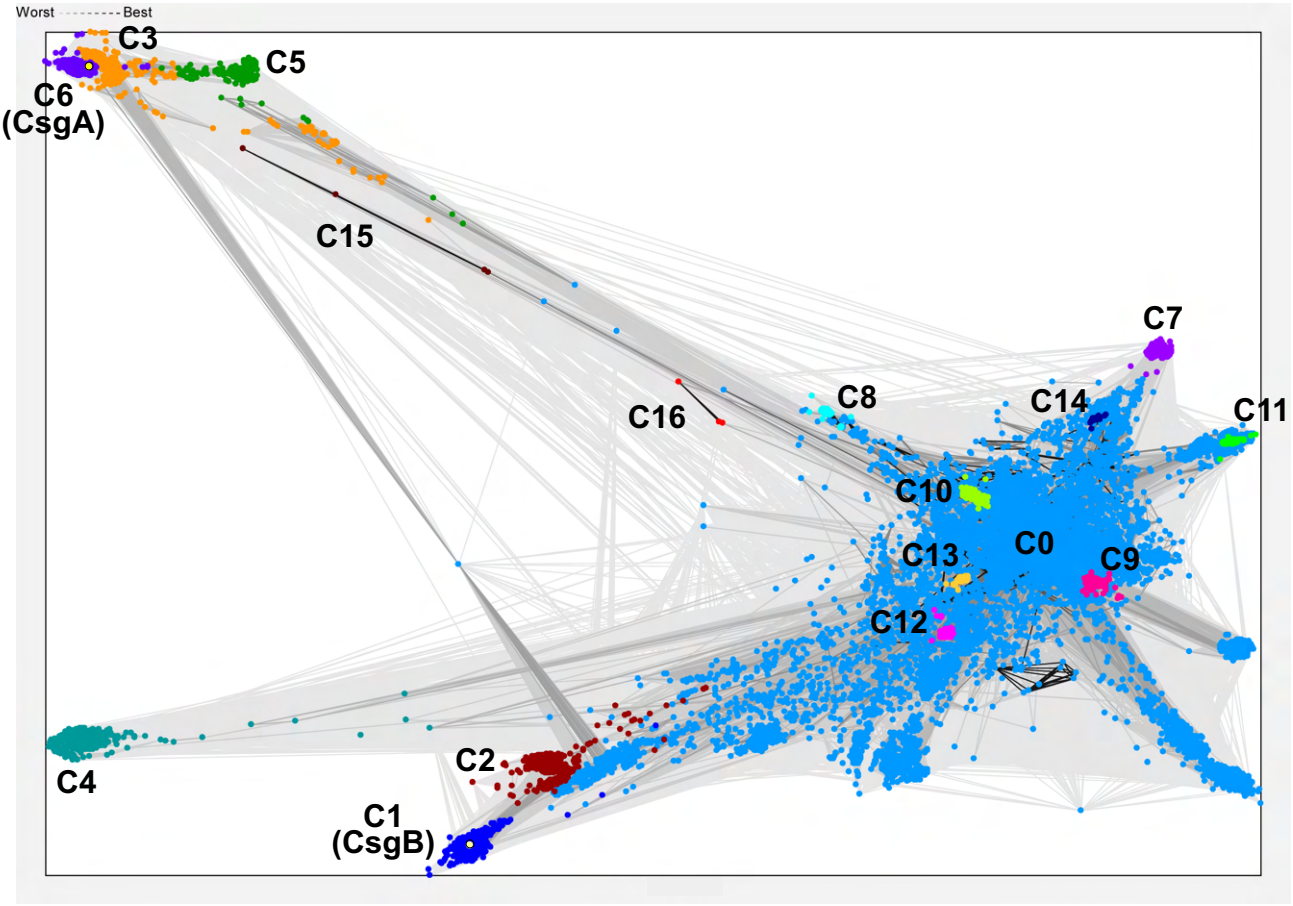

Figure 29: **Analysis of CsgA and CsgB homologs in CLANS, showing 17 main groups.** The analyzed sequences are represented by vertices connected by edges reflecting attractive forces proportional to the negative logarithm of the P-value. The grayness intensity of the connections is proportional to these forces. Yellow circles represent the reference CsgA and CsgB sequences.

Applying the next step of clustering, we separated the most numerous cluster C0 into smaller ones, whose number was also 17 (Tab.15, Fig. 30). They formed distinct groups in the CLANS plot, and the number of their members was comparable with those found in the previous step. The 33 clusters determined in these analyses were subjected to further studies.

Table 15: **Clusters and their count identified for the most numerous cluster C0 in CLANS.**

Cluster	Number	Percent
C0_0	2896	30.04
C0_1	1491	15.46
C0_2	950	9.85
C0_3	720	7.47
C0_4	686	7.11
C0_5	638	6.62
C0_6	472	4.90
C0_7	458	4.75
C0_8	457	4.74
C0_9	260	2.70
C0_10	152	1.58
C0_11	130	1.35
C0_12	130	1.35
C0_13	97	1.01
C0_14	60	0.62
C0_15	37	0.38
C0_16	8	0.08

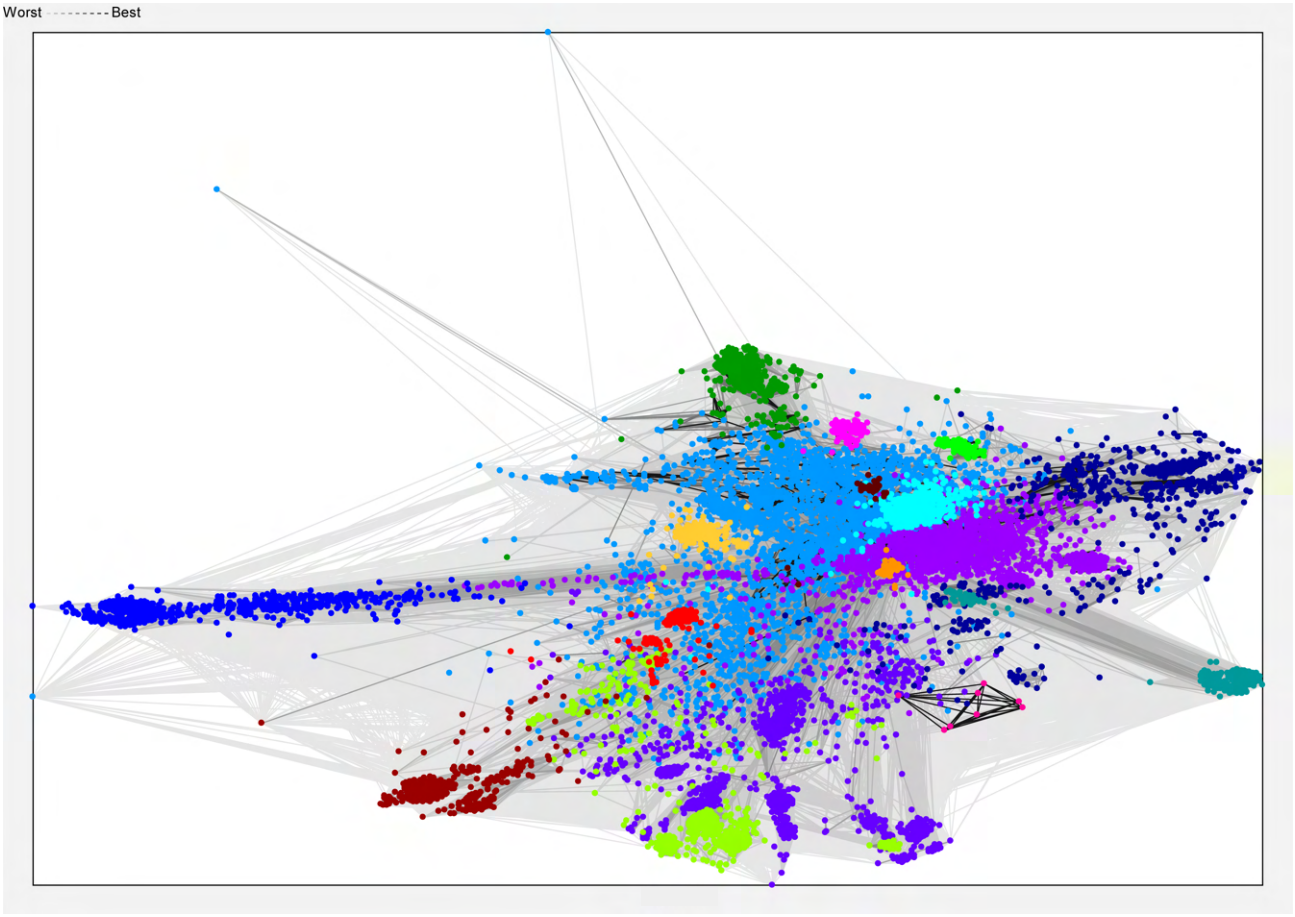

Figure 30: **Re-analysis of CsgA and CsgB C0 cluster in CLANS, showing 17 main groups.** The analyzed sequences are represented by vertices connected by edges reflecting attractive forces proportional to the negative logarithm of the P-value. The grayness intensity of the connections is proportional to these forces. Different colors represent different groups.

5.3.7 Phylogenetic relationships between clusters and sequences of curli homologs

To infer evolutionary relationships between the identified 33 clusters, we calculated a consensus phylogenetic tree based on HMM profiles produced from alignments of sequences from each cluster (Fig. 31).

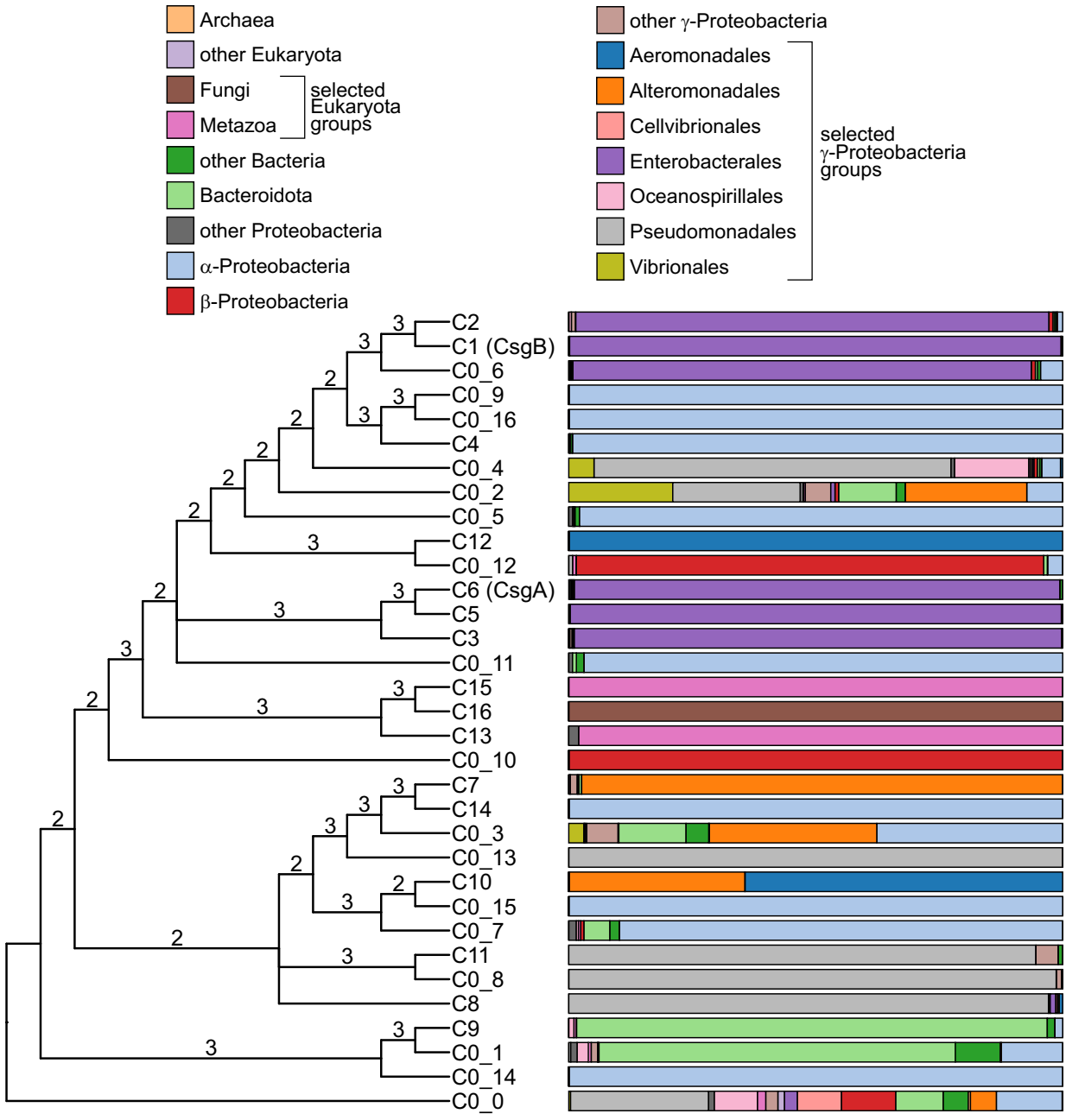

Figure 31: **The consensus tree is based on HMM profiles grouping clusters of CsgA and CsgB homologs.** On the right, a taxonomic distribution of a given cluster was presented. Numbers at branches indicate the number of trees, out of three, that produced a given branching pattern.

Most groupings in the tree were congruently inferred by three methods. Some inconsistencies were for deep branches because one method produces a different topology. Clusters for CsgA and CsgB are clearly separated and significantly grouped with the clusters that were adjacent to them also in the CLANS result. These clusters are dominated by representatives of *Enter-*

obacterales, which are not present in such a high abundance in other clusters. The CsgB cluster with its relatives is grouped with three others, including almost exclusively *α-Proteobacteria* representatives. Nevertheless, it should be noticed that the CsgA and CsgB clusters are in some way related when considering the whole tree. They are collectively present in a subtree, including almost half of all clusters, and separated from the rest. Besides two separated groups of *Enterobacterales*, there are also clusters of other taxonomic groups that do not always form monophyletic clades and are separated in the tree. It concerns, e.g., *α-Proteobacteria*, *β-Proteobacteria*, and *Pseudomonadales*. There are also heterogeneous clusters including sequences from various taxonomic groups. Three clusters comprise eukaryotic sequences from fungi and animals (Metazoa) in a significant monophyletic clade. The separation of sequences from the main groups of *Proteobacteria* indicates that the CsgA and CsgB homologs duplicated before the divergence of these groups.

Based on the results, we selected to further analyses sequences from 15 clusters (C1-C6, C12, C0_2, C0_4-C0_6, C0_9, C0_11, C0_12, C0_16) that were grouped together and included the CsgA and CsgB clusters in the HMM profile phylogeny. Using their alignment, we inferred a maximum likelihood phylogenetic tree (Fig. 32). It revealed the presence of the clade including many *α-Proteobacteria* sequences and also some *β-Proteobacteria*, which are clearly separated from other sequences coming mainly from *γ-Proteobacteria*. This split is significant, with 95%. In the second clade, we can identify four main lineages. The first group with 100% support contains almost exclusively representatives of *Enterobacterales* including the reference CsgA sequence. The second clade contains mostly *Bacteroidota* with 93% support. The third, supported in 96%, comprises predominantly other *γ-Proteobacteria* sequences, whereas the fourth clade clusters with 80% support *Enterobacterales* including the CsgB reference and other *γ-Proteobacteria* representatives.

This topology demonstrates a significant partition on two *Enterobacterales* clades including the reference curli proteins and suggests that the separation of CsgA and CsgB occurred after the divergence of *γ-Proteobacteria* from *α-* and *β-Proteobacteria* but before the differentiation of *γ-Proteobacteria* lineages. In the main groups marked in the tree (Fig. 32), there are also placed individual sequences or their small bunches assigned to other taxonomic groups, e.g., among *γ-Proteobacteria* sequences there are also *Bacteroidota* as well as *α-* and *β-Proteobacteria*

representatives, which can suggest a horizontal gene transfer. Similarly, the closer relation of the second clade comprising *Bacteroidota* with γ-*Proteobacteria* than the latter with α-*Proteobacteria* can suggest that the *Bacteroidota* acquired a CsgA homolog via horizontal gene transfer just from γ-*Proteobacteria*.

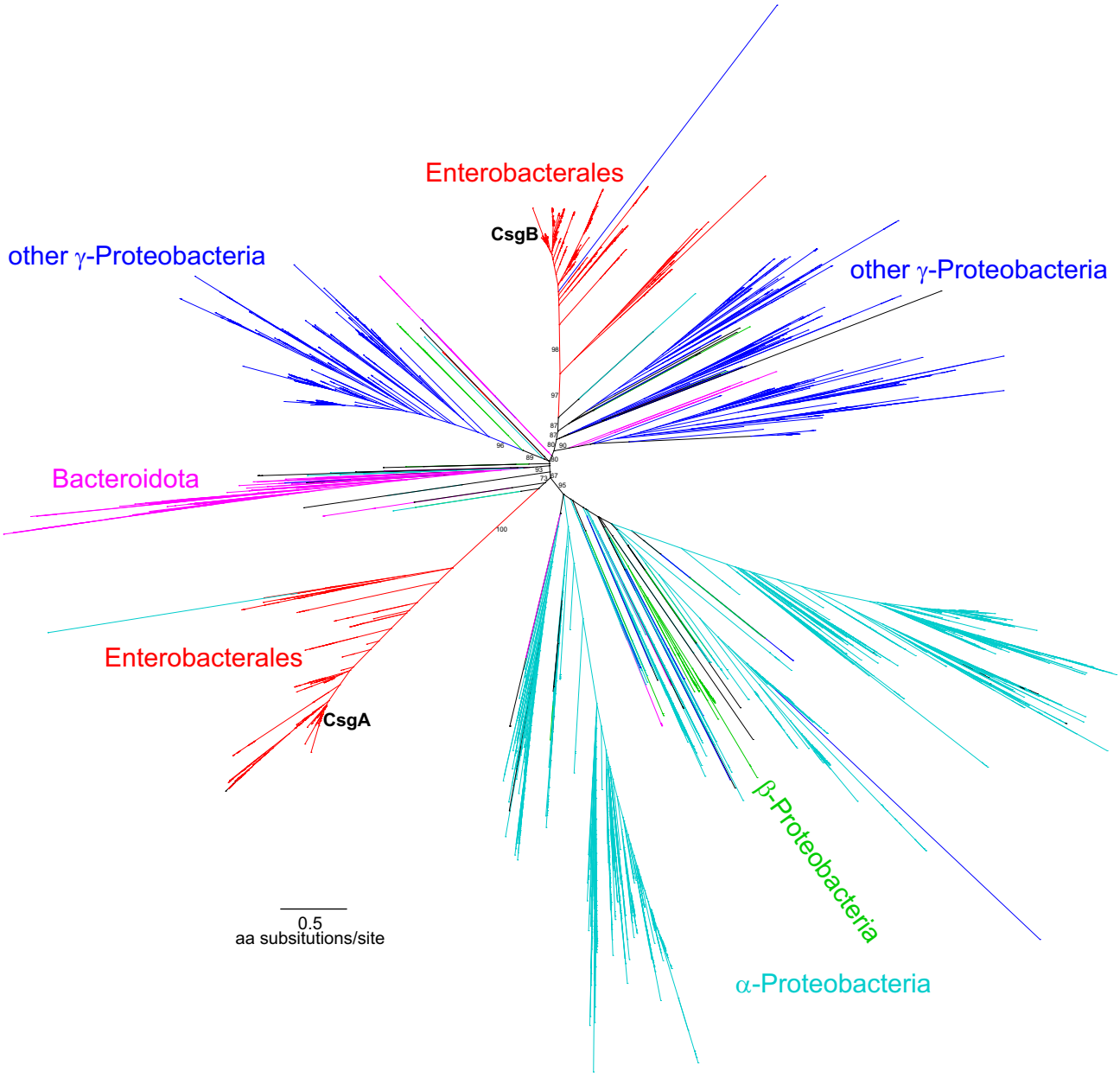

Figure 32: **The maximum likelihood tree based on the alignment of 5764 sequences shows a close similarity to the reference CsgA and CsgB sequences.** Numbers at nodes correspond to support values calculated by SH-aLRT procedure. Only selected support values at deep branches were shown. Branches of the most numerous bacterial groups were colored and labeled.

5.3.8 Phylogenetic relationships between *Enterobacterales* curli homologs

In order to analyze in detail the phylogeny of CsgA and CsgB proteins in *Enterobacterales*, we inferred separate trees for appropriate sequences. They were derived from clusters C3, C5 and C6 in the case of CsgA as well as C1, C2 and C06 for CsgB.

In the phylogenetic tree of CsgA homologs (Fig. 33), we can notice a big clade including almost all sequences from *Enterobacteriaceae*, which is significantly separated from the other group. It comprises sequences assigned to *Yersiniaceae*, which do not form a monophyletic group but separate sequentially on the tree. The next diverging lineages belong to *Kluyvera* and *Shimwellia* classified to *Enterobacteriaceae* and are sisters to a smaller monophyletic clade of *Budviciaceae*. Interestingly, within *Enterobacteriaceae* is placed a sequence assigned to a fungus *Astraeus odoratus*, and among *Yersiniaceae*, sequences annotated to *Pseudomonas reactans* from *Pseudomonadales*. It suggests a horizontal gene transfer from *Enterobacter* to *Astraeus odoratus* and from *Ewingella* to *Pseudomonas reactans*. However, these sequences come from the draft and whole genome sequencing, so should be verified in terms of contamination.

The tree of CsgB homologs also contains a significant clade grouping many *Enterobacteriaceae* taxa (Fig. 34). The second main clade also comprises members of *Budviciaceae* and *Yersiniaceae*, but the former is here monophyletic as the latter. Interestingly, an additional clade appears, which includes representatives of *Hafniaceae* and is a sister to *Yersiniaceae*. Similarly, to the CsgA tree, there are *Enterobacteriaceae* sequences from *Kluyvera* and also *Klebsiella*, which are separated from the main clade but are significantly grouped with the second one. We can also notice *Pseudomonas reactans* sequence coming from the same sequencing project as its CsgA sequence. It is also clustered with *Ewingella* suggesting a horizontal gene transfer. Another case of transfer can demonstrate a sequence assigned to *Bacteroidales bacterium*, which was obtained from metagenomic sequencing of the coral skeleton [Cárdenas et al., 2022].

In these two trees, some *Enterobacteriaceae* genera (*Kluyvera, Shimwellia* and *Klebsiella*) are clearly separated from the main clade, which can suggest a horizontal gene transfer to them from other *Enterobacterales* families. Alternatively, these genera should not be classified to the current family but assigned to another. In both trees, there are sequences assigned to individual genera and species, which are clustered into monophyletic clades, but there are exceptions. They can be incorrectly assigned to the current taxon.

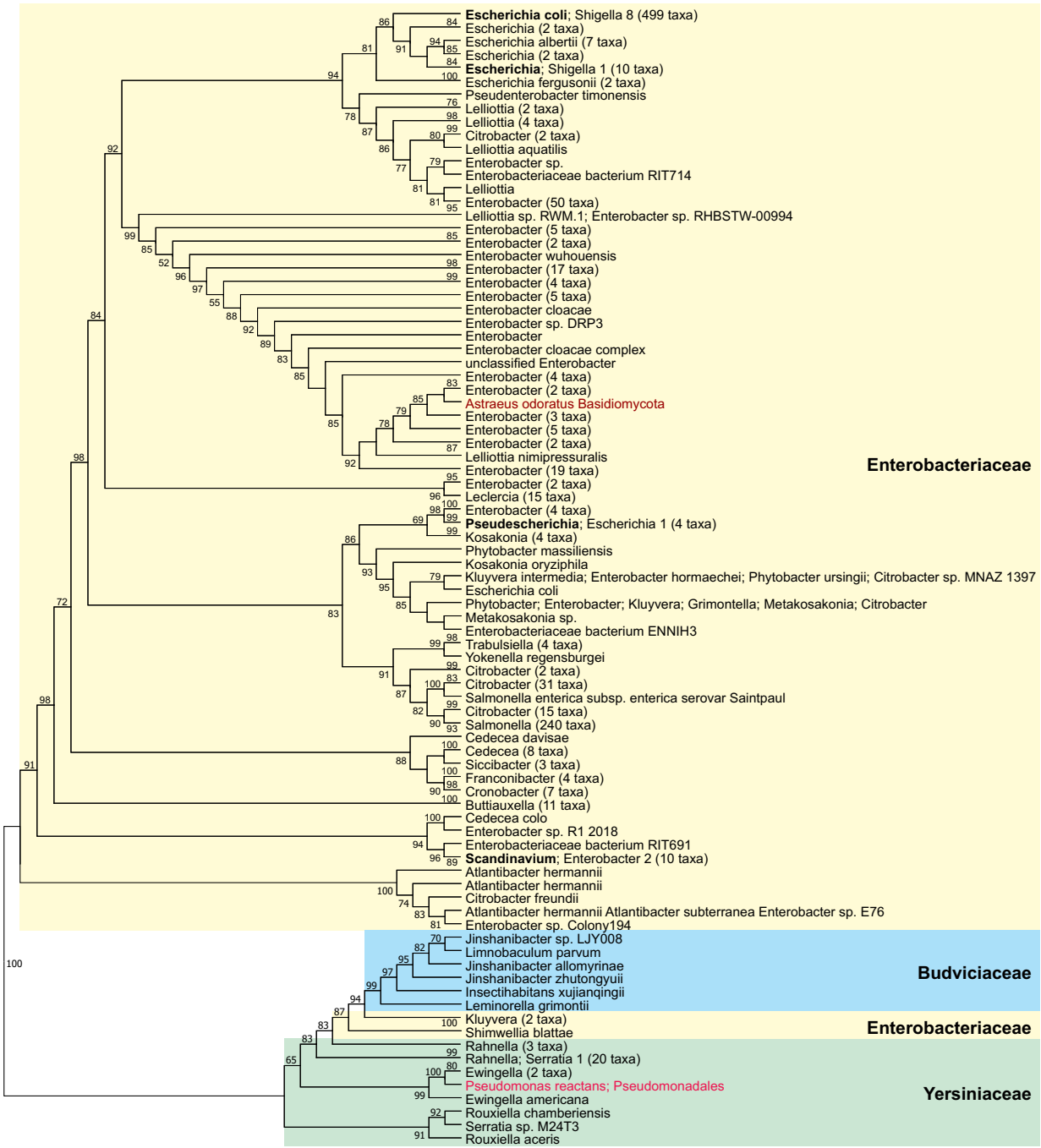

Figure 33: **The maximum likelihood tree based on the alignment of 1083 sequences derived from clusters C3, C5, and C6 shows a close similarity to the reference CsgA sequence.** Numbers at nodes correspond to support values calculated by SH-aLRT procedure. Many branches including representatives of the same genus or species were compressed. The most abundant taxon was bolded. The number of taxa in minority was shown, as well as the total number of sequences in the clade.

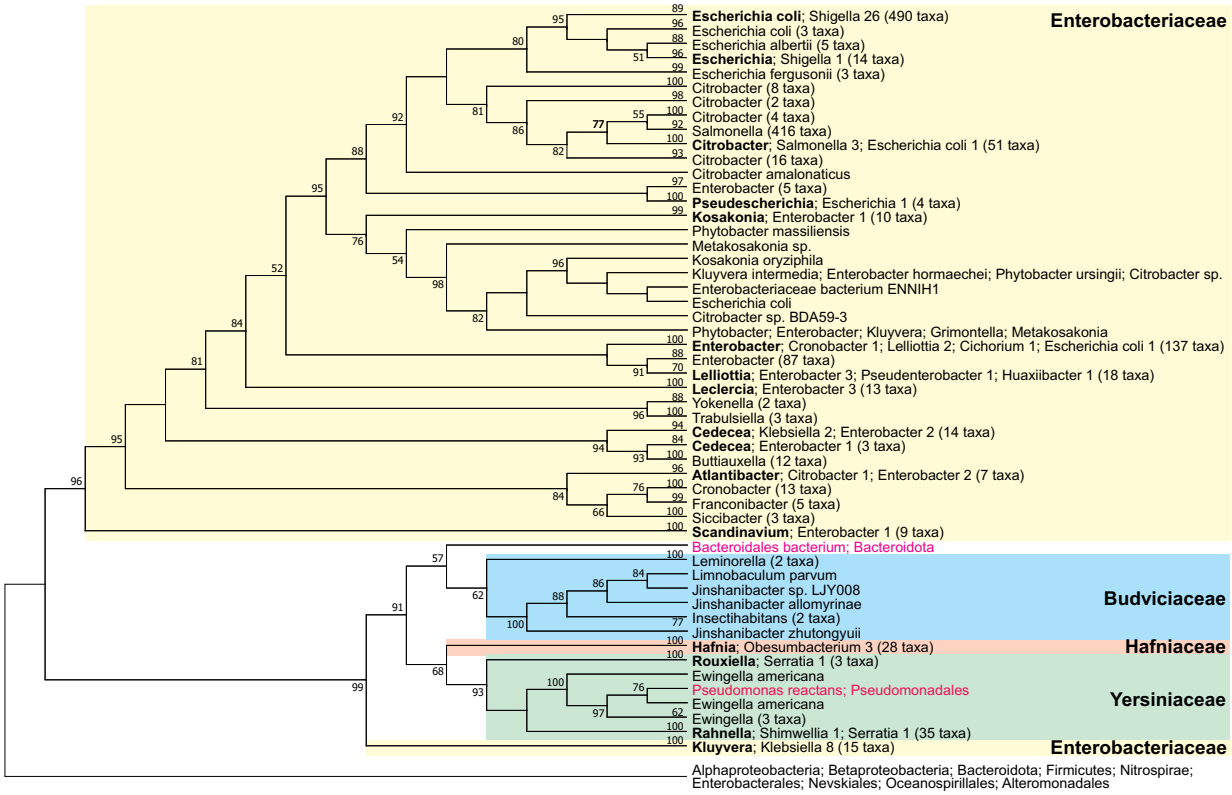

Figure 34: **The maximum likelihood tree based on the alignment of 1083 sequences derived from clusters C1, C2, and C06 shows a close similarity to the reference CsgB sequence.** Numbers at nodes correspond to support values calculated by the SH-aLRT procedure. Many branches including representatives of the same genus or species were compressed. The most abundant taxon was bolded. The number of taxa in minority was shown, as well as the total number of sequences in the clade. Sequences obtained from the clusters contained representatives not only from *Enterobacterales* but also from other bacterial taxonomic groups, which were significantly grouped in one clade and used as an outgroup.

5.3.9 Variation of duplicated regions in *Enterobacterales* curli homologs

In order to study the evolution of repeating regions in curli proteins, we derived them from alignments of the *Enterobacterales* sequence and inferred their phylogenetic relationships based on their HMM profiles (Fig. 35). All three approaches produced very similar topology. Regions derived from a given type of curli protein, CsgA or CsgB, are grouped together. In the case of CsgA regions, R3 is closely related with R5 and R1 with R4. R2 is the sister to the latter. CsgB regions showed a different clustering. R3 grouped with R4 and next clustered with R1. R2 and R5 are joined together in two minimum evolution methods, whereas in FM R2 is sister to R3,

R4 and R1 clade. It can be also noticed that region 5 from CsgB shows the largest divergence in comparison to others. Among CsgA regions, R1 is the most divergent.

These relationships can present a potential order of duplication of these regions. The results indicate that the regions were duplicated in a different order in these two curli proteins and regions in one protein are more closely related with themselves than with regions in the other protein.

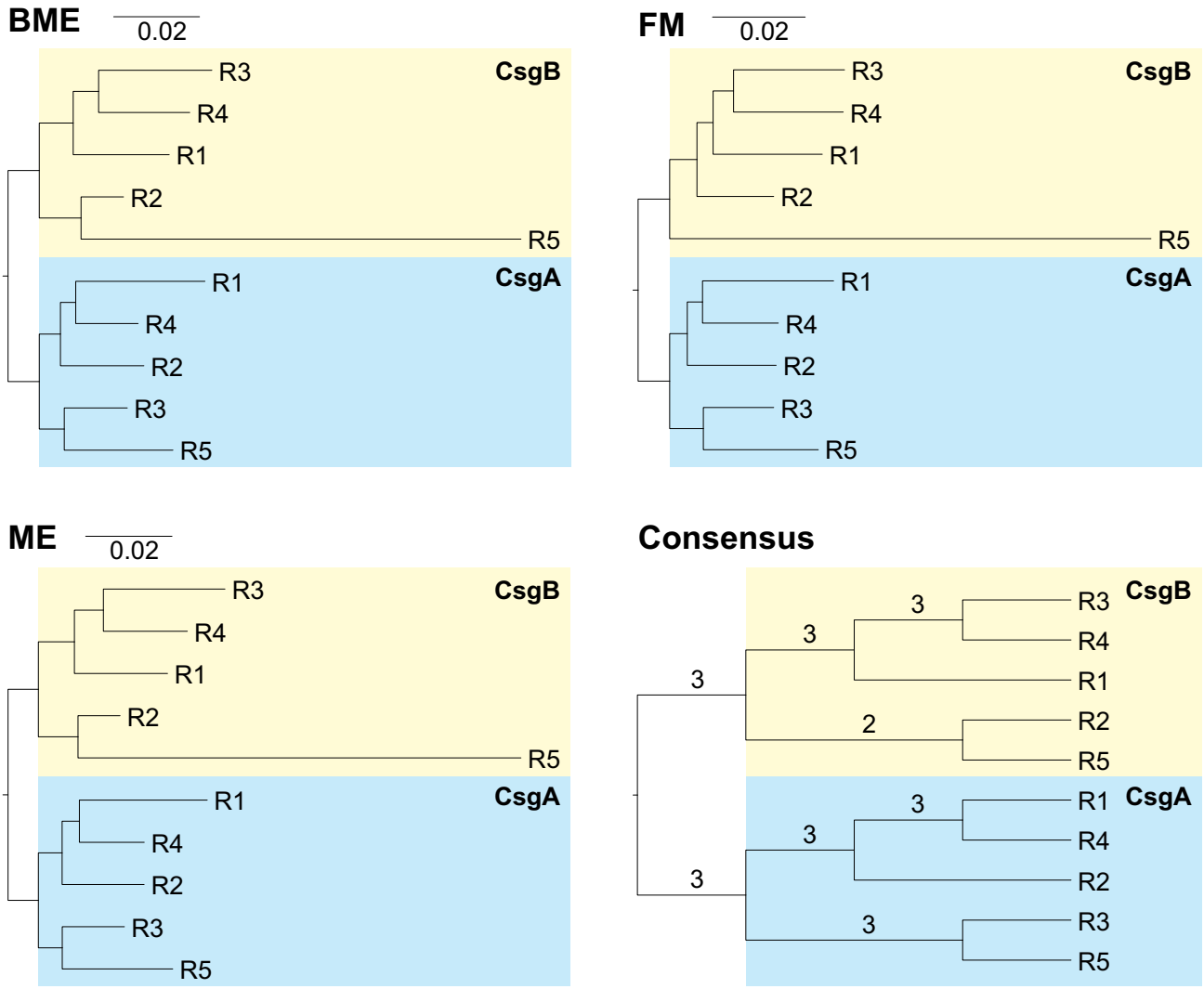

Figure 35: **Trees produced by balanced minimum evolution (BME), minimum evolution (ME) and Fitch-Margoliash criterion (FM) as well as a consensus tree based on HMM profiles grouping duplicated regions of CsgA and CsgB homologs from *Enterobacterales*.** Numbers at branches in the consensus tree indicate the number of trees, out of three, that produced a given branching pattern.

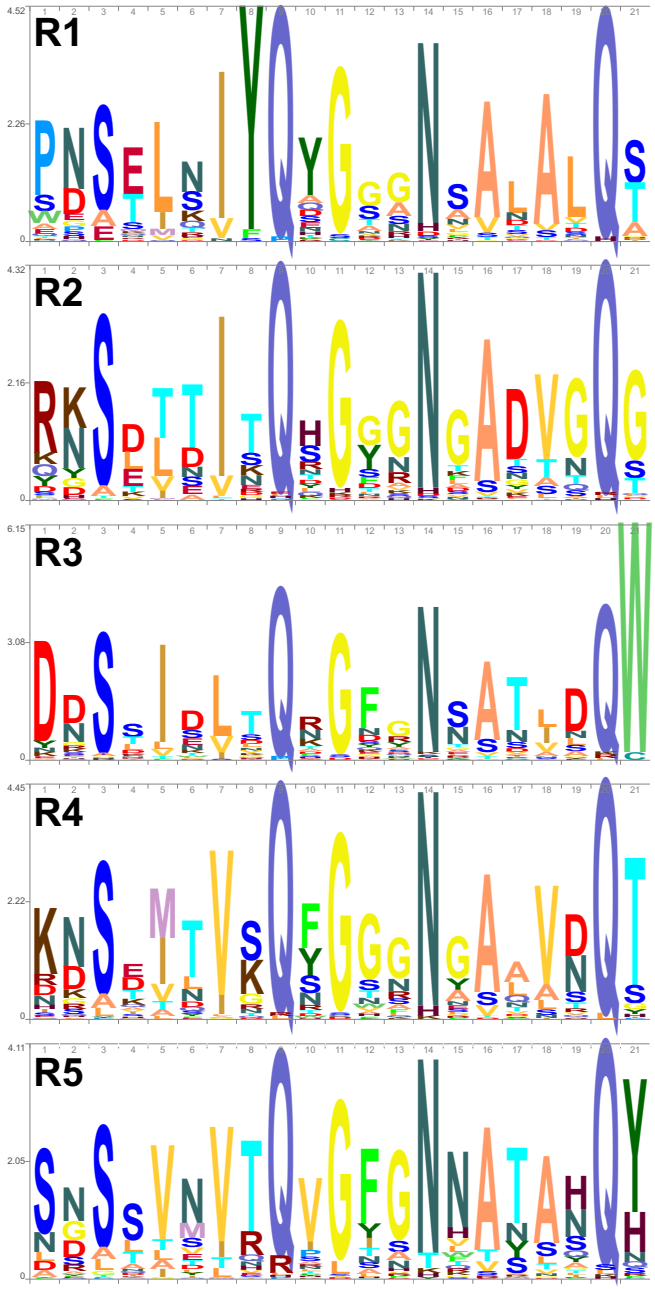

Figure 36: **Sequence logos based on HMM profiles of five duplicated regions from *Enterobacterales* CsgA sequences.**

In Fig. 36 and 37, we presented sequence logos derived from the HMM profiles of five duplicated regions from *Enterobacterales* curli proteins. All CsgA regions share conserved glutamine in the 9th and 20th positions, as well as asparagine in the 14th position. Quite conserved is also the third position with serine, the 11th position with glycine and the 16th position with alanine.

In the 13th position, there is also dominated glycine. The 7th position can be also considered conserved in respect of the presence of hydrophobic residues. Some sites share similar residues only between some regions that discriminate them from others, e.g., R1, R3, R4 and R5 have mostly hydrophobic residues, whereas R2 polar tyrosine in the 5th position; R3 and R5 have phenylalanine, whereas R1, R2 and R4 glycine in the 12th position; R2, R3 and R5 have threonine, whereas R1 tyrosine and R4 serine in the 8th position. The positions 6, 15, and 18 can group some regions in pairs containing the same dominated amino acid: R2+R4 and R1+R5; R1+R3 and R2+R4; R2+R4 and R1+R5, respectively. The same can be applied to positions 17 and 19, which can cluster R3+R5 and R3+R4, respectively. Three positions, i.e., 1, 10, and 21 are unique for each region in terms of the most common residue.

The CsgB regions also contain conserved residues in many positions, i.e., glutamine in the 10th and 21st positions, alanine in the 6th position, hydrophobic residues in the 8th and 19th positions, as well as polar residues in the 13th position. In three positions, i.e., 12, 15 and 17, there is the same dominant amino acid in four regions, from R1 to R4. They have predominantly glycine, asparagine and alanine, whereas R5 has glutamine, methionine and isoleucine, respectively. In turn, in the 7th position R1, R2, R4 and R5 contain hydrophobic residues, whereas R3 polar tyrosine. Glycine is also dominated in the first and 14th positions in R1, R2 and R4 as well as R2, R3 and R5, respectively. The 5th position can cluster R4 with R5 due to common threonine and R2 with R3 due to common leucine. In the 4th position, R3 and R4 share asparagine in contrast to others. The same amino acid is also present in R4 and R5 in positions 9 and 20. Six positions, 2, 3, 11, 16, 18, and 21, are distinct across all these regions in terms of preferred residues.

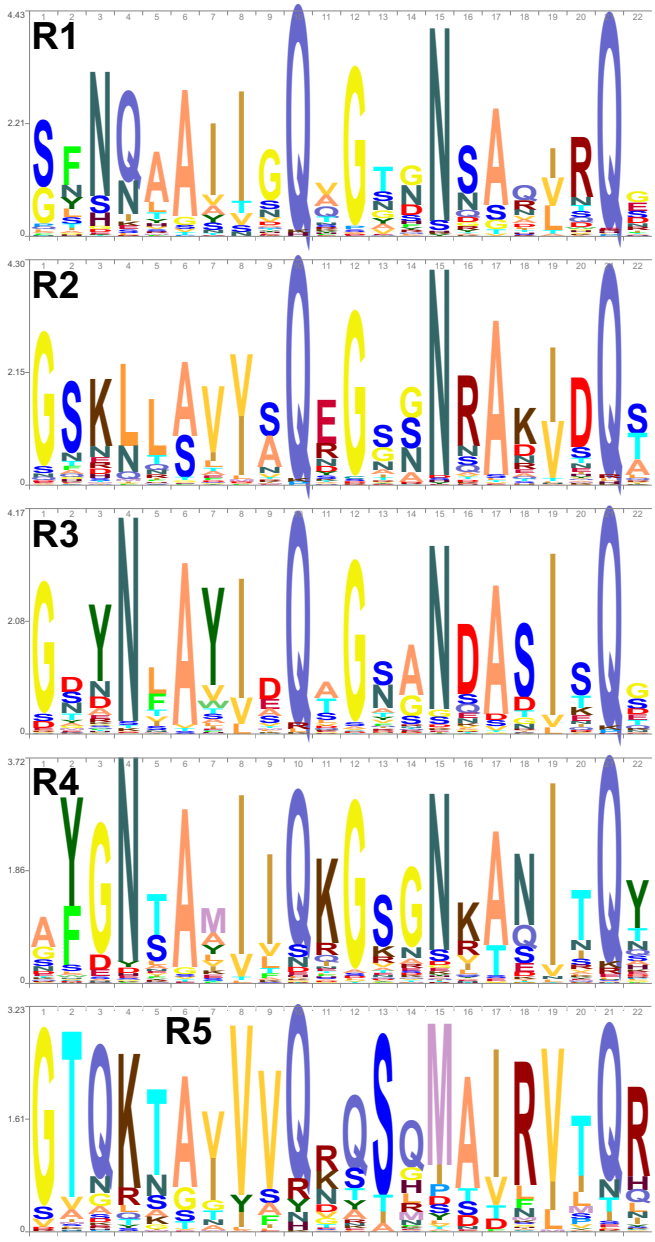

Figure 37: **Sequence logos based on HMM profiles of five duplicated regions from *Enterobacterales* CsgA sequences.**

Pairwise comparisons of *Enterobacterales* sequences for the individual regions indicate that they evolved at a different rate (Tab. 16). Generally, CsgA regions showed a larger variation (median 0.24) than CsgB (0.23). Considering the individual regions in CsgA homologs, the smallest fraction of different positions (p-distance) revealed region R5, with a median of 0.19 (Fig. 38). R2 regions were more different (0.24), R1 and R3 showed identical median values of

0.29, whereas the most divergent occurred in R4 (0.38). In the case of CsgB regions, R5 and R4 accumulated the smallest number of substitutions (Fig. 38). Their median value was the same, i.e. 0.14. A greater distance showed R1 and R3, i.e. 0.27, whereas the largest difference was R2 (0.3). All differences were statistically significant with p-value $<$ 2.2e-16.

Table 16: **Median and 25% and 75% quartiles for pairwise distance (i.e. a fraction of different positions) between regions of CsgA and CsgB homologs.**

Region	CsgA	CsgB
R1	0.286 [0.048-0.381]	0.273 [0.182-0.364]
R2	0.238 [0.048-0.286]	0.318 [0.182-0.409]
R3	0.286 [0.048-0.333]	0.273 [0.182-0.364]
R4	0.381 [0.095-0.429]	0.136 [0.091-0.227]
R5	0.191 [0.048-0.333]	0.136 [0.046-0.182]
All	0.238 [0.095-0.381]	0.227 [0.091-0.364]

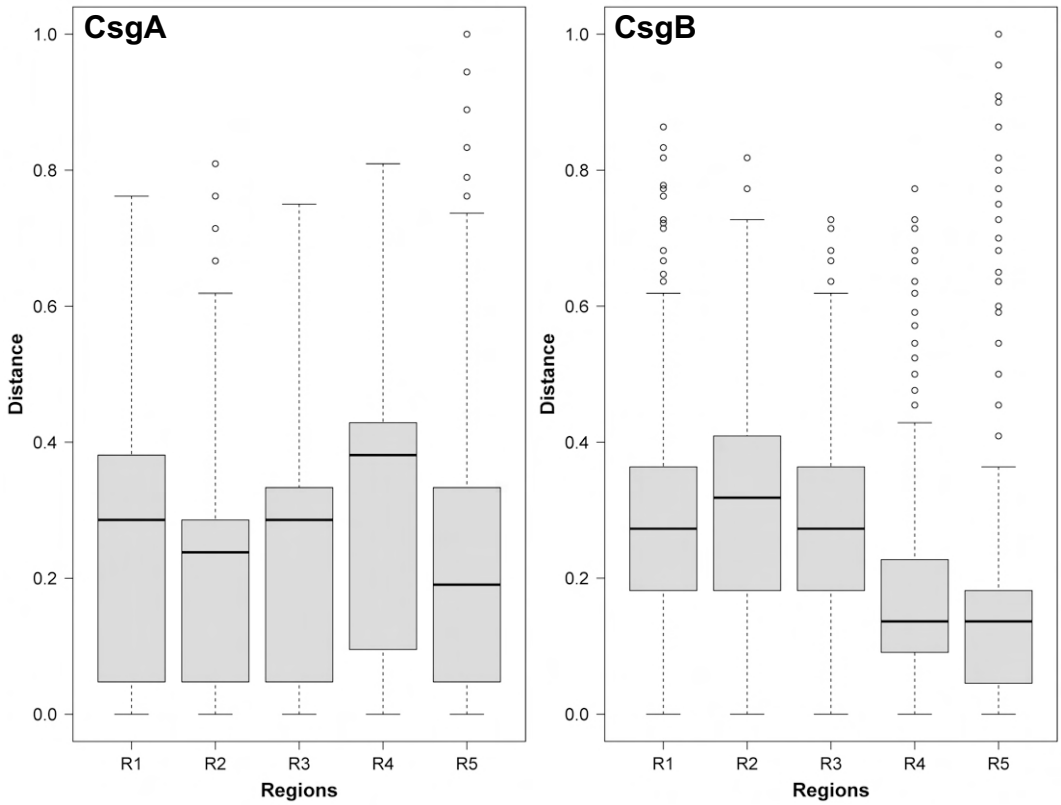

Figure 38: **Box-plots of pairwise distance (i.e. fraction of different positions) between sequences for the individual regions of CsgA and CsgB homologs.** The thick line indicates the median, the box shows the quartile range, and the whiskers denote the range without outliers.

We also checked how the differences between the regions are correlated during evolution (Tab. 17). Interestingly, they showed quite high significant correlations. The largest correlations showed the regions in CsgA homologs, from 0.83 to 0.89, whereas in CsgB homologs, they were smaller, from 0.76 to 0.84. The most correlated occurred R1 with R3 as well as R3 with R4 in CsgA, whereas in CsgB, R1 with R4 and R3 with R4. All these correlations were statistically significant with p-value $<$ 2.2e-16. Generally, correlations between adjacent regions are larger than those between more distant regions. The median for these regions type is 0.875 vs 0.866 for CsgA and 0.793 vs 0.775 for CsgB. The more coordinated evolution in CsgA regions can be related to more important interactions between them in this protein. The interactions of individual regions in CsgA are necessary to create amyloid fibrils, whereas CsgB is only an initiator of this process [Hammer et al., 2007, Shu et al., 2012]. All these correlations were statistically significant with p-value $<$ 2.2e-16.

Table 17: **Spearman correlation coefficients between p-distances (fraction of different positions) calculated for pairwise region comparisons for CsgA (the upper triangle) and CsgB homologs (the lower triangle).**

	R1	**R2**	**R3**	**R4**	**R5**
R1	-	0.883	0.892	0.884	0.865
R2	0.800	-	0.868	0.867	0.862
R3	0.782	0.762	-	0.890	0.833
R4	0.837	0.757	0.844	-	0.850
R5	0.759	0.800	0.767	0.786	-

6 Structural variability of CsgA and CsgB variants

6.1 Research objectives

The other goal of this investigation was to verify the importance of the various repeating units of functional amyloids, i.e., CsgA and CsgB proteins, participating in amyloid fibril formation (Fig. 3). We would like to find out which regions are involved in aggregation and how they affect its rate. In addition, we wanted to verify if the different variants of these proteins have altered aggregation characteristics and how the reactions between these regions can change. To achieve that, we designed, manufactured, and purified selected CsgA and CsgB variants, which were verified by ThT assay and AFM.

6.2 Materials and Methods

6.2.1 Cloning of csgA and csgB

Six variants of CsgA (Tab. 18) and six of CsgB (Tab. 19) proteins were selected according to the bioinformatic results and scientific papers. Based on other work [Wang et al., 2008], we decided to use proteins with deletions of particular regions that, although very similar, do not have the same functions. We then analyzed the sequences we designed using our AmyloGram predictor to see if these proteins still have amyloidogenic properties. We then wanted to validate this prediction experimentally and check on HDX-MS which regions interact with each other. Besides the wild type (WT), we studied variants that are characterized by the deletion of one of the five repeating units. Moreover, in each protein, we removed the signal peptide region and added His-tag to be able to purify it on the column. Based on them, appropriate nucleotide sequences were prepared, which were cloned and expressed by standard genetic procedures in *Escherichia coli* BL21 strain [Zhou et al., 2012b, Andreasen et al., 2019b]. In the case of deletions in the curli gene region, we used the PCR overlapping technique [Bryksin and Matsumura, 2010] and PIPE (Polymerase Incomplete Primer Extension) technique [Klock and Lesley, 2009].

Table 18: **List of studied variants of CsgA proteins.**

Variant	**Sequence**
CsgA WT	GVVPQYGGGGNHGGGGNNSGPNSELNIYQYGGGNSALALQT DARNSDLTITQHGGGNGADVGQGSDDSSIDLTQRGFGNSATL DQWNGKNSEMTVKQFGGGNGAAVDQTASNSSVNVTQVGFG NNATAHQYEHHHHHH
CsgA ΔSPΔR1 6xHis	GVVPQYGGGGNHGGGGNNSGPNSDLTITQHGGGNGADVGQG SDDSSIDLTQRGFGNSATLDQWNGKNSEMTVKQFGGGNGAA VDQTASNSSVNVTQVGFGNNATAHQYEHHHHHH
CsgA ΔSPΔR2 6xHis	GVVPQYGGGGNHGGGGNNSGPNSELNIYQYGGGNSALALQT DARNSSIDLTQRGFGNSATLDQWNGKNSEMTVKQFGGGNGA AVDQTASNSSVNVTQVGFGNNATAHQYEHHHHHH
CsgA ΔSPΔR3 6xHis	GVVPQYGGGGNHGGGGNNSGPNSELNIYQYGGGNSALALQT DARNSDLTITQHGGGNGADVGQGSDDSEMTVKQFGGGNGA AVDQTASNSSVNVTQVGFGNNATAHQYEHHHHHH
CsgA ΔSPΔR4 6xHis	GVVPQYGGGGNHGGGGNNSGPNSELNIYQYGGGNSALALQT DARNSDLTITQHGGGNGADVGQGSDDSSIDLTQRGFGNSAT LDQWNGKNSSVNVTQVGFGNNATAHQYEHHHHHH
CsgA ΔSPΔR5 6xHis	GVVPQYGGGGNHGGGGNNSGPNSELNIYQYGGGNSALALQT DARNSDLTITQHGGGNGADVGQGSDDSSIDLTQRGFGNSAT LDQWNGKNSEMTVKQFGGGNGAAVDQTASNEHHHHHH

To extract *E. coli* genomic DNA, the colony was added to 35 μl QuickExtract DNA Extraction Solution (Lucigen), mixed by vortexing, and transferred to a heat block at 65°C for 6 minutes and 98°C for 2 minutes. Genomic DNA was used to amplify the curli gene sequence in PCR, without the region coding for the signal peptide. Specific primers with overlaps for restriction enzymes were used.

The PCR product was purified with QIAquick PCR Purification Kit (Qiagen). For the construction of protein variants with removed selected regions overlapping PCR and PIPE methods were used. Primers had a length of approximately 30 nt. All primers that were used in the reactions are included in Tab. 20 and 21. The PCR products were examined on an agarose gel, and proper length bands were extracted with GeneJET Gel Extraction Kit (Thermo).

Table 19: **List of studied variants of CsgB proteins.**

Variant	**Sequence**
CsgB WT	AGYDLANSEYNFAVNELSKSSFNQAAIIGQAGTNNSAQLRQGG SKLLAVVAQEGSSNRAKIDQTGDYNLAYIDQAGSANDASIS QGAYGNTAMIIQKGSGNKANITQYGTQKTAIVVQRQSQMAI RVTQREHHHHHH
CsgB ΔSPΔR1 6xHis	AGYDLANSEYNFAVNELSKSSFNLLAVVAQEGSSNRAKIDQTG DYNLAYIDQAGSANDASISQGAYGNTAMIIQKGSGNKANITQ YGTQKTAIVVQRQSQMAIRVTQREHHHHHH
CsgB ΔSPΔR2 6xHis	AGYDLANSEYNFAVNELSKSSFNQAAIIGQAGTNNSAQLRQGG SKNLAYIDQAGSANDASISQGAYGNTAMIIQKGSGNKANITQ YGTQKTAIVVQRQSQMAIRVTQREHHHHHH
CsgB ΔSPΔR3 6xHis	AGYDLANSEYNFAVNELSKSSFNQAAIIGQAGTNNSAQLRQGG SKLLAVVAQEGSSNRAKIDQTGDYNTAMIIQKGSGNKANITQ YGTQKTAIVVQRQSQMAIRVTQREHHHHHH
CsgB ΔSPΔR4 6xHis	AGYDLANSEYNFAVNELSKSSFNQAAIIGQAGTNNSAQLRQGG SKLLAVVAQEGSSNRAKIDQTGDYNLAYIDQAGSANDASISQ GAYGKTAIVVQRQSQMAIRVTQREHHHHHH
CsgB ΔSPΔR5 6xHis	AGYDLANSEYNFAVNELSKSSFNQAAIIGQAGTNNSAQLRQGG SKLLAVVAQEGSSNRAKIDQTGDYNLAYIDQAGSANDASISQ GAYGNTAMIIQKGSGNKANITQYGTQEHHHHHH

The pET24d plasmid was extracted from bacteria possessing it, using GeneJET Plasmid Miniprep Kit (Thermo). The PCR products and plasmids were cut with restriction enzymes. Additionally, we used 10x FastDigest Buffer (Thermo). The cut plasmid was validated on an agarose gel and extracted using GeneJET Gel Extraction Kit (Thermo).

The cut PCR products and plasmids were ligated using T4 DNA Ligase with 10x T4 DNA Ligase Buffer (Thermo). The plasmids with genes encoding curli proteins were used to transform *E. coli* BL21 with the heat shock technique. The transformed bacteria were transferred to an LB-agar plate. After overnight culture, colony PCR was performed. To verify the successful transformation, plasmids including the correct insert were extracted with GeneJET Gel Extraction Kit (Thermo) and sent for Sanger sequencing.

Table 20: **List of primers, constructed by overlap extension method, for amplification of selected regions.**

Primer name	**Primer sequence**
csga_reg_F	ATGGGTGTTGTTCCTCAGTACGG
csga_reg1_del_R	AATAGTCAAGTCAGAATTTGGGCCGCTATTATTACCG
csga_reg1_del_F	AATAGCGGCCCAAATTCTGACTTGACTATTACCCAGCA
csga_reg2_del_R	CAGATCGATTGAGCTGTTACGGGCATCAGTTTGCA
csga_reg2_del_F	ACTGATGCCCGTAACAGCTCAATCGATCTGACCCA
csga_reg3_del_R	AACCGTCATTTCAGAGTCATCTGAGCCCTGACCA
csga_reg3_del_F	CAGGGCTCAGATGACTCTGAAATGACGGTTAAACAGTTCG
csga_reg4_del_R	CACGTTGACGGAGGAATTTTTGCCGTTCCACTGATCA
csga_reg4_del_F	TGGAACGGCAAAAATTCCTCCGTCAACGTGACT
csga_reg5_del_R	GTTAGATGCAGTCTGGTCAACTG
csga_reg_R	GTACTGATGAGCGGTCGC
reg_F_NcoI	CGGCCCATGGGTGTTGTTCCTCAGTACGG
reg5_del_R_XcoI	GCGCTCGAGGTTAGATGCAGTCTGGTCAACTG
reg_R_XcoI	GCGCTCGAGGTACTGATGAGCGGTCGC

Table 21: **List of primers, constructed by PIPE method, for amplification of selected regions.**

Primer name	Primer sequence
csgB full F NcoI	CCATGGATGAAAAACAAATTGTTATTTATGATGTTAACAA TACTGGG
csgB SP N22 r4 R	AATCATCGCAGTATTATTAAATGAAGACTTACTCAATTCATT TACCGCG
csgB r3 R1 r5 F	CAAGGTGCTTATGGTCAGGCAGCCATAATTGGTCAAGC
csgB r3 R1 r5 R	TACAATTGCCGTTTTTTTTGAGCCTCCCTGCCG
csgB r3 R2 r5 F	CAAGGTGCTTATGGTCTTTTGGCGGTTGTTGCGC
csgB r3 R2 r5 R	TACAATTGCCGTTTTATAATCTCCTGTCTGGTCAATCTTTGCC
csgB r3 R3 r5 F	CAAGGTGCTTATGGTAACCTTGCATATATTGATCAGGCG
csgB R3 R	ACCATAAGCACCTTGCGAAATAC
csgB R3 r3 R	GATCAATATATGCAAGGTTACCATAAGCACCTTGCGAAATAC
csgB r3 R3 r5 R	TACAATTGCCGTTTTACCATAAGCACCTTGCGAAATAC
csgB n22 R4 F	AAGTCTTCATTTAATAATACTGCGATGATTATCCAGAAAGG
csgB r3 R5 r5 F	CAAGGTGCTTATGGTAAAACGGCAATTGTAGTGCAGAG
csgB full R his stop PstI	CTGCAGTTAGTGGTGGTGGTGGTGGTGACGTTGTGTCACG CGAATAGC
csgB r3 R5 r5 R	TACAATTGCCGTTTTACGTTGTGTCACGCGAATAG
csgB r5 R5 F	CGCGTGACACAACGTAAAACGGCAATTGTAGTGCAGAG
v-PIPE pET24d 6xHis csgB	AACGTCACCACCACCACCACCACTG
v-PIPE pET24d N22 csgB	AGCTAAATCATAACCATGGTATATCTCCTTCTTAAAGTTAAAC
i-PIPE csgB ATG N22 pET24d	AGAAGGAGATATACCATGGGTTATGATTTAGCTAATTCAGAAT ATAACT
i-PIPE csgB 6xHis pET24d	GATCTCAGTGGTGGTGGTGGTGGTGACG

6.2.2 Expression and purification of CsgA and CsgB variants using Cobalt Resin for HDX-MS

It should be emphasized the expression and purification of amyloid proteins is a difficult task due to their quick aggregation. Therefore, for the expression and purification of these

proteins, we had to combine and modify protocols developed by Zhou et al. [2012a], Andreasen et al. [2019a] and Perov et al. [2019]. *E. coli* BL21 overnight culture was transferred to 1L TB medium and incubated in 37°C, with shaking 180 rpm to OD = 0.8-0.9. Protein expression was induced by adding IPTG to the final concentration of 0.5 mM and incubated for 1.5 h. Cells were harvested by centrifugation at 4000 g for 25 min in 4°C. The supernatant was discarded. The 35 ml of the solubilization buffer (8M GdnHCl, 50 mM potassium phosphate buffer [7.3]) was added to the pellet and sonicated for 1 min with power 10. After gentle rocking for 18-24 h at room temperature, the solution was centrifuged at 10000 g for 10 min in 4°C. The pellet was discarded.

The 1.5 ml of HisPur cobalt resin (Thermo) was added to the supernatant and incubated for 1 h at room temperature with gentle rocking. The resin was collected by centrifugation for 2 min at 700 g and transferred to a polypropylene column (Qiagen). The cobalt resin was washed with: 1) 10 ml of cold potassium phosphate buffer [7.3] (CPPB) and 2) 6 BV (bed volumes) of ice-cold 12.5 μM imidazole in CPPB or 1 BV of ice-cold 125 μM imidazole in CPPB (in the case of CsgA). The eluted proteins were loaded onto the Amicon 30 kDa concentration tube and centrifuged to remove aggregation seeds and ribosomal proteins. The example results of purification can be seen in Fig. 39, 40 and 41. Proteins were precipitated using the chloroform-methanol method and stored at -80°C.

6.2.3 CsgA expression and purification using Ni-NTA Resin

The *E. coli* BL21 overnight culture of 30 ml was added to 600 ml of LB medium with 50 mM Kanamycin. Bacterial cultures were incubated in 37°C, with shaking 180 rpm to OD = 0.6-0.7. Protein expression was induced by adding IPTG to the final concentration of 0.5 mM and incubated for 1 h. Cells were harvested by centrifugation at 6000 g for 20 min in 4°C. The supernatant was discarded, and the remaining pellet was resuspended in 15 ml 8M GdnHCl and sonicated on ice for 15 mins, 30 s sonication, 30 s break with the amplitude 40%. The solution was transferred to a falcon tube and left overnight at room temperature with gentle rocking and centrifuged at 18000 g for 20 mins at room temperature. The pellet was discarded, and the supernatant was filtered with the filter paper and 0.45 μm filter.

Ni-NTA resin was equilibrated with 1 BV of 8M GdnHCl, then the sample was added and

washed with an additional 1 BV of 8M GdnHCl. The resin was washed with: 1) 10 ml of cold potassium phosphate buffer [7.3] (CPPB), 2) 2 BV of ice-cold 12.5 mM imidazole in CPPB and 3) 2 BV of ice-cold 300 mM imidazole in CPPB. Samples were run through Amicon 30 kDa cutoff buffer and concentrated on the Amicon 10 kDa column in 4°C. The sample was desalted with Zeba Spin Desalting Columns on LC.

6.2.4 Thioflavin T assay

ThT was dissolved in 50 mM phosphate buffer with pH = 7.3 to a final concentration of 20 μM. Protein samples purified with Ni-NTA resin were diluted in the buffer mentioned above to the concentration of 0.625, 1.25, 2.5, 5, and 10 μM if it was possible. Each sample, together with the ThT control, was transferred to a 96-well plate in triplicates. Protein samples purified with Cobalt resin and dehydrated were resuspended in 8M GdnHCl buffer, sonicated, run through a 30 kDa cutoff concentrator, and desalted with desalting columns. The concentration of samples was set to approximately 4 μM. Fluorescence was measured in CLARIOstar Plus Microplate Reader at room temperature in 10 min intervals and 30s of shaking before measurements.

6.2.5 Atomic Force Microscopy

AFM measurements were performed as previously mentioned. Each protein sample was dissolved to the concentration of 2 μM. The protein solution was applied on the mica for 5 min, gently washed with 2 ml MiliQ water, and dried to change the polarity of the surface for the easier binding of negatively-charged CsgA proteins. The mica was pretreated with APTES. AFM images were processed using Gwyddion software [Nečas and Klapetek, 2012].

6.3 Results

Despite many attempts, producing and purifying enough amounts of CsgB protein and its variants have failed. We were able to identify our product after a small-scale production in only wild type variant, both on the SDS page gel and Western Blot (Fig. 41 and 42). We also tried to produce all six variants of CsgB in a greater amount but after purification through Cobalt resin, we were not able to find any product both on the SDS page and Western Blot. It means

that the proteins either stayed on the column with resin or were eluted much earlier. In the case of CsgA, we had no such problems. The wild type during small-scale production can be seen in Fig. 39 and 40.

Therefore, further analyses were carried out on variants of the CsgA protein. The planned protein analyses using HDX-MS did not take place, due to the fact that the precipitation, freezing, and resuspension of the protein purified by Cobalt resin, showed deviating results of reaction kinetics.

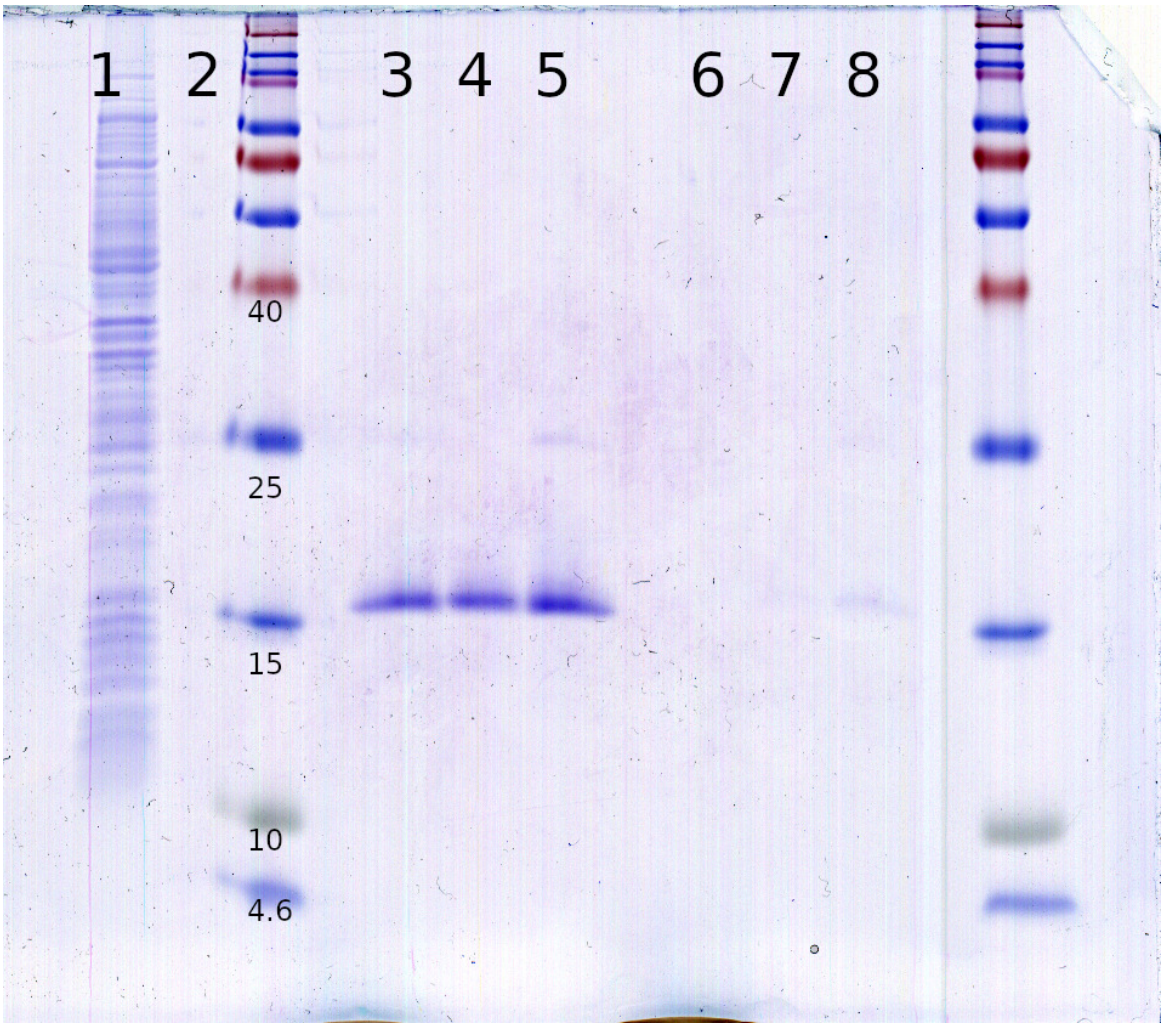

Figure 39: **Expression and purification of CsgA.** Results of CsgA (ca. 15 kDa) purification. 1) Flow through 1 ml/10 ml, 2) Flow through 10ml/10ml, 3) CsgA after the first elution, 4) CsgA from the first elution after Amicon 30 kDa concentration filter, 5) Proteins from the first elution that remained on the filter, 6) CsgA after the second elution, 7) CsgA from the second elution after Amicon 30 kDa concentration filter, 8) Proteins from the second elution that remained on the filter.

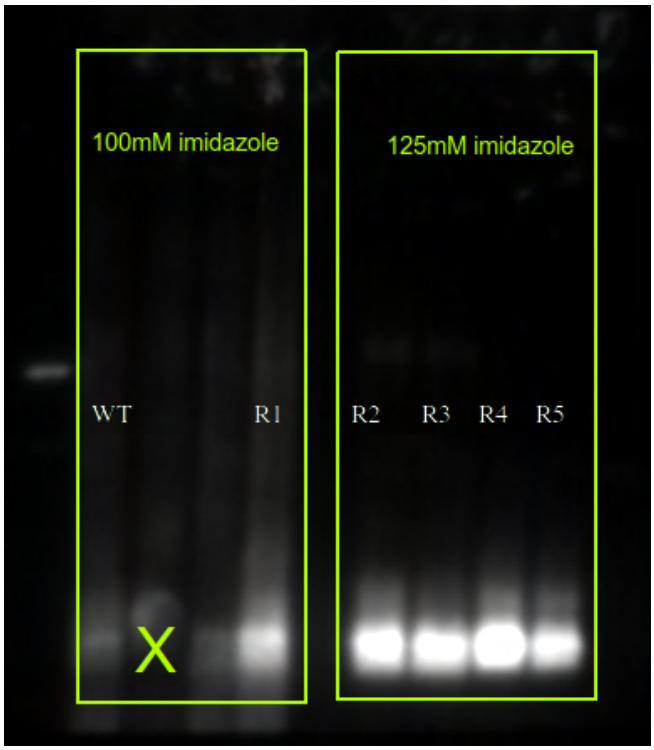

Figure 40: **Western Blot of CsgA eluates.** The samples were eluted with the buffer containing 100 and 125 mM imidazole. The CsgA bands are at the height of ∼15 kDa. X indicates a poorly formed well.

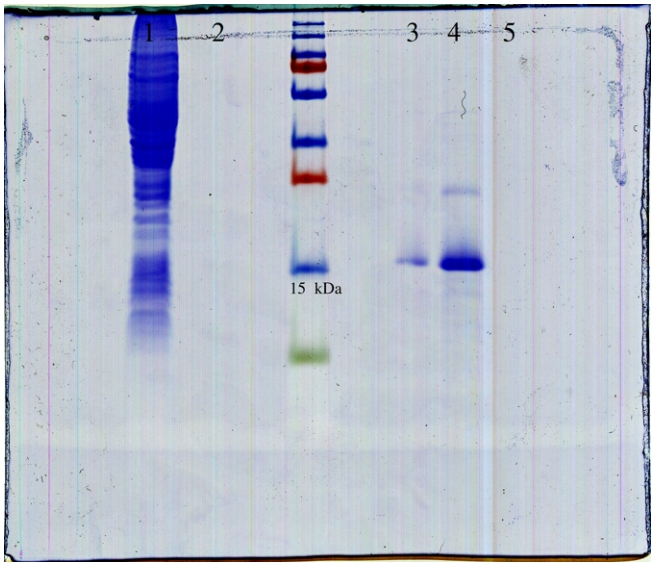

Figure 41: **Expression and purification of CsgB.** Results of CsgB (ca. 15 kDa) purification. 1) Flow through 1 ml/10 ml, 2) Flow through 10ml/10ml, 3) CsgB after the first elution, 4) CsgB from the first elution after Amicon 30 kDa concentration filter, 5) CsgB after the third elution.

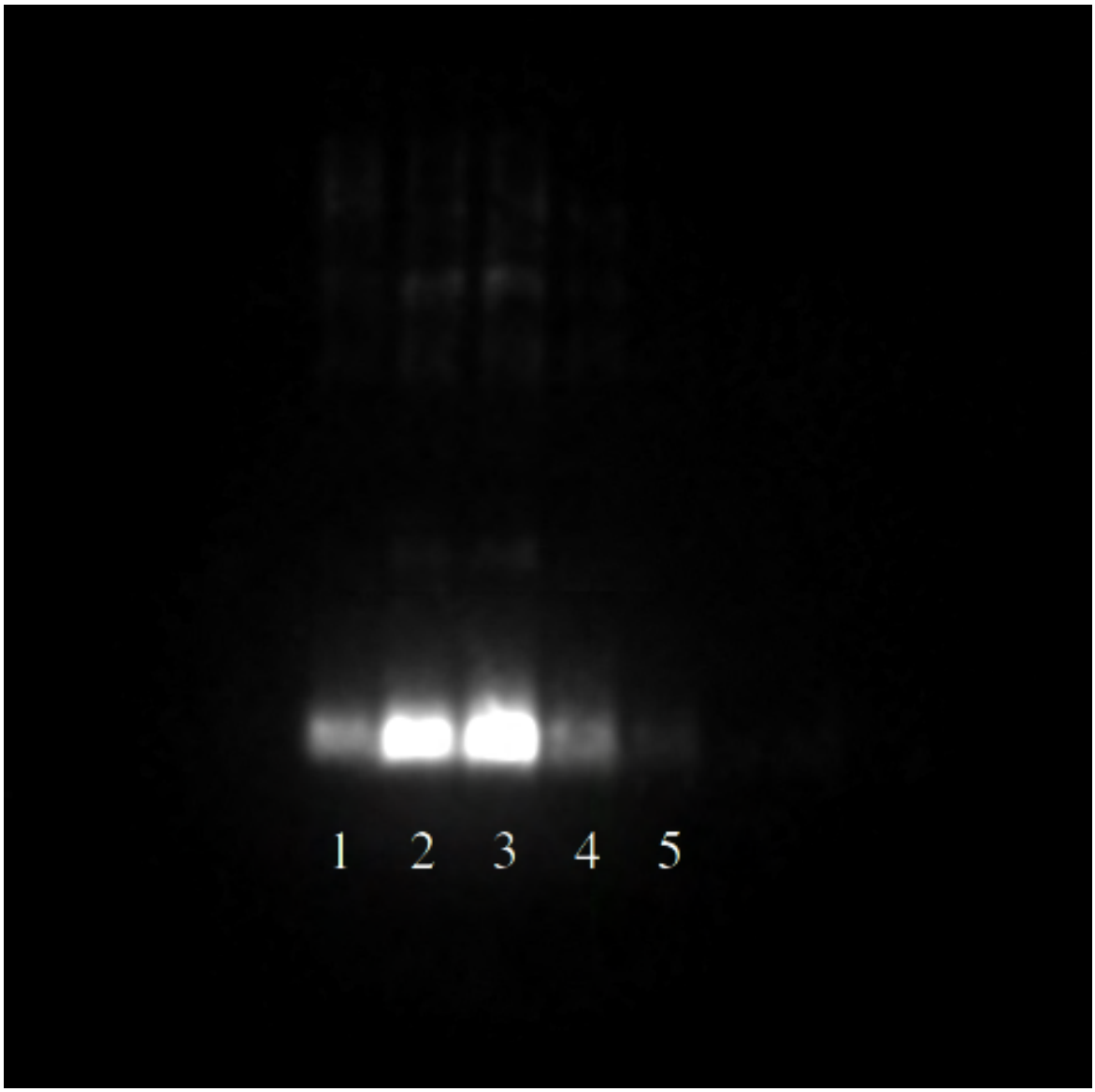

Figure 42: **Western Blot of CsgB eluates.** The samples were eluted with the buffer containing 125 mM imidazole. 1 and 4) samples from the first elution, 2 and 3) samples after concentration with Amicon 30 kDa concentration filter, 5) control with only buffer.

6.3.1 ThT assay

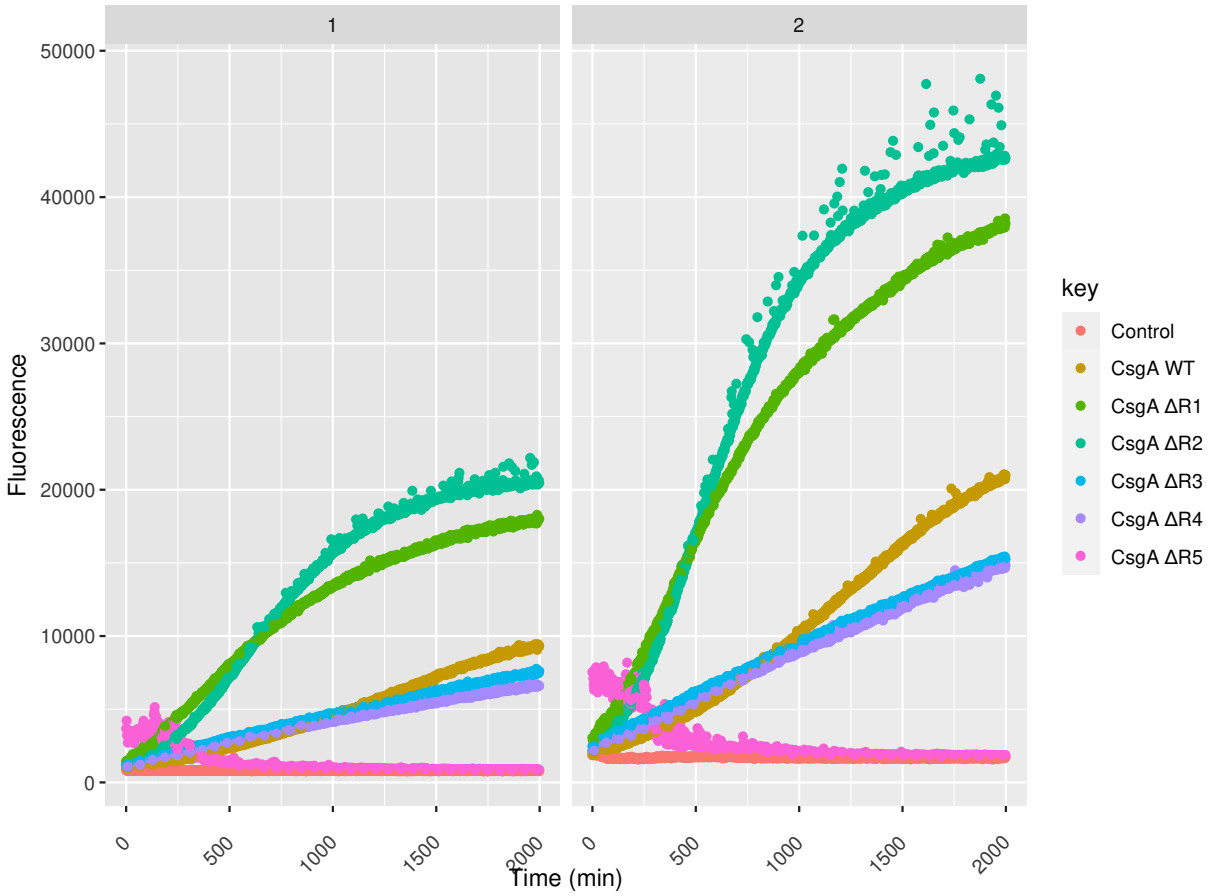

Figure 43: **ThT assay of CsgA variants purified with Cobalt resin.** Samples were run on two different plate readers marked by 1 and 2. The concentration of each sample was approximately 4 μM.

Proteins that were prepared for HDX-MS and dissolved in 8M GdnHCl show a similar aggregation curve (Fig. 43). It is likely that the proteins, despite precipitation and storage at -80 ^{o}C have aggregated. Dissolving them and then sonicating have broken down existing amyloid fibrils, but this did not result in changes in aggregation kinetics. CsgA proteins, with the exception of the ΔR5 variant, showed a similar increase in fluorescence at the same time (Fig. 43). In contrast, the ΔR5 variant showed a temporary increase in fluorescence levels, followed by a sudden decrease. Variants ΔR1 and ΔR2 demonstrated the highest fluorescence values, whereas ΔR3 and ΔR4 smaller. WT CsgA presented rather intermediate values between them.

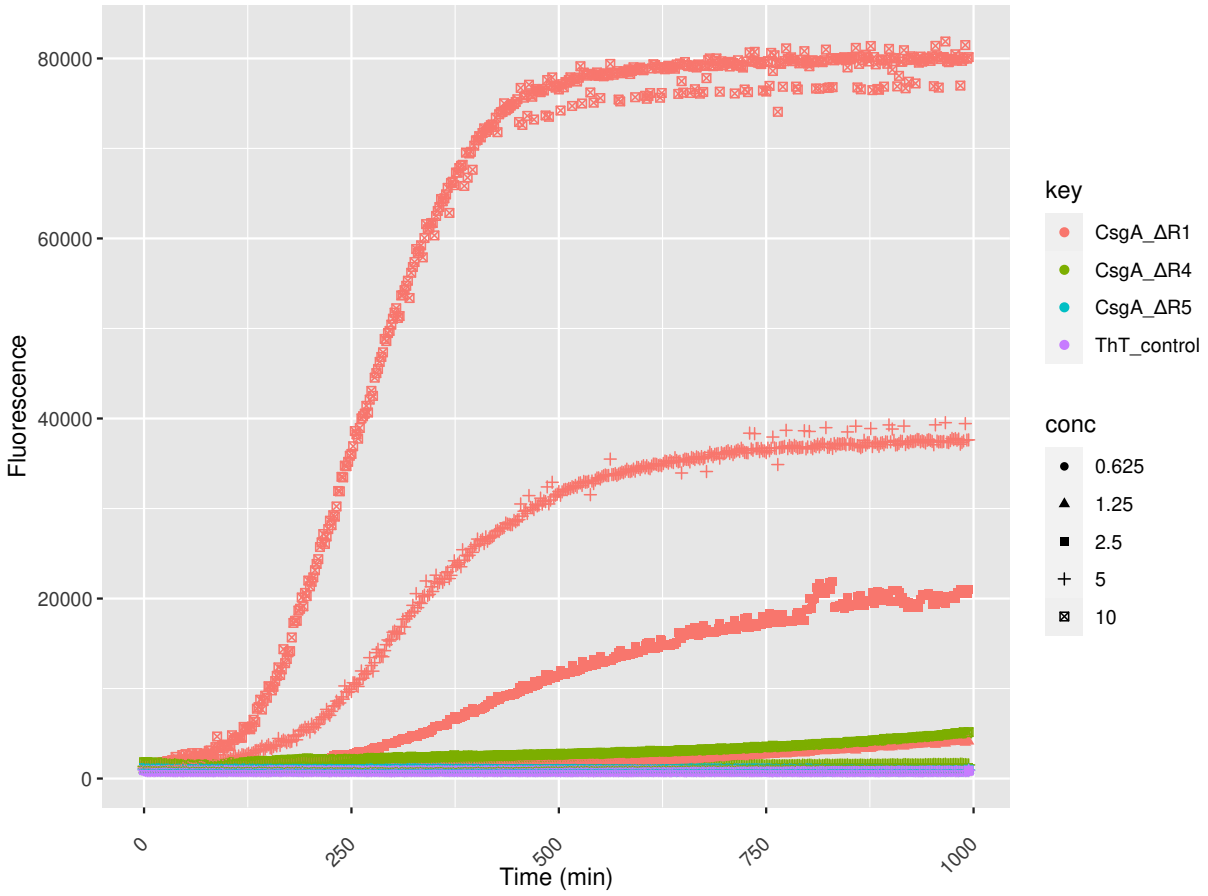

Figure 44: **ThT assay of ΔR1, ΔR4, and ΔR5 CsgA variants purified with Ni-NTA resin.** All proteins have concentrations of 0.625, 1.25, 2.5, 5, and 10 μM.

Other aggregation kinetics analyses of the ΔR1, ΔR4, and ΔR5 variants of the CsgA protein purified with the Ni-NTA deposit and tested immediately after purification were shown in (Fig. 44). Similar to the previous experiment, the ΔR1 variant occurred in the most aggregating form, which may indicate that the R1 region of the CsgA protein can probably control the speed of the aggregation process. The ΔR4 and ΔR5 variants of the protein showed no clear increase in fluorescence intensity, which may indicate that they need much more time to start the process. The presented results also clearly show that the higher the protein concentration, the aggregation process starts faster and causes a higher fluorescence.

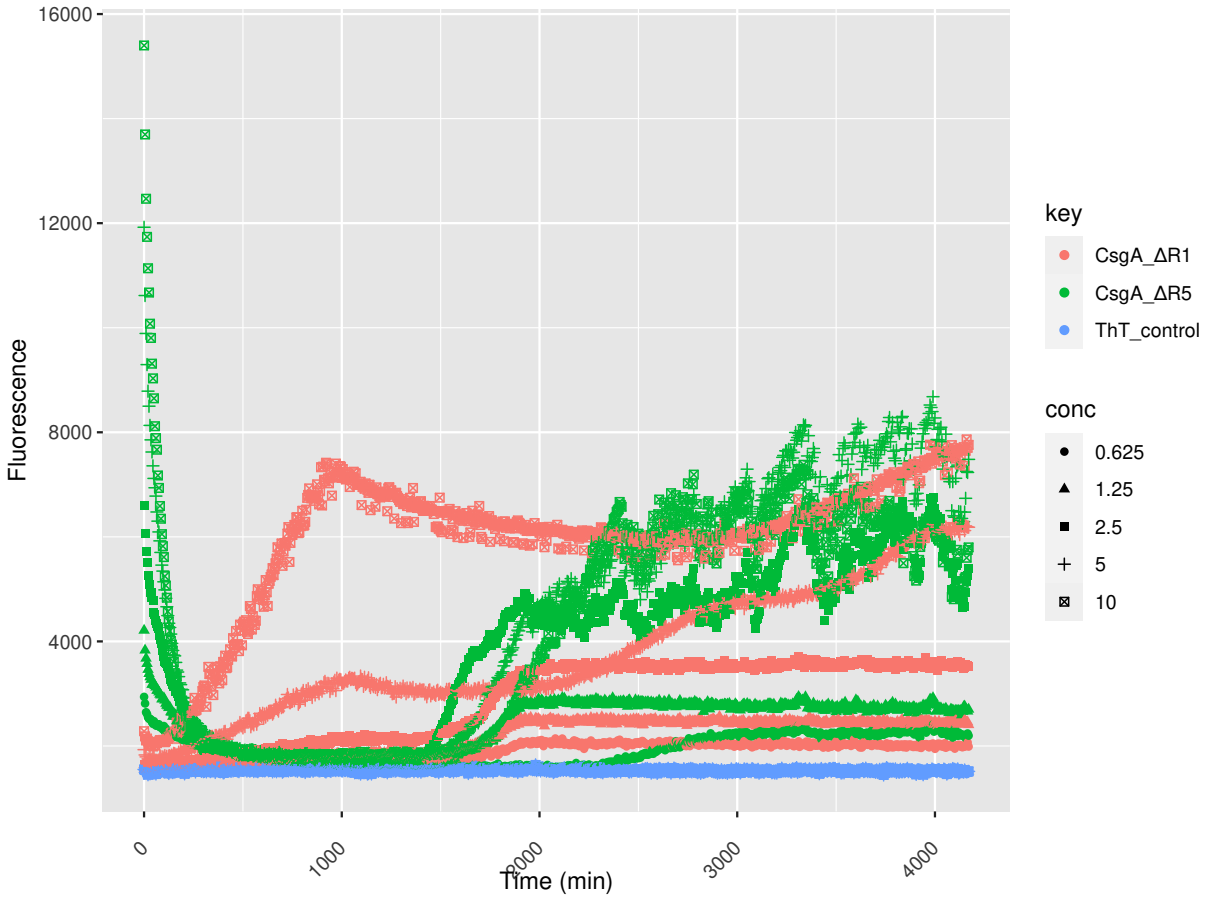

Figure 45: **ThT assay of ΔR1 and ΔR5 CsgA variants purified with Ni-NTA resin.** All proteins have concentrations of 0.625, 1.25, 2.5, 5 and 10 μM.

Re-examination of the kinetics of ΔR1 and ΔR5 variants (Fig. 45) confirmed the earlier findings. Variant ΔR5 needs significantly more time to start the aggregation process than variant ΔR1. The longer incubation needed for aggregation confirms the observations of the previous analysis, that the R5 region is much more important in this process than R1 despite the fact that both are necessary in the formation of amyloid fibrils of the CsgA protein.

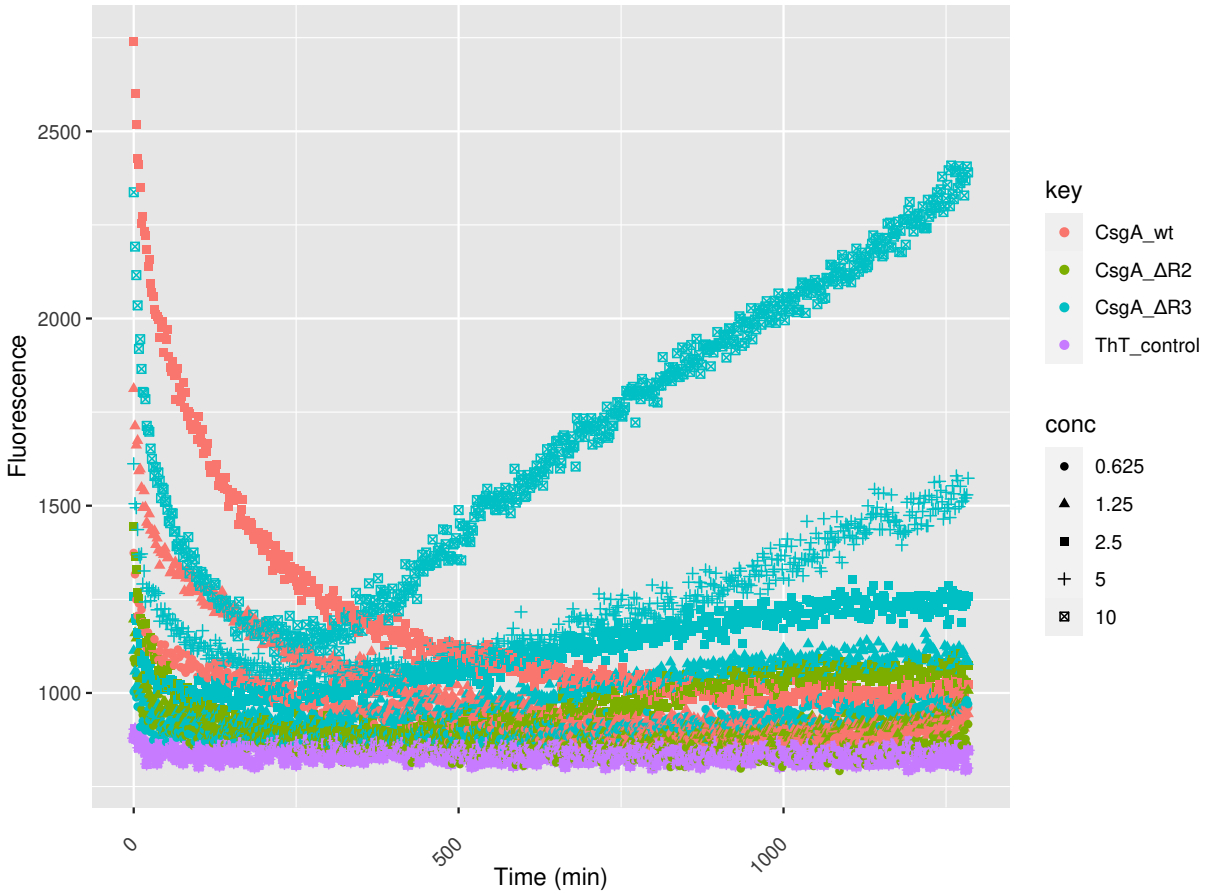

Figure 46: **ThT assay of WT, ΔR2 and ΔR3 CsgA variants purified with Ni-NTA resin.** All proteins have concentrations of 0.625, 1.25, 2.5, 5 and 10 μM.

Additional analyses of CsgA variants can be seen in Fig. 46. We can notice that the ΔR3 variant aggregates faster than the ΔR2 variant and WT. This may indicate that the R3 region of CsgA may be responsible for controlling the aggregation process or has no effect on aggregation. In contrast, the WT and ΔR2 variants of the CsgA protein need more time to aggregate.

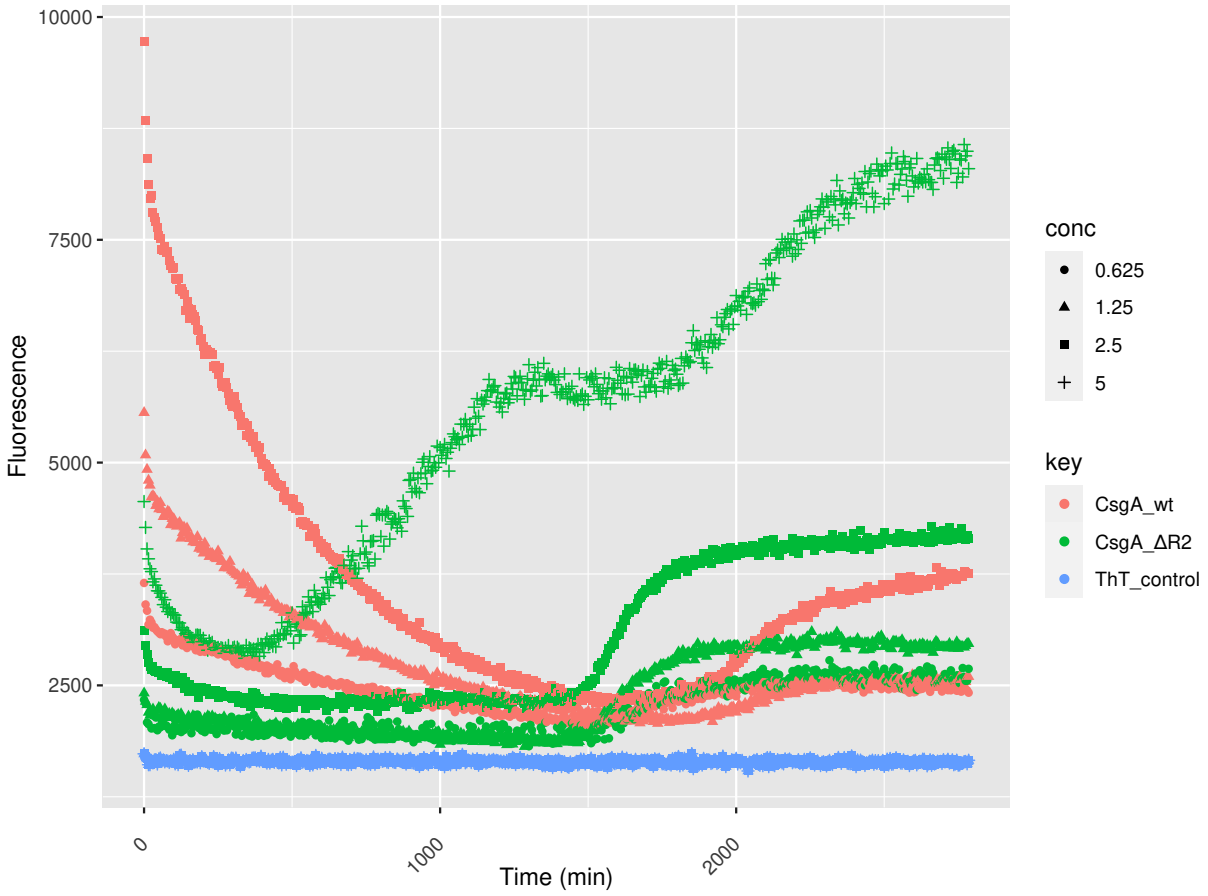

Figure 47: **ThT assay of WT and ΔR2 CsgA variants purified with Ni-NTA resin.** CsgA WT protein has concentrations of 0.625, 1.25, 2.5 μM and CsgA R2 0.625, 1.25, 2.5, 5 μM.

A re-analysis of the WT and ΔR2 variants (Fig. 47) confirms our previous observations. Yet, we were not able to obtain similar concentrations during purification. Nevertheless, the ΔR2 variant aggregated faster than in the previous analysis, which may indicate problems with the correct purification of the protein in the previous experiment. The WT variant was characterized by a longer lag phase to start the aggregation process than ΔR2 and showed a lower fluorescence, which may be related with the later start of the process.

To summarize, the kinetic studies showed that the ΔR1 variant appeared the most aggregating. It may be associated with the regulatory role of R1 region in the aggregation of CsgA protein. This region can decrease the speed of this process. On the other hand the ΔR4 variant occurred not as much reactive in aggregation as others, which may suggest that this region is more important in the aggregation. The ΔR5 is also poorly aggregated and need more time in

this process. It indicates that this region has also a decisive influence on the aggregation rate.

6.3.2 Atomic Force Microscopy

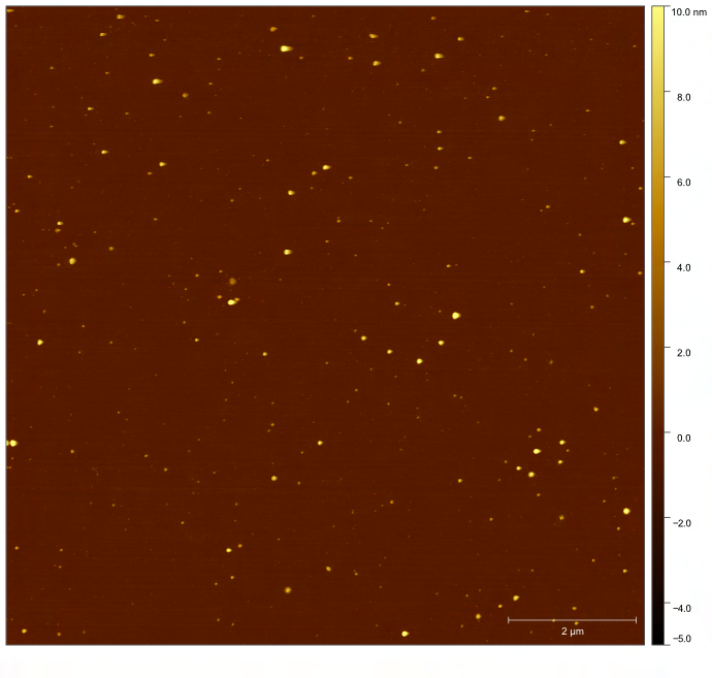

Figure 48: **CsgA WT variant after resuspension under AFM.**

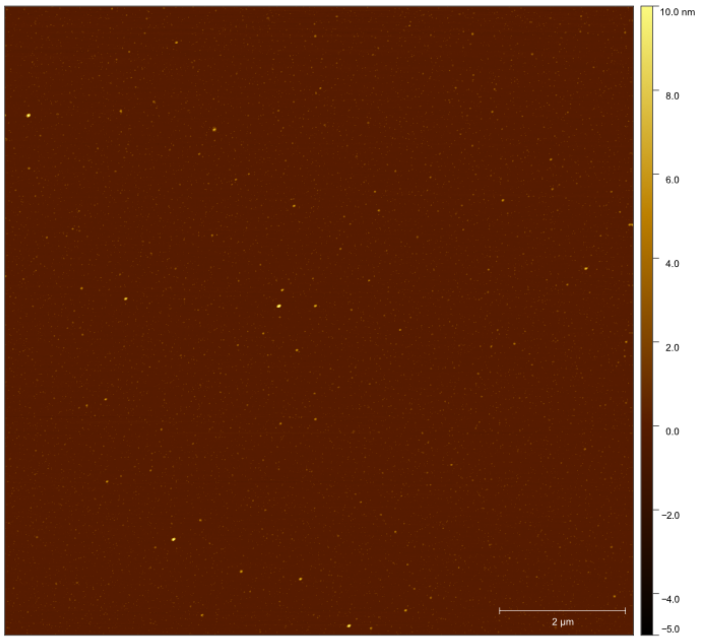

Figure 49: **CsgA ΔR1 variant after resuspension under AFM.**

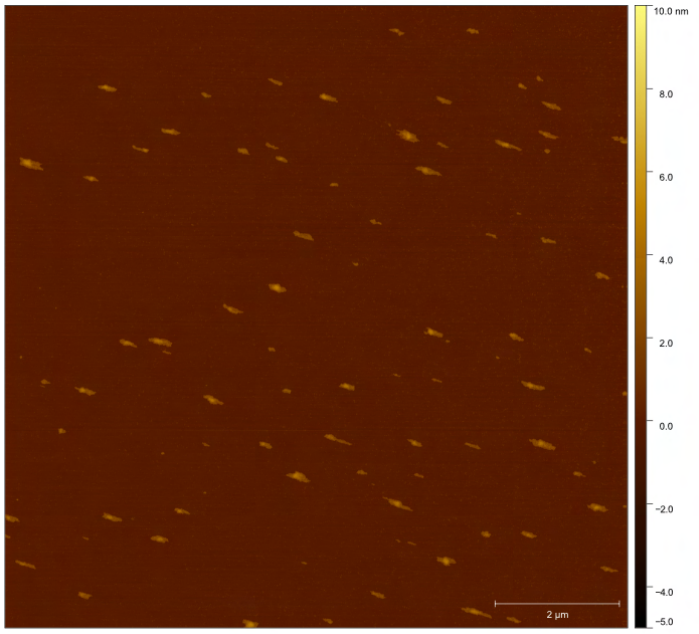

Figure 50: **CsgA ΔR2 variant after resuspension under AFM.**

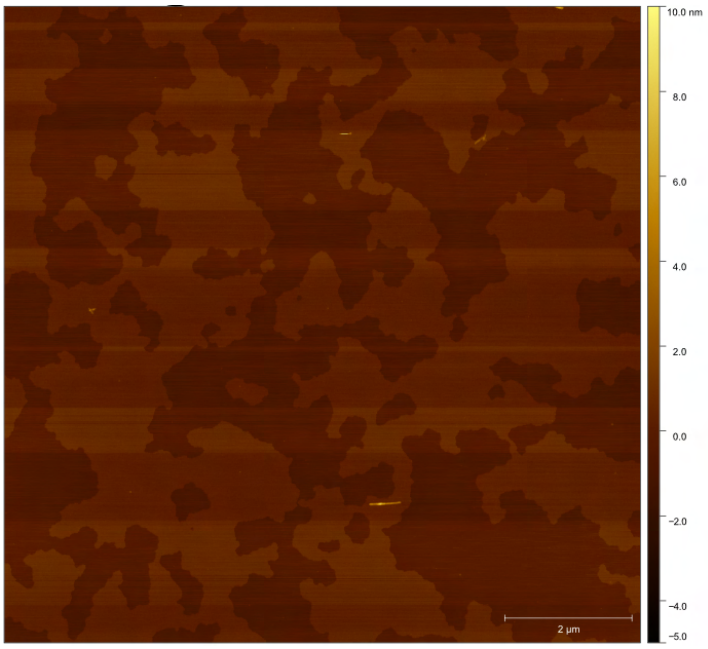

Figure 51: **CsgA ΔR3 variant after resuspension under AFM.**

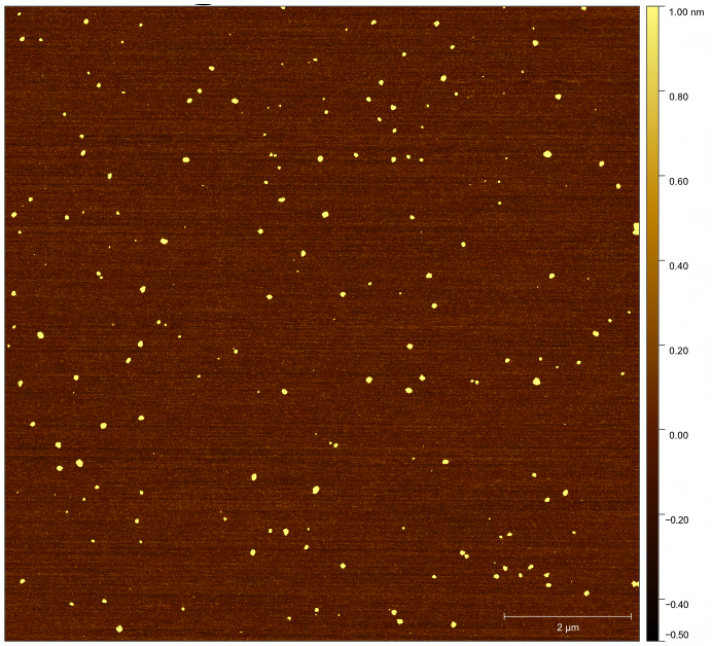

Figure 52: **CsgA ΔR4 variant after resuspension under AFM.**

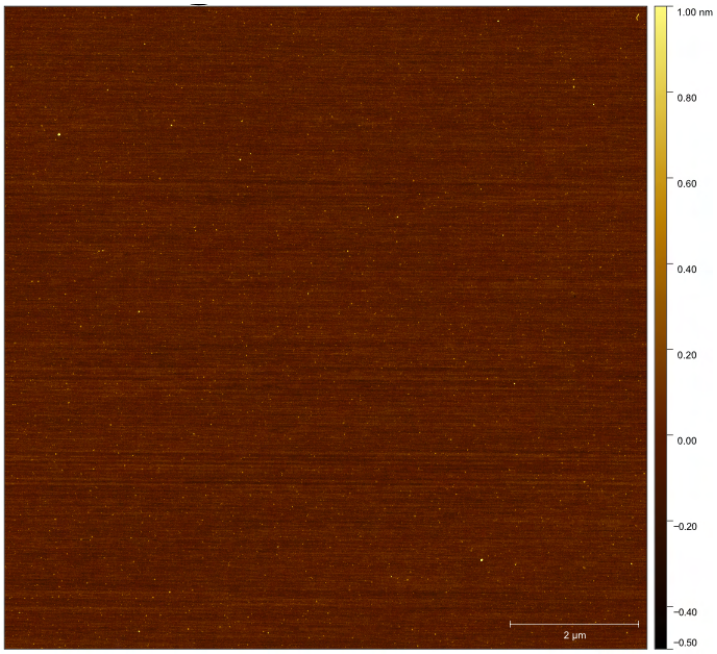

Figure 53: **CsgA ΔR5 variant after resuspension under AFM.**

Using AFM, we attempted to find amyloid fibrils of individual CsgA protein variants that were purified with cobalt resin resuspended (Fig. 48-53). Unfortunately, we did not find any amyloid fibrils. Instead, as shown in Fig. 48, 49, and 52, we observed some kind of oligomeric

aggregates. The amount of them in ΔR4 variants might indicate that there are problems with starting aggregation ass our previous observation. There is a lot of building material, but fibrils do not form. Fig. 51 might present fibrils, but they are quite small.

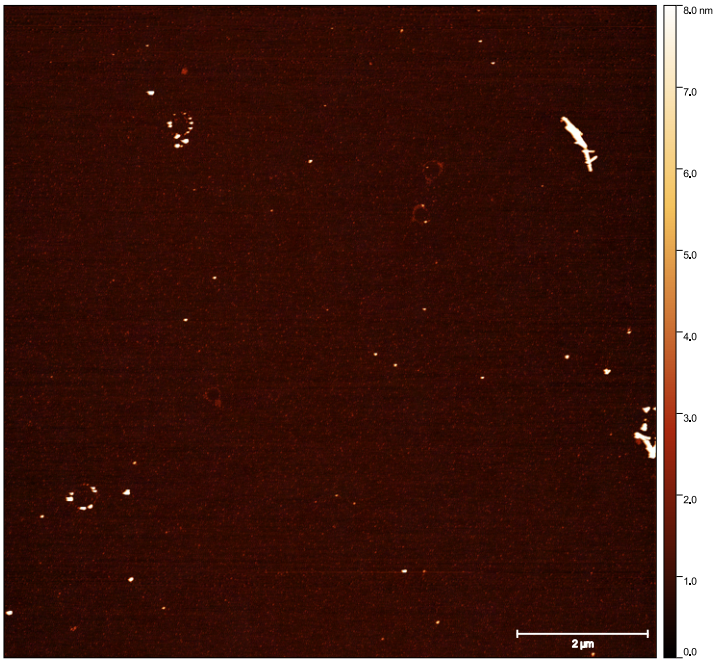

Figure 54: **CsgA R1 variant after 1 week of incubation.**

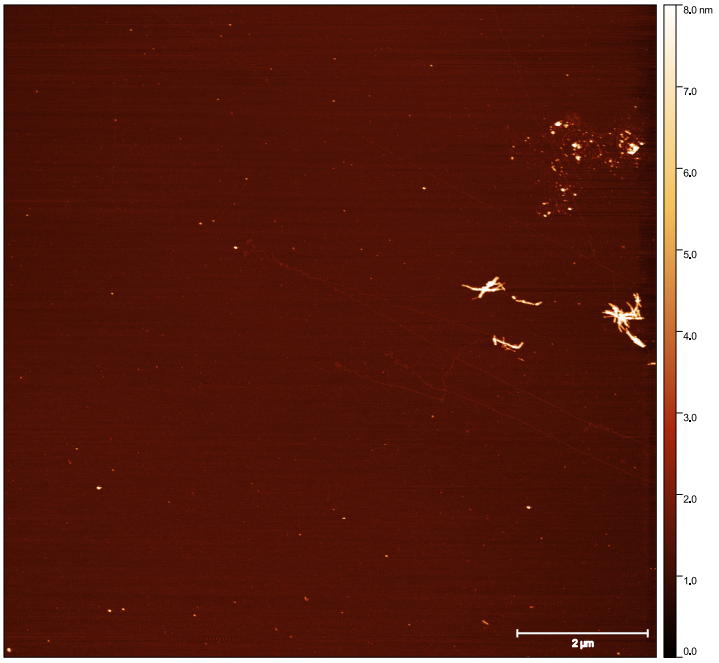

Figure 55: **CsgA R4 variant after 1 week of incubation.**

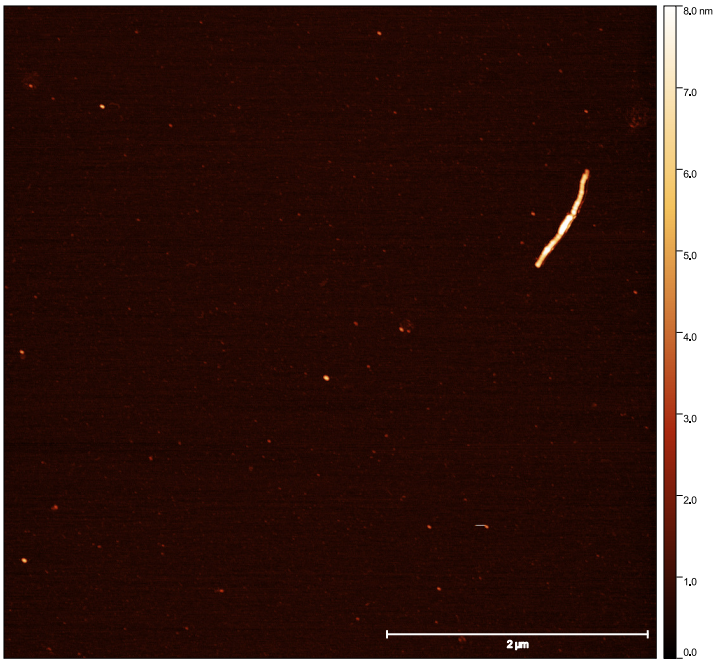

Figure 56: **CsgA R4 fibrils after 1 week of incubation.**

However, in the case of protein variants purified by Ni-NTA, we were able to find some kind of fibrils in the ΔR1 and ΔR4 variants, although the fibrils are very small, which is unusual. In Fig. 54 and 55 one can see clusters of fibrils but also numerous forms, which are probably oligomeric in the form of larger dots. In Fig. 56 a single fibril is visible, which is much shorter than amyloid fibrils in the other figures.

7 Comparison of sequence features between functional and non-functional amyloids

7.1 Research objectives

After the detailed analyses of functional amyloids, CsgA and CsgB, we decided to identify specific sequence characters that could distinguish functional amyloids in general from non-functional ones. Distinguishing these classes of amyloids can be crucial in verifying whether the dysfunctional protein plays a role in an organism or is the result of incorrect folding. So far, computational approaches for the detection of functional amyloids were not elaborated. Therefore, we compared functional and non-functional amyloids to select features that could be used in the elaboration of a robust model to distinguish these sequences.

7.2 Materials and Methods

7.2.1 Dataset preparation

We collected functional and non-functional amyloids based on literature and UniProt database [UniProt Consortium, 2018] searches. These proteins were gathered in Tab. 1, and their sequences were downloaded from the UniProt database.

For each protein, we run BLASTP [Camacho et al., 2009] to find homologous sequences in the UniProt database. In the search, we used default values: matrix auto, filter none, gapped yes, hits 1000, HSPs per hit All. The scoring matrix was automatically selected according to the sequence length and is presented in Tab. 22. From each protein, we have chosen up to top 500 significant homologs with very low E-values, 1e-50 and 1e-15, depending on a protein (Tab. 22).

The initial datasets contained 1789 functional and 5963 non-functional amyloid sequences. We rejected sequences shorter than six amino acids and used CD-HIT [Fu et al., 2012], assuming a 0.7 threshold and word size 2, to cluster them and remove similar sequences. It resulted in 1214 functional and 941 non-functional amyloids. The functional amyloids were considered as a positive set and the non-functional ones as a negative set in classification approaches.

Table 22: **List of proteins with their UniProt id used in the designing predictor of functional amyloids.** Matrix and E-value threshold used in the selection of their homologs were also included.

Protein name	**UniProt id**	**Matrix**	**E-value**
AIMP2	Q13155	BLOSUM62	1e-50
Albumin	P02768	BLOSUM62	1e-50
α-crystallin	P24623	BLOSUM62	1e-50
α-lactalbumin	P00714	BLOSUM62	1e-50
α-S2-casein	P02663	BLOSUM62	1e-50
α-synuclein	P37377	BLOSUM62	1e-50
Amyloid β	P05067	BLOSUM62	1e-50
Apolipoprotein A-I	P02647	BLOSUM62	1e-50
Apolipoprotein E	P02649	BLOSUM62	1e-50
β-casein	P05814	BLOSUM62	1e-50
β-crystallin	P53674	BLOSUM62	1e-50
β-lactoglobulin	P02754	BLOSUM62	1e-50
β-parvalbumin	P20472	BLOSUM62	1e-50
β2-microglobulin	P61769	BLOSUM62	1e-50
Bri2	Q9Y287	BLOSUM62	1e-50
CRES	O60676	BLOSUM62	1e-50
CsgA	P28307	BLOSUM62	1e-50
CsgB	P0ABK7	BLOSUM62	1e-50
Cystatin C	P21460	BLOSUM62	1e-50
Cytochrome C	P00427	BLOSUM62	1e-50
Delta-toxin	P0C1V1	BLOSUM62	1e-50
DJ-1	Q9VA37	BLOSUM62	1e-50
FapC	C4IN70	BLOSUM62	1e-20
Fibroin	P21828	BLOSUM62	1e-50
FUS	P35637	BLOSUM62	1e-50
γ-crystallin	P07315	BLOSUM62	1e-50
GroES	P0A6F9	BLOSUM62	1e-50
HET-s	Q03689	BLOSUM62	1e-50
IAPP	P10997	BLOSUM62	1e-50
Insulin	P01308	BLOSUM62	1e-50
Kappa-casein	P07498	BLOSUM62	1e-50
Lysozyme	P61626	BLOSUM62	1e-50

Table 22: **List of proteins with their UniProt id used in the designing predictor of functional amyloids. (Continued).** Matrix and E-value threshold used in the selection of their homologs were also included.

Protein name	UniProt id	Matrix	E-value
Medin (AMed)	Q08431	PAM70	1e-50
Myoglobin	P02144	BLOSUM62	1e-50
New1 (NU+)	Q08972	BLOSUM62	1e-50
p53	P04637	BLOSUM62	1e-50
p73	O15350	BLOSUM62	1e-50
Pmel17	P40967	BLOSUM62	1e-50
Polyglutamine (polyQ)	O60828	BLOSUM62	1e-50
proSP-C	P11686	BLOSUM62	1e-50
PrP	P04156	BLOSUM62	1e-50
PSMα1	A9JX05	PAM30	1e-50
PSMα2	A9JX06	PAM30	1e-50
PSMα3	A9JX07	PAM30	1e-50
PSMα4	A9JX08	PAM30	1e-50
PSMβ1	A0A068FPX1	PAM70	1e-15
PSMβ2	A0A068FLK9	PAM70	1e-15
Rnq1	P25367	BLOSUM62	1e-50
S100A9	P06702	BLOSUM62	1e-50
Sericin	P07856	BLOSUM62	1e-50
Serum amyloid A	P0DJI8	BLOSUM62	1e-50
Sup35	P05453	BLOSUM62	1e-50
Tau	P10636	BLOSUM62	1e-50
TDP-43	Q13148	BLOSUM62	1e-50
Transthyretin	P02766	BLOSUM62	1e-50
Tubulin	P0DPH7	BLOSUM62	1e-50

7.2.2 Sequence descriptors

For each sequence, we calculated several descriptors (features), i.e. various numerical representation schemes: amino acid composition (AAC), dipeptide composition (DC), AAindex, Normalized Moreau-Broto Autocorrelation (MoreauBroto), Moran Autocorrelation (Moran), Geary Autocorrelation (Geary), Composition (CTDC), Transition (CTDT), Distribution (CTDD), Conjoint Triad (CTriad), Sequence-Order-Coupling Number (SOCN), Quasi-Sequence-Order

Descriptors (QSO), Pseudo-Amino Acid Composition (PAAC) and Amphiphilic Pseudo-Amino Acid Composition (APAAC).

AAindex is set of numerical indices representing various physicochemical and biochemical properties of amino acids [Kawashima et al., 2008]. Autocorrelation descriptors (MoreauBroto, Moran and Geary) are defined based on the distribution of amino acid properties along the sequence using various types of amino acid indices. CTDC, CTDT and CTDD use amino acids grouped into classes based on hydrophobicity, normalized van der Waals volume, polarity and polarizability. They calculate their composition, transition and distribution in a sequence. CTriad was applied to model protein-protein interactions based on the classification of amino acids [Shen et al., 2007]. In this case, each protein sequence is represented by a vector space consisting of descriptors of amino acids, which are clustered into several classes according to their dipoles and volumes of the side chains. SOCN and QSO were derived from the distance matrices between amino acids, i.e. Schneider-Wrede physicochemical distance matrix [Schneider and Wrede, 1994] and Grantham chemical distance matrix [Grantham, 1974]. PAAC and APAAC use the original hydrophobicity values, the original hydrophilicity values and the original side chain masses of amino acids.

Since many indices from AAIndex database can be redundant, we removed the highly correlated indices for discriminant and prediction analyses assuming the correlation coefficient threshold 0.8 and the variance inflation factor (VIF) as the criterion for excluding variables among those that are correlated. Thereby, the number of the indices was reduced from 544 to 74. In the calculation of descriptors MoreauBroto, Moran and Geary, we assumed the following properties: CIDH920105 (Normalized Average Hydrophobicity Scales), BHAR880101 (Average Flexibility Indices), CHAM820101 (Polarizability Parameter), CHAM820102 (Free Energy of Solution in Water), CHOC760101 (Residue Accessible Surface Area in Tripeptide), BIGC670101 (Residue Volume) and CHAM810101 (Steric Parameter). Moreover, we applied the value of 15 for the maximum lag in the case of descriptors MoreauBroto, Moran, Geary, SOCN and QSO as well as for the lambda in PAAC and APAAC. These descriptors include relations between properties of amino acids located in various distances (defined by the lag and lambda) in a studied sequence.

7.2.3 Statistical and prediction analyses

Differences between the functional and non-functional amyloids in selected features were compared in the non-parametric unpaired Wilcoxon test. We applied the Bonferroni method to correct the p-value due to multiple testing. P-values smaller than 0.05 were regarded as statistically significant. Due to the multidimensionality of data, the data set was studied and visualized in Correspondence Analysis (CA) and Principal Component Analysis (PCA), where the variables were scaled to the unit variance. Moreover, we applied Linear Discriminant Analysis (LDA) to find a linear combination of features that characterizes and separates the studied protein sequences. Before conducting LDA, we removed highly correlated variables with a correlation coefficient $> |0.8|$ and the highest VIF. Moreover, we normalized the variables by applying the best normalizing transformations on the basis of the Pearson P test statistic for normality.

For selected descriptors based on the LDA results, we build a random forest classification model to classify these two types of proteins. We conducted 100 runs of the model splitting randomly the data into a training set and a test set in the ratio of 3:1. For each iteration, we individually searched for the optimal value (with respect to Out-of-Bag error estimate) of mtry, i.e. the number of variables randomly sampled as candidates at each split. We assumed the number of trees to grow for 1000. The model was learned on the training set with 5-fold cross-validation. Finally, the tested set was predicted using the trained model.

Based on the 100 iterations, the mean, the minimum and the maximum of the following measures were calculated, both for the cross-validation and predicting step: precision (PRE), sensitivity (SEN), specificity (SPE), accuracy (ACC), Matthews correlation coefficient (MCC) and AUC (area under the receiver operating characteristic curve).

The parameters are expressed by the formulas:

$$PRE = \frac{TP}{(TP + FP)}$$

$$SEN = \frac{TP}{(TP + FN)}$$

$$SPE = \frac{TN}{(TN + FP)}$$

$$ACC = \frac{TP + TN}{(TP + FP + TN + FN)}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}}$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives and FN is the number of false negatives.

Sensitivity is also called recall or true positive rate (TPR), whereas specificity is also named true negative rate (TNR). AUC corresponds to the area under the receiver operating characteristic curve depicted in a plot of the sensitivity against the false positive rate (FPR), i.e. (1 −specificity) at various threshold settings. AUC is a statistics that is used in the comparison of different models. MCC is generally regarded as a balanced measure that can be used even if the classes are of very different sizes and is used to measure the quality of classifications. MCC takes values from -1 to 1, whereas the other measures from 0 to 1. The higher the values, the better the model distinguishes the analyzed groups.

The analyses were conducted in R software [RStudio Team, 2020] using various packages: stats, bestNormalize [Peterson, 2021], bio3d [Grant et al., 2021], fuzzySim [Barbosa, 2015], FactoMineR [Lê et al., 2008], MASS [Venables and Ripley, 2002], protr [Xiao et al., 2015], randomForest [Cutler and Wiener, 2022] and seqinr [Charif et al., 2022].

7.3 Results

7.3.1 Statistical analyses

Easily interpretable descriptors such as amino acid composition (AAC), dipeptide composition (DC) and AAindex were subjected to statistical testing. The functional and non-functional amyloids differ significantly in the composition of 17 amino acid residues: R, M, K, T, H, S, C, Y, L, Q, P, E, F, W, D, N and V. The functional amyloid sequences are characterized by a higher content of small hydroxylated residues, threonine and serine (Fig. 57). Considering mean values for the whole set, the frequency of T and S was 2 and 2.5 times higher in functional amyloids. In turn, non-functional amyloids show 1.5 to almost 1.8 times increase in basic lysine, arginine and histidine, aromatic tryptophan and tyrosine as well as sulphur-containing methionine and cysteine.

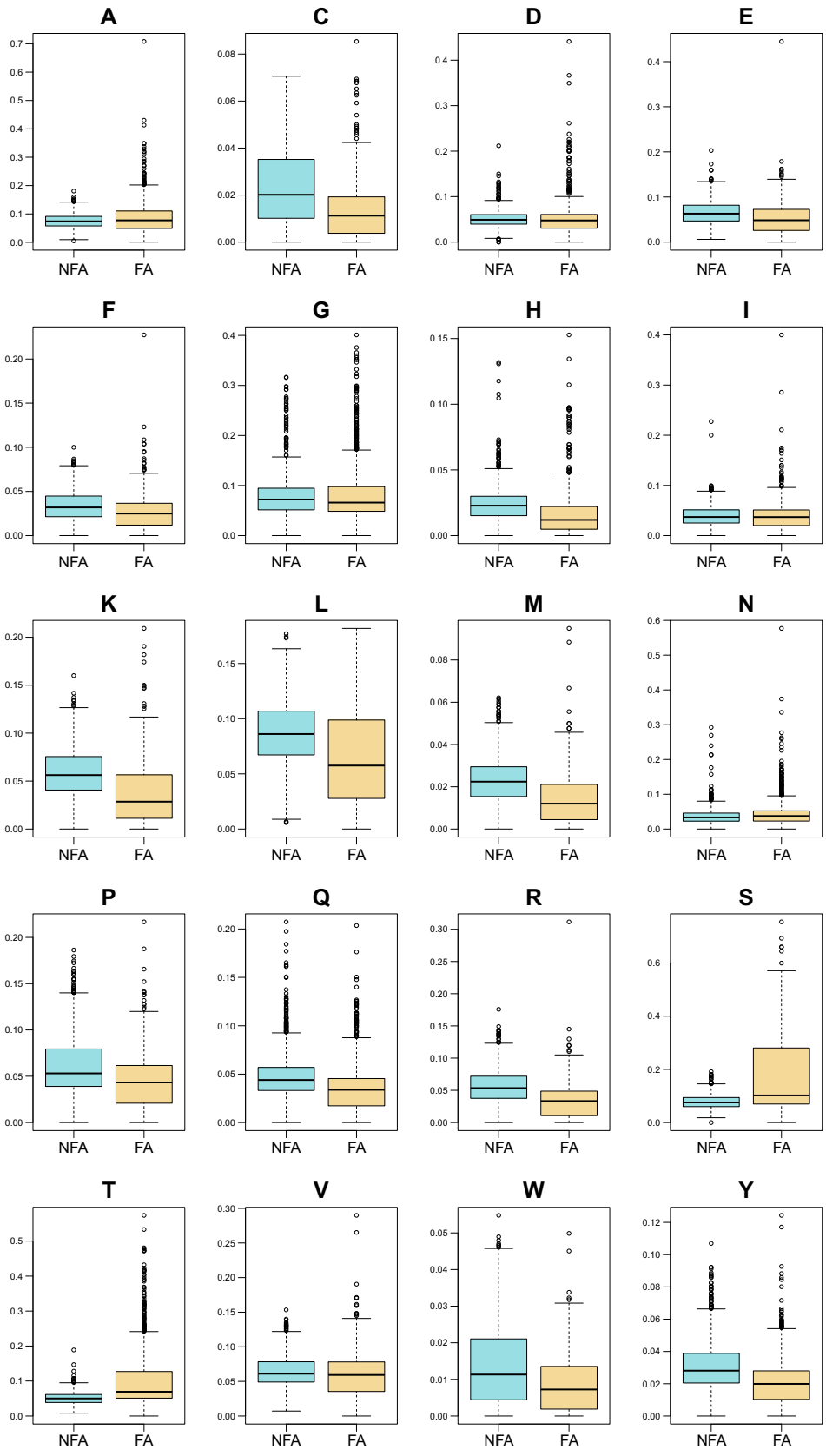

Figure 57: **Box-plots of amino acid composition in functional amyloids (FA) and non-functional amyloids (NFA).** The thick line indicates the median, the box shows the quartile range and the whiskers denote the range without outliers.

CA (Correspondence Analysis) and PCA (Principal Component Analysis) quite well separate the compared amyloid proteins (Fig. 58). In the plots for two principal coordinates or components explaining the largest fraction of variance, we can notice two main sets separated by the first coordinate. A bigger set contains exclusively functional amyloids and the second comprises both types of proteins. There is also a small group isolated by the second principal component. In the case of CA, this group includes only non-functional amyloids but in PCA, the representatives of both amyloid proteins are present. In the results of CA (Tab. 23), the largest weights in the separation revealed serine, alanine, glycine and threonine. In PCA, the most positively correlated variables with the first component are leucine, methionine, phenylalanine, lysine, arginine, tyrosine, tryptophan and cysteine, whereas high negative correlation coefficients show serine and threonine. Considering the second component, the highest positive correlation is demonstrated by glycine and negative by valine.

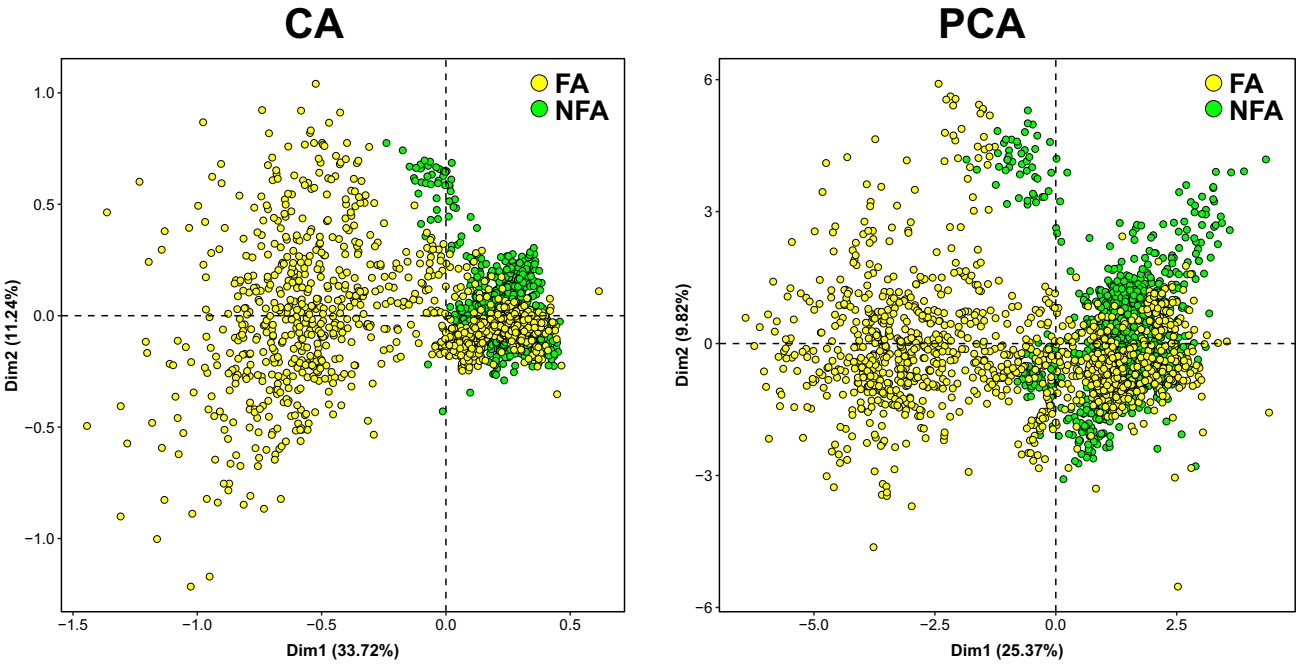

Figure 58: **Correspondence Analysis (CA) and Principal Component Analysis (PCA) of functional amyloids (FA) and non-functional amyloids (NFA) for amino acid composition.**

Table 23: **Weights from Correspondence Analysis and correlation coefficients with two components (CC1 and CC2) from Principal Component Analysis for the amino acid composition of functional amyloids and non-functional amyloids.**

Amino acid	Weights	CC1	CC2
A	0.082	-0.19	-0.47
C	0.018	0.52	0.16
D	0.050	0.13	-0.17
E	0.057	0.43	-0.35
F	0.029	0.63	0.05
G	0.080	-0.33	0.52
H	0.019	0.42	0.34
I	0.039	0.44	-0.28
K	0.046	0.60	-0.29
L	0.073	0.76	-0.24
M	0.018	0.68	0.05
N	0.042	-0.08	0.30
P	0.052	0.22	-0.02
Q	0.041	0.31	0.41
R	0.043	0.59	0.29
S	0.134	-0.83	0.13
T	0.079	-0.63	-0.18
V	0.061	0.36	-0.54
W	0.011	0.52	0.21
Y	0.026	0.53	0.48

In the case of dipeptide content, 215 out of 400 parameters occurred statistically significant between the compared protein sequences. The functional amyloids are characterized on average by a high increase in dipeptides containing serine and threonine but also VW dipeptide, whereas non-functional amyloids by dipeptides, which are rich in cysteine, aspartic acid, methionine, arginine and histidine (Tab. 24).

Table 24: **Selected dipeptides whose mean frequency was at least 2.5 times higher in a given amyloid group and the difference was statistically significant.**

Dipeptide	Ratio of means	Group
WV	3.7	non-functional amyloids
PM	3.5	non-functional amyloids
CA	3.4	non-functional amyloids
RG	3.2	non-functional amyloids
FH	3.1	non-functional amyloids
MD	3.1	non-functional amyloids
FQ	3.0	non-functional amyloids
RC	2.9	non-functional amyloids
WC	2.9	non-functional amyloids
ER	2.9	non-functional amyloids
IC	2.9	non-functional amyloids
CK	2.7	non-functional amyloids
HP	2.7	non-functional amyloids
MC	2.7	non-functional amyloids
DM	2.7	non-functional amyloids
KH	2.7	non-functional amyloids
DY	2.6	non-functional amyloids
YE	2.6	non-functional amyloids
DR	2.5	non-functional amyloids
SS	5.8	functional amyloids
TT	5.8	functional amyloids
ST	5.6	functional amyloids
TS	5.2	functional amyloids
SG	3.1	functional amyloids
SA	3.0	functional amyloids
AS	3.0	functional amyloids
DS	2.9	functional amyloids
GS	2.9	functional amyloids
TG	2.9	functional amyloids
SN	2.7	functional amyloids
VW	2.5	functional amyloids

CA and PCA plots for the dipeptide composition also clearly distinguish many functional

amyloids but some of them are still grouped with non-functional ones according to the first principal component (Fig. 59). In PCA, we can also notice the differentiation of non-functional amyloids according to the second component. In Tab. 25, we selected dipeptides showing the highest weights and/or correlations with two components. Most often, they are combinations of serine, leucine, glycine, alanine, and threonine.

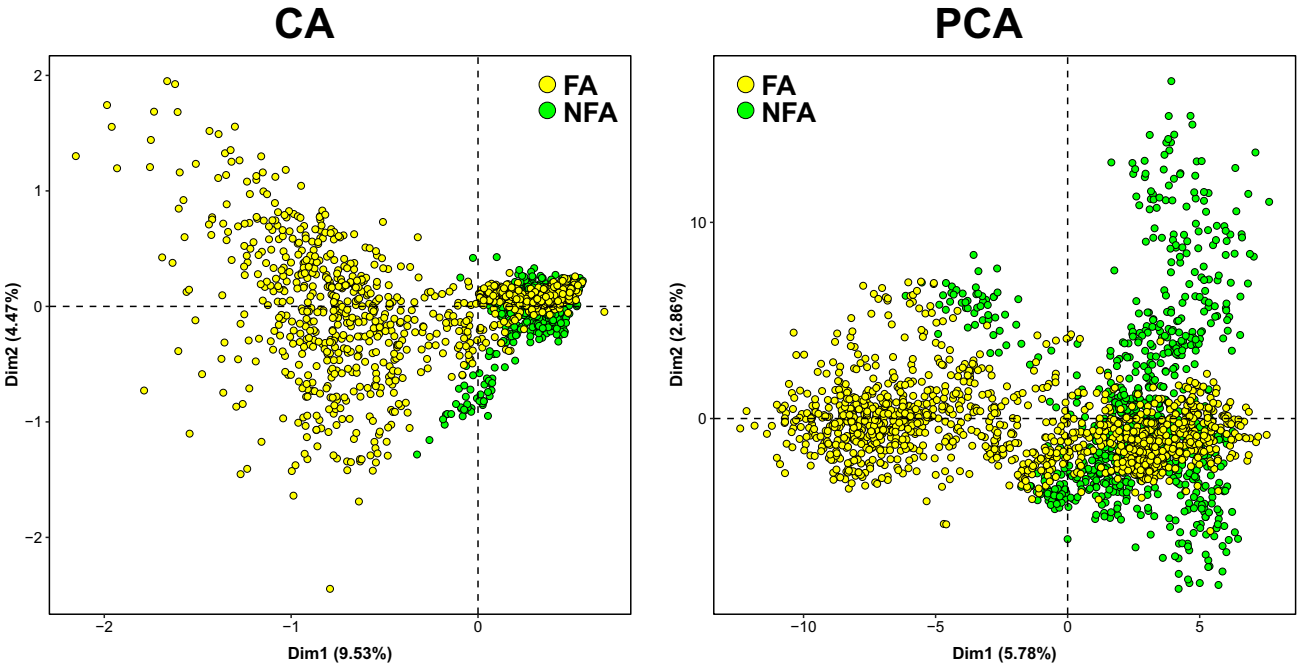

Figure 59: **Correspondence Analysis (CA) and Principal Component Analysis (PCA) of functional amyloids (FA) and non-functional amyloids (NFA) for dipeptide composition.**

Table 25: **Weights from Correspondence Analysis and correlation coefficients with two components (CC1 and CC2) from Principal Component Analysis for selected dipeptides of functional and non-functional amyloids.**

Dipeptide	Weights	CC1	CC2
AA	0.010	-0.17	-0.13
AS	0.010	-0.48	-0.05
AW	0.001	0.24	0.51
CN	0.001	0.19	0.55
EL	0.004	0.53	-0.15
FL	0.003	0.45	0.06
GC	0.001	0.23	0.54
GG	0.009	-0.24	0.18
GS	0.014	-0.53	0.15
LE	0.004	0.49	-0.26
LK	0.004	0.46	-0.20
LL	0.008	0.54	0.05
LS	0.007	0.00	-0.05
SA	0.010	-0.46	-0.04
SG	0.013	-0.48	0.17
SS	0.033	-0.57	0.05
ST	0.016	-0.60	-0.03
TG	0.006	-0.41	-0.01
TS	0.013	-0.56	-0.06
TT	0.011	-0.44	-0.07

AAindex measures also occurred to differentiate the studied amyloids (Tab. 26). The most distinguishing indices cover various physicochemical properties including secondary structure. Among them, there are weights for β-sheet and coil structures, optimized relative partition energies, information measure for loop and turn, as well as scales for hydrophobicity, polarity and net charge. Functional amyloids showed larger mean values for optimized relative partition energies as well as indices for β-sheet, turn and loop, whereas non-functional amyloids for hydrophobicity, net charge and weights for coil region. Due to reverse scaling, high values in RADA880108 index mean that a given amino acid is in fact hydrophobic.

Table 26: **Selected amino acid indices that showed the largest percentage difference between mean values calculated for functional amyloids (FA) and non-functional amyloids (NFA) and significantly differentiate these groups.**

AAIndex	**Description**	**Mean for NFA**	**Mean for FA**
QIAN880116	Weights for β-sheet at the window position of -4	-0.0575	0.0005
BAEK050101	Linker index	-0.0003	-0.0244
ROBB760108	Information measure for turn	-0.3310	0.0063
KLEP840101	Net charge	-0.0007	-0.0308
QIAN880125	Weights for β-sheet at the window position of 5	-0.0019	0.0779
MIYS990104	Optimized relative partition energies - method C	-0.0208	0.0007
QIAN880114	Weights for β-sheet at the window position of -6	-0.0024	0.0720
ROBB760109	Information measure for N-terminal turn	-0.0072	0.1283
CIDH920103	Normalized hydrophobicity scales for α+β-proteins	-0.0082	-0.1598
MIYS990105	Optimized relative partition energies - method D	-0.0178	0.0012
QIAN880139	Weights for coil at the window position of 6	0.0572	0.0040
MIYS990103	Optimized relative partition energies - method B	-0.0186	-0.0016
NAKH900106	Normalized composition from animal	-0.0057	0.0550
COWR900101	Hydrophobicity index, 3.0 pH	0.0771	0.0073
QIAN880126	Weights for β-sheet at the window position of 6	0.0135	0.1027
RADA880108	Mean polarity	0.0137	-0.0696
ROBB760113	Information measure for loop	-0.3106	0.0723
CIDH920101	Normalized hydrophobicity scales for α-proteins	-0.0385	-0.2386
QIAN880138	Weights for coil at the window position of 5	0.0419	-0.0107
QIAN880137	Weights for coil at the window position of 4	-0.0072	-0.0416

PCA performed for amino acid indices also separated two groups, although the distance between them is smaller than in the case of amino acid and dipeptide compositions (Fig. 60). Most functional amyloids are in one group, but others overlap non-functional ones. In contrast to the plots for the compositions, the two groups have comparable ranges in the plot. Amino acid indices the most correlated with the principal components are associated with various features related for example to secondary structures, especially for β-sheets, hydrophobicity, buriability and protein stability (Tab. 27).

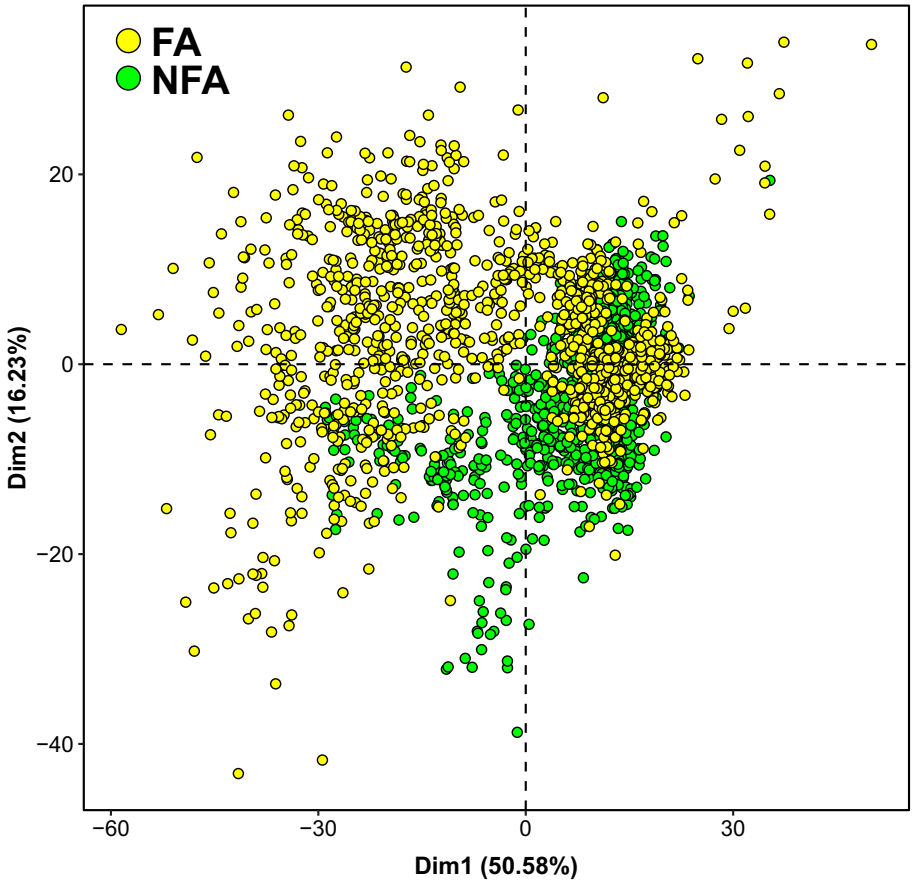

Figure 60: **Principal Component Analysis of functional amyloids (FA) and non-functional amyloids (NFA) for selected amino acid indices.**

Table 27: **Correlation coefficients with two components (CC1 and CC2) from Principal Component Analysis for selected amino acid indices calculated for functional and non-functional amyloids.**

AAIndex	CC1	CC2	Description
BASU050101	0.98	0.16	Interactivity scale obtained from the contact matrix
BASU050102	0.98	0.07	Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins
CHOP780203	-0.97	-0.05	Normalized frequency of β-turn
CHOP780211	-0.97	0.12	Normalized frequency of C-terminal non β region
CIDH920102	0.97	-0.02	Normalized hydrophobicity scales for β-proteins
CIDH920105	0.97	0.06	Normalized average hydrophobicity scales
DESM900102	0.12	0.96	Average membrane preference: AMP07
GEIM800111	-0.98	0.04	Aperiodic indices for α/β-proteins
GUOD860101	0.97	0.12	Retention coefficient at pH 2
MEIH800101	-0.97	-0.14	Average reduced distance for C-α
MIYS990101	-0.97	-0.16	Relative partition energies derived by the Bethe approximation
OOBM770101	-0.09	-0.95	Average non-bonded energy per atom
PARJ860101	-0.97	-0.11	HPLC parameter
PLIV810101	0.97	0.08	Partition coefficient
RACS770101	-0.98	-0.01	Average reduced distance for C-α
SUEM840101	0.97	-0.13	Zimm-Bragg parameter s at 20 C
TAKK010101	0.97	-0.04	Side-chain contribution to protein stability (kJ/mol)
VINM940102	-0.97	-0.04	Normalized flexibility parameters (B-values) for each residue surrounded by none rigid neighbours
ZHOH040101	0.97	-0.05	The stability scale from the knowledge-based atom-atom potential
ZHOH040103	0.97	0.15	Buriability

7.3.2 Discriminant analysis

The application of LDA (Linear Discriminant Analysis) showed a variable accuracy of classification depending on the used descriptors (Tab. 28). SOCN was excluded from the study because all variables occurred too highly correlated. The accuracy was the smallest for MoreauBroto (0.69) and basic composition descriptors, i.e. CTDC, CTDT and AAC (0.70 to 0.80).

The best discriminator appeared dipeptide composition (DC) with an accuracy of almost 0.99. At the top, there were also CTriad, AAindex and APAAC (0.96 – 0.90).

Table 28: **Accuracy obtained in Linear Discriminant for various protein sequence descriptors for functional and non-functional amyloids. The numbers of initial variables and those after the exclusion of the most correlated were also presented.**

Descriptor	Initial number of parameters	Number of parameters without correlated	Accuracy
DC	400	396	0.986
CTriad	343	342	0.958
AAindex	544	74	0.898
APAAC	50	49	0.890
QSO	70	48	0.872
Moran	105	75	0.850
Geary	105	75	0.850
PAAC	35	21	0.846
CTDD	105	69	0.844
AAC	20	20	0.805
CTDT	21	14	0.735
CTDC	21	8	0.705
MoreauBroto	105	6	0.693

AAC - amino acid composition, APAAC - Amphiphilic Pseudo-Amino Acid Composition, CTDC - Composition, CTDD - Distribution, CTDT - Transition, CTriad - Conjoint Triad, DC - dipeptide composition, Geary - Geary Autocorrelation, Moran - Moran Autocorrelation, MoreauBroto - Normalized Moreau-Broto Autocorrelation, PAAC - Pseudo-Amino Acid Composition, QSO - Quasi-Sequence-Order Descriptors

LDA identified only one discriminant function for each of the descriptors. Histograms for this function and selected descriptors are presented in Fig. 61. In terms of amino acid composition, shown for comparison, the functional and non-functional amyloids are rather poorly separated and characterized by a large overlap in the distribution of function values. Much better separation is visible for amino acid indices, whereas the best one for CTriad and dipeptide composition. In the case of dipeptide composition, the most positively correlated (0.17 to 0.12) with the discriminant function occurred dipeptides ST, TT, AS, SA SS, AT, NG, TG, and NT, whereas the most negatively correlated (-0.16 to -0.12) RG, FQ, ER, AK, DY, CA, KV, LK, and MD. The correlation coefficients of amino acid indices with the discriminant function are

larger in terms of absolute value (Tab. 29). Among these indices, there are those related to the free energy of protein conformation as well as secondary structures: β-sheet, coil regions, and helix.

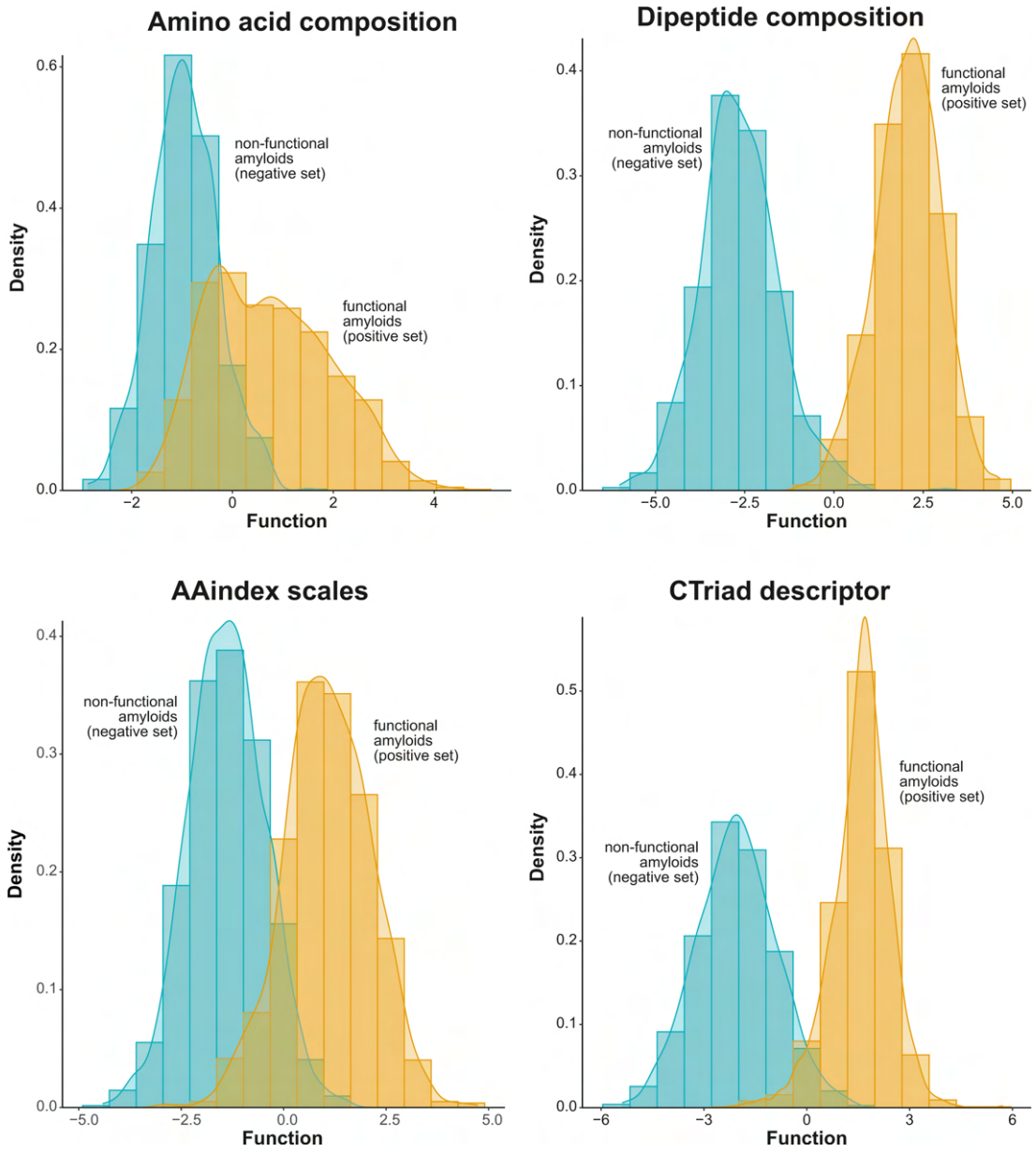

Figure 61: **Histograms for discriminant function and selected descriptors calculated for functional and non-functional amyloids.**

Table 29: **Correlation coefficient (CC) with discriminant function for selected amino acid indices calculated for functional and non-functional amyloids.**

AAIndex	CC	Description
WERD780102	-0.470	Free energy change of epsilon(i) to epsilon(ex)
CHAM830102	-0.441	A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of β-sheet
NAKH900103	-0.412	AA composition of mt-proteins
WERD780103	-0.386	Free energy change of α(Ri) to α(Rh)
TANS770108	-0.306	Normalized frequency of zeta R
LAWE840101	0.302	Transfer free energy, CHP/water
ROSM880103	0.308	Loss of Side chain hydropathy by helix formation
QIAN880137	0.382	Weights for coil at the window position of 4
RICJ880116	0.421	Relative preference value at C'
CHAM830108	0.452	A parameter of charge transfer donor capability

To summarize, the results of the statistical and discriminant analyses indicate that functional and non-functional amyloids show many sequence features that can differentiate them. However, CA and PCA plots demonstrated that the functional amyloids are clearly a heterogeneous group, which can be separated into two sets. The non-functional amyloids are not fully homogenous, either, and some smaller subgroups can be recognized. Nevertheless, there are distinctive features that discriminate these sets. Taking together of the results, we can say that functional amyloids are characterized by a high content of amino acids S and T, as well as dipeptides ST, TS, TT, SS, AS, SA, SG, TG, AT, GS, SN, GT and DS. In turn, non-functional amyloids are rich in R, K, M, L, C, H and Y as well as LK, AK, KV, LL, RV, FQ, ER, KE, EL, EK, VR, LR and DL. Larger values of amino acid indices for functional amyloids are associated mainly with loop, turn, flexibility, β-sheet, β-turn, β region, optimized relative partition energies, whereas those for non-functional amyloids with buriability, hydrophobicity, net charge, protein stability and free energy change.

7.3.3 Prediction analyses using random forest

Based on the LDA results, we selected the most promising descriptions to verify their usefulness in a random forest model. To this study, we selected three top descriptors, i.e. DC, CTriad and AAindex as well as APAAC, which was characterized by the smallest number of parameters.

To assess how accurately the predictive model can perform in practice, we conducted five-fold cross-validation by dividing functional amyloid and non-amyloid training data sets into five groups, four training and one tested. This assessment is presented in Tab. 30. Generally, the values of all the applied measures were higher than 0.8 and mostly 0.9. Dipeptide composition occurred as the best predicting feature in terms of precision (mean 0.983), specificity (mean 0.979) and AUC (mean 0.996), whereas AAIndex for sensitivity (mean 0.961), accuracy (mean 0.962) and MCC (mean 0.922). CTriad and APAAC descriptions appeared weaker.

Table 30: **Cross-validation results for random forest model based on selected descriptors in the prediction of functional and non-functional amyloids.** The highest value for a given measure was bolded. The mean as well as the minimum and the maximum values in parentheses calculated for 100 runs were also presented.

Measure	**DC**	**AAIndex**	**CTriad**	**APAAC**
Precision	**0.983 [0.976-0.988]**	0.970 [0.962-0.978]	0.979 [0.972-0.987]	0.968 [0.957-0.977]
Sensitivity	0.945 [0.926-0.963]	**0.961 [0.955-0.969]**	0.887 [0.862-0.907]	0.938 [0.930-0.947]
Specificity	**0.979 [0.970-0.986]**	0.962 [0.950-0.972]	0.975 [0.965-0.985]	0.960 [0.947-0.972]
Accuracy	0.960 [0.949-0.970]	**0.962 [0.956-0.968]**	0.926 [0.909-0.934]	0.948 [0.939-0.953]
AUC	**0.996 [0.994-0.997]**	0.992 [0.991-0.994]	0.988 [0.986-0.991]	0.990 [0.987-0.991]
MCC	0.920 [0.898-0.939]	**0.922 [0.911-0.936]**	0.856 [0.825-0.870]	0.894 [0.878-0.905]

APAAC - Amphiphilic Pseudo-Amino Acid Composition, CTriad - Conjoint Triad, DC - dipeptide composition

Next, we tested the model based on an independent set. The predictions turned out very good (Tab. 31). The measures varied from 0.796 to 0.996. Dipeptide composition and amino acid indices outperformed other descriptors. DC achieved the highest mean sensitivity (0.980) and AUC (0.962), whereas AAIndex precision (0.950), specificity (0.961), accuracy (0.961) and MCC (0.922).

Table 31: **Performance of random forest model based on selected descriptors in the prediction of functional and non-functional amyloids.** The highest value for a given measure was bolded. The mean as well as the minimum and the maximum values in parentheses calculated for 100 runs were also presented.

Measure	**DC**	**AAIndex**	**CTriad**	**APAAC**
Precision	0.932 [0.874-0.980]	**0.950 [0.914-0.976]**	0.871 [0.808-0.938]	0.922 [0.863-0.959]
Sensitivity	**0.980 [0.945-0.996]**	0.962 [0.923-0.995]	0.974 [0.945-0.996]	0.962 [0.907-0.991]
Specificity	0.944 [0.893-0.984]	**0.961 [0.930-0.980]**	0.889 [0.848-0.952]	0.937 [0.892-0.967]
Accuracy	0.960 [0.933-0.987]	**0.961 [0.935-0.980]**	0.926 [0.894-0.963]	0.948 [0.920-0.968]
AUC	**0.962 [0.936-0.987]**	0.961 [0.933-0.981]	0.931 [0.906-0.965]	0.950 [0.920-0.969]
MCC	0.920 [0.868-0.974]	**0.922 [0.866-0.959]**	0.856 [0.796-0.925]	0.896 [0.839-0.937]

APAAC - Amphiphilic Pseudo-Amino Acid Composition, CTriad - Conjoint Triad, DC - dipeptide composition

In Tab. 32 and Tab. 33, we selected the top twenty dipeptides that showed the largest importance for random forest classification. Their importance was assessed by two measures: Mean Decrease Accuracy, which describes how much the model accuracy decreases if we drop a given variable, and Mean Decrease Gini, which is based on the Gini impurity index used for the calculation of splits in trees and describes the inconsistency of classification by nodes in trees. Twelve dipeptides were chosen at the same time by these two criteria: AK, CA, DI, ER, FQ, HP, NG, PG, RG, SS, ST and TS. The dipeptides selected by these two criteria are a combination of mainly serine, threonine, glycine, arginine, alanine and isoleucine.

Table 32: **Top twenty dipeptides that showed the biggest Mean Decrease in Accuracy in random forest prediction of functional and non-functional amyloids.** The mean as well as the minimum and the maximum of this measure calculated for 100 runs were presented.

Dipeptide	**Mean [Min-Max]**
SS	20.98 [13.72-35.02]
RG	18.12 [14.18-24.77]
ST	15.82 [11.42-22.1]
NG	15.23 [12.01-22.57]
CA	15.05 [11.16-19.71]
FQ	14.92 [11.85-19.48]
ER	13.83 [10.85-17.43]
TS	13.7 [11.06-16.66]
PG	13.68 [10.85-15.93]
DI	13.54 [10.11-20.26]
GS	12.4 [9.34-15.94]
IC	11.73 [9.03-16.98]
TT	11.57 [9.34-14.8]
SG	11.52 [8.97-14.25]
AK	11.49 [8.29-14.26]
SA	11.39 [9.37-14.66]
HP	10.96 [8.01-13.59]
IE	10.88 [8.41-15.29]
MD	10.87 [8.67-13.24]
AS	10.71 [8.87-13.36]

Table 33: **Top twenty dipeptides that showed the biggest Mean Decrease in Gini in random forest prediction of functional and non-functional amyloids.** The mean as well as the minimum and the maximum of this measure calculated for 100 runs were presented.

Dipeptide	**Mean [Min-Max]**
SS	26.51 [12.2-56.74]
RG	21.14 [10.89-38.32]
NG	17.92 [9.69-38.28]
ST	17.29 [9.32-31.78]
FQ	14.5 [8.58-24]
ER	12.09 [6.66-18.71]
CA	11.38 [6.44-19.9]
TS	10.56 [6.89-16.07]
DI	9.47 [5.02-18.45]
PG	9.4 [5.73-15.09]
AK	9.12 [3.99-16.44]
SI	7.75 [3.41-12.6]
RR	6.89 [3.51-18.43]
RV	6.32 [2.11-12.2]
NT	6.32 [2.61-11.76]
KV	6.04 [2.34-10.11]
VW	6.02 [2.97-12.57]
TY	5.84 [3.07-10.12]
DT	5.55 [2.06-10.47]
HP	5.52 [2.58-9.22]

Likewise, we gather the top twenty amino acid indices that turned out the most important in the prediction (Tab. 34 and 35). Twelve indices were also agreeably chosen by these two criteria: AURR980118, CHAM830104, CHAM830105, DAYM780201, EISD860102, FAUJ880111, GRAR740101, LIFS790102, MITS020101,l NAKH900103, PONP800104 and QIAN880122. The indices selected by these two criteria refer mostly to parameters of the side chain, hydrophobicity, amphiphilicity, and charge as well as various protein conformations and structures, i.e. α-helix, β-strand, β-sheet and coil.

Table 34: **Top twenty amino acid indices that showed the biggest Mean Decrease Accuracy in random forest prediction of functional and non-functional amyloids.** The mean as well as the minimum and the maximum of this measure calculated for 100 runs were presented.

AAIndex	**Description**	**Mean [Min-Max]**
AURR980118	Normalized positional residue frequency at helix termini C"	17.83 [12.92-27.08]
GRAR740101	Composition	17.81 [13.77-26.92]
PONP800104	Surrounding hydrophobicity in α-helix	17.74 [13.96-25.57]
CHAM830105	The number of atoms in the side chain labelled 3+1	16.6 [13.54-21.53]
MITS020101	Amphiphilicity index	15.39 [12.19-18.87]
FAUJ880110	Number of full nonbonding orbitals	15.29 [12.56-20.63]
EISD860102	Atom-based hydrophobic moment	14.95 [13.48-17.47]
DAYM780201	Relative mutability	14.79 [11.71-18.37]
NAKH900103	AA composition of mt-proteins	14.78 [12.04-17.46]
FAUJ880111	Positive charge	14.6 [12.47-16.86]
SUYM030101	Linker propensity index	14.45 [12.47-16.67]
FAUJ880105	STERIMOL minimum width of the side chain	14.29 [12.68-16.83]
QIAN880122	Weights for β-sheet at the window position of 2	14.12 [12.34-16.23]
TANS770108	Normalized frequency of zeta R	13.94 [11.8-17.25]
LIFS790102	Conformational preference for parallel β-strands	13.92 [11.62-16.28]
FASG760102	Melting point	13.86 [11.36-17.87]
AURR980106	Normalized positional residue frequency at helix termini N1	13.76 [11.78-15.55]
CRAJ730102	Normalized frequency of β-sheet	13.73 [11.02-17.72]
CHAM830104	The number of atoms in the side chain labelled 2+1	13.69 [12.13-15.76]
ISOY800106	Normalized relative frequency of helix end	13.54 [11.45-16.1]

Table 35: **Top twenty amino acid indices that showed the biggest Mean Decrease in Gini in random forest prediction of functional and non-functional amyloids.** The mean as well as the minimum and the maximum of this measure calculated for 100 runs were presented.

AAIndex	**Description**	**Mean [Min-Max]**
AURR980118	Normalized positional residue frequency at helix termini C"	27.08 [11.74-71.18]
CHAM830105	The number of atoms in the side chain labelled 3+1	22.45 [13.48-46.25]
MITS020101	Amphiphilicity index	20.12 [10.78-41.48]
DAYM780201	Relative mutability	18.98 [10.2-36.24]
FAUJ880111	Positive charge	18.53 [10.2-29.66]
EISD860102	Atom-based hydrophobic moment	14.5 [9.33-21.13]
QIAN880129	Weights for coil at the window position of -4	12.77 [8.13-18.02]
GRAR740101	Composition	12.19 [8.68-22.5]
QIAN880123	Weights for β-sheet at the window position of 3	11.87 [7.65-17.9]
CHAM830104	The number of atoms in the side chain labelled 2+1	11.6 [6.54-15.18]
SNEP660103	Principal component III	10.91 [7.2-14.86]
PONP800104	Surrounding hydrophobicity in α-helix	9.94 [6.7-16.51]
AURR980120	Normalized positional residue frequency at helix termini C4'	9.81 [5.99-13.74]
NAKH900103	AA composition of mt-proteins	9.5 [6.38-13.23]
LIFS790102	Conformational preference for parallel β-strands	8.81 [6.03-13.9]
SNEP660104	Principal component IV	8.75 [4.65-12.27]
FAUJ880104	STERIMOL length of the side chain	8.64 [4.72-11.64]
QIAN880122	Weights for β-sheet at the window position of 2	8.34 [6.12-11.71]
OOBM770102	Short and medium range non-bonded energy per atom	8.22 [3.66-11.24]
QIAN880139	Weights for coil at the window position of 6	7.78 [5.16-11.48]

8 AmyloGraph - amyloid interaction database

8.1 Research objectives

Due to the lack of databases containing standardized descriptions of amyloid interactions, including proteins under this study, i.e., CsgA and CsgB, we decided to create a database to fill this gap. In our work on AmyloGraph [Burdukiewicz et al., 2022], we designed detailed definitions of amyloid interactions (Fig. 62). Six of them deal with amyloid-amyloid interactions and assume that there are only two participants in each interaction and that the interactor modulates the self-assembly of the interactee. We have also developed three descriptors to more rigorously describe the details of the scenarios, which provide specific determinations of possible states. We also would like to emphasize that the designed descriptors do not replace existing terminology, but rather standardize it.

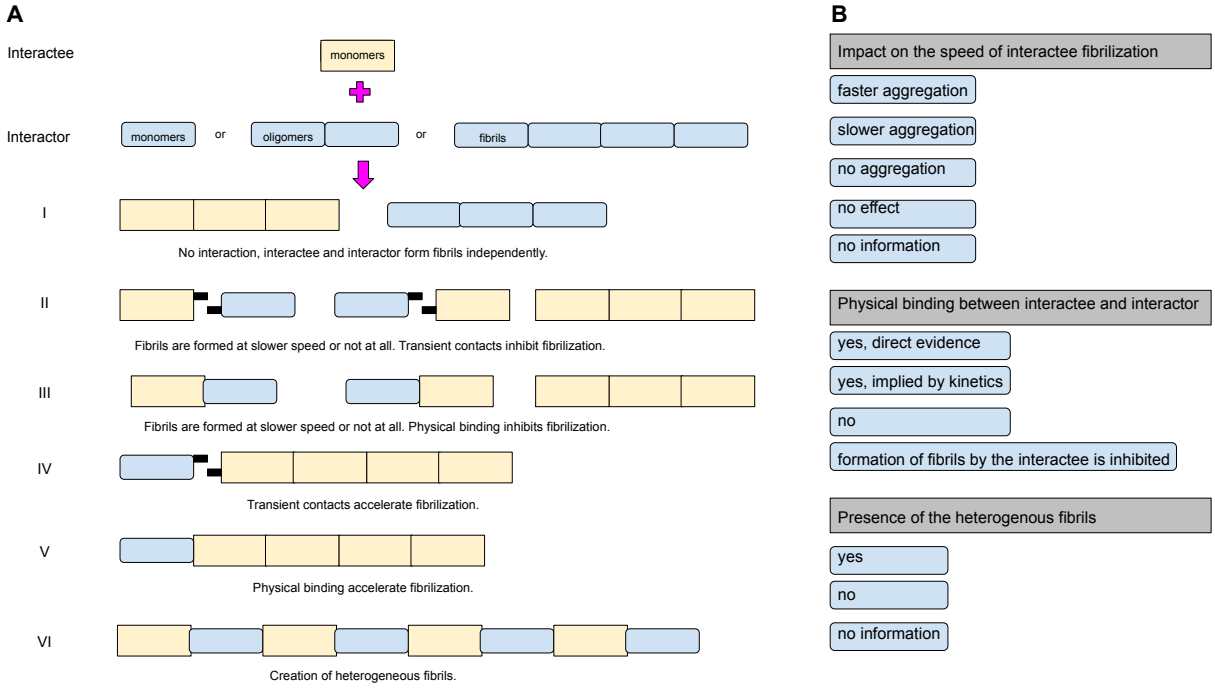

Figure 62: **Definitions of amyloid interactions developed for the AmyloGraph.** A) Six scenarios of amyloid-amyloid interactions. Colors represent different amyloid molecules taking a part in the interactions. Roman numerals denote different interaction scenarios. B) Three descriptors of AmyloGraph. Grey rectangles represent descriptors, blue rectangles with round edges represent the levels of the descriptor above.

8.2 Materials and Methods

8.2.1 Systematization of terminology on interactions between amyloids

We created a precisely controlled vocabulary to describe the amyloid-amyloid interactions. We assumed that there are only two participants in each interaction, interaction and interactee. Next, we developed three categories of descriptors to better describe the interaction. They were based on the fibrilization speed, the presence of physical binding between both interacting proteins, and the appearance of heterogeneous fibrils (Fig. 62B).

Descriptors for fibrilization speed:

- 1. Faster aggregation:
  - a) the maximum ThT emission of the reaction of the interactee and interactor is higher than the emission for the interactee alone.
  - b) the slope of the kinetic curve is steeper.
  - c) the lag phase is shorter.
  - d) the time required for the amyloid reaction to reach 50% of the final fluorescence intensity is lower.
- 2. Slower aggregation:
  - a) the maximum ThT emission observed at the end of the reaction of the interactee and interactor is lower than the emission for the interactee alone.
  - b) the slope of the kinetic curve is less steep.
  - c) the lag phase is longer.
- 3. No aggregation:
  - a) there is no confirmed fibrilization after the interaction.
- 4. No effect:
  - a) the slopes of kinetic curves are similar.
  - b) the maximum ThT emission is similar.

  - c) the lag phase is similar.

- 5. No information:

  - a) there were no kinetic assays.

Descriptors for physical binding between interactee and interactor. Are there physical bonds between interactee and interactor?:

- 1. Yes, direct evidence:

  - a) there is experimental evidence that fibrils consist of two different amyloids.
  - b) there is a visible colocalization of an interactee and an interactor in microscopic images.

- 2. Yes, implied by kinetics:

  - a) seeding is implied by kinetic experiments results and is interpreted as such by the authors of the publication.

- 3. No:

  - a) there is no effect on the elongation of interactees fibrils.

- 4. Formation of fibrils by the interactee is inhibited:

  - a) the formation of interactees aggregates was slowed or halted by the interactor.

- 5. No information:

  - a) there is no experimental evidence, the seeding is not implied by kinetic experiments results.

Descriptors for indicating the presence of the heterogenous fibrils, which consist of interactor and interactee molecules. Are heterogenous fibrils formed, composed of interactor and interactee?:

- 1. Yes:

  - a) experimental evidence that fibrils consist of two different amyloids
  - b) the mature fibrils are structurally different from fibrils formed in the presence of an interactor
  - c) the term coaggregation, heterogeneous or hybrid fibrils is used to describe the aggregation process.

- 2. No, amyloid fibrils have the same dimension, which matches interactee alone:
  - a) the same structure of interactee and interactor fibrils confirmed by a microscopy technique.
  - b) there are no fibrils at all.
  - c) interactee and interactor are the same protein.

- 3. No information:
  - a) no experimental evidence, seeding not implied by kinetic experiments.

8.2.2 Datasets preparation and curation

In order to create the interaction database, we needed to collect the data and curate it. We have designed a three-stage pipeline, which includes the pre-screen of manuscripts, manual curation, and independent final validation. The first two steps have been supported through the design of the applicable forms, which standardized annotations. We managed to expand the collection of many publications, and through labor-intensive efforts. In total, we analyzed 562 manuscripts. However, only 364 were potentially suitable for the database.

Having the collected manuscripts and data, we, along with other curators, attempted to manually curate the database. We extracted information on amyloid interactions from each paper, without reinterpreting the data and conclusions provided by their authors, except when the authors did not describe the results or the description was limited. During the initial data curation, we focused on annotating the manuscripts using the descriptors and collecting protein sequences of interacting amyloids. The next step was the validation of our annotations by other curators, who were not involved in the validation of a specific record during the initial

curation. The final collections of manuscripts consisted of 172 publications and 883 amyloid-amyloid interactions. Then, we tried to contact the authors of the publications to validate our database entries, however, only 11 of them responded, confirming 81 interactions.

8.2.3 R package, shiny web server, and the database

AmyloGraph is available as an online web server. The front end of it is built using the Shiny package [Chang et al., 2022]. However, due to the fact that the application relies on external servers, which reduces their persistence [Veretnik et al., 2008], we have developed AmyloGraph as an R package [Team et al., 2021]. The package itself needs only core knowledge of R to be able to run it on a local PC.

8.3 Results

The results of the study were published in Burdukiewicz et al. [2022], the amyloids interaction database is available at https://amylograph.com/. One of the main results of the amyloid interaction database was the identification of the 48 interactions of CsgA and 14 interactions of CsgB with other amyloid proteins (Fig. 63).

	interactor_name	interactee_name	aggregation_speed	elongates_by_attaching	heterogenous_fibers	doi
1	CsgA	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
2	CsgA	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
3	CsgA	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
4	CsgA	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
5	CsgA	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
6	CsgA	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
7	CsgA	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
8	CsgB	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
9	CsgB	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
10	CsgB	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
11	CsgB	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
12	CsgA	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
13	CsgA	CsgA	No effect	No	No information	10.1074/jbc.M112.383737
14	CsgA	CsgA	No effect	No	No information	10.1074/jbc.M112.383737
15	CsgA	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
16	CsgA	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
17	CsgB	CsgA	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
18	Amyloid beta	CsgA	No effect	No	No information	10.1074/jbc.M112.383737
19	Sup35	CsgA	Slower aggregation	Formation of fibrils by the interactee is inhibited	No information	10.1074/jbc.M112.383737
20	CsgB	Sup35	No effect	No	No information	10.1074/jbc.M112.383737
21	CsgA	Amyloid beta	No effect	No	No information	10.1074/jbc.M112.383737
22	CsgA	Sup35	Faster aggregation	Yes; implied by kinetics.	No information	10.1074/jbc.M112.383737
23	CsgA	CsgA	No effect	No	No information	10.1074/jbc.M112.383737
24	CsgA	Alpha-synuclein	Faster aggregation	Yes; implied by kinetics.	No	10.7554/eLife.53111
25	CsgA	Tau	No effect	No	No	10.7554/eLife.53111

Showing 1 to 56 of 56 entries Previous 1 Next

Figure 63: **Known interactions of CsgA and CsgB peptides in AmyloGraph.**

On a broader level, AmyloGraph contains 883 interactions between 46 proteins reported in 172 manuscripts. Moreover, we integrated into the database a user-friendly graph (Fig. 64A), where nodes represent individual amyloids and their edges illustrate interactions between them.

In addition to the graphical representation of interactions, we also included a table with interactees and interactors as well as links to UniProt records of these proteins (Fig. 64B). The AmyloGraph table is dynamic, searchable, and the user can download selected rows. Downloaded data contain all available information, including the sequences of amyloid proteins participating in the interactions.

Table as well as the graph representations of data can be filtered out using various criteria. The filters cover all three descriptors. Moreover, the edges on the graph can be colored, according to the levels of a chosen descriptor. Amino acid sequences can be used to filter the information presented both in the graph and in the table. We have implemented a set of regular expressions inspired by the POSIX system to facilitate more advanced searches of sequence motifs that should appear in either interactor's or interactee's sequence.

By analyzing the interactions that occur between CsgA and CsgB and other amyloids, we can determine that CsgA and CsgB react with essentially the same five proteins. The exceptions are three proteins that react only with CsgA, namely Tau, α-synuclein and the lysosome (Fig. 65). It is also worth noting that these other amyloid proteins also interact with each other.

Based on the database, we found also information about the character of these interactions. CsgA can normally accelerate their own aggregation, but there are reports that also cannot. The ones that do not impact on aggregation usually have modified sequences. They differ by 38 amino acids along the sequence, when compared to CsgA K12 model. It speeds up the aggregation of IAPP, SEVI, and Sup35 but does not have any effect on the aggregation of Tau. CsgA can have no effect or accelerate the aggregation of α-synuclein, depending on the sequence variation of CsgA. Its impact on amyloid β can be also different from no effect to acceleration. CsgA proteins speed up the aggregation of amyloid β, even if they lack a signal peptide. However, amyloid β, which is only two amino acids longer than the interacting one, does not react to CsgA. Transthyretin either has no effect, slows down, or inhibits the aggregation of CsgA, depending on the variation of the sequences. There are 13 records of transthyretin impacting the aggregation of CsgA with different protein sequences.

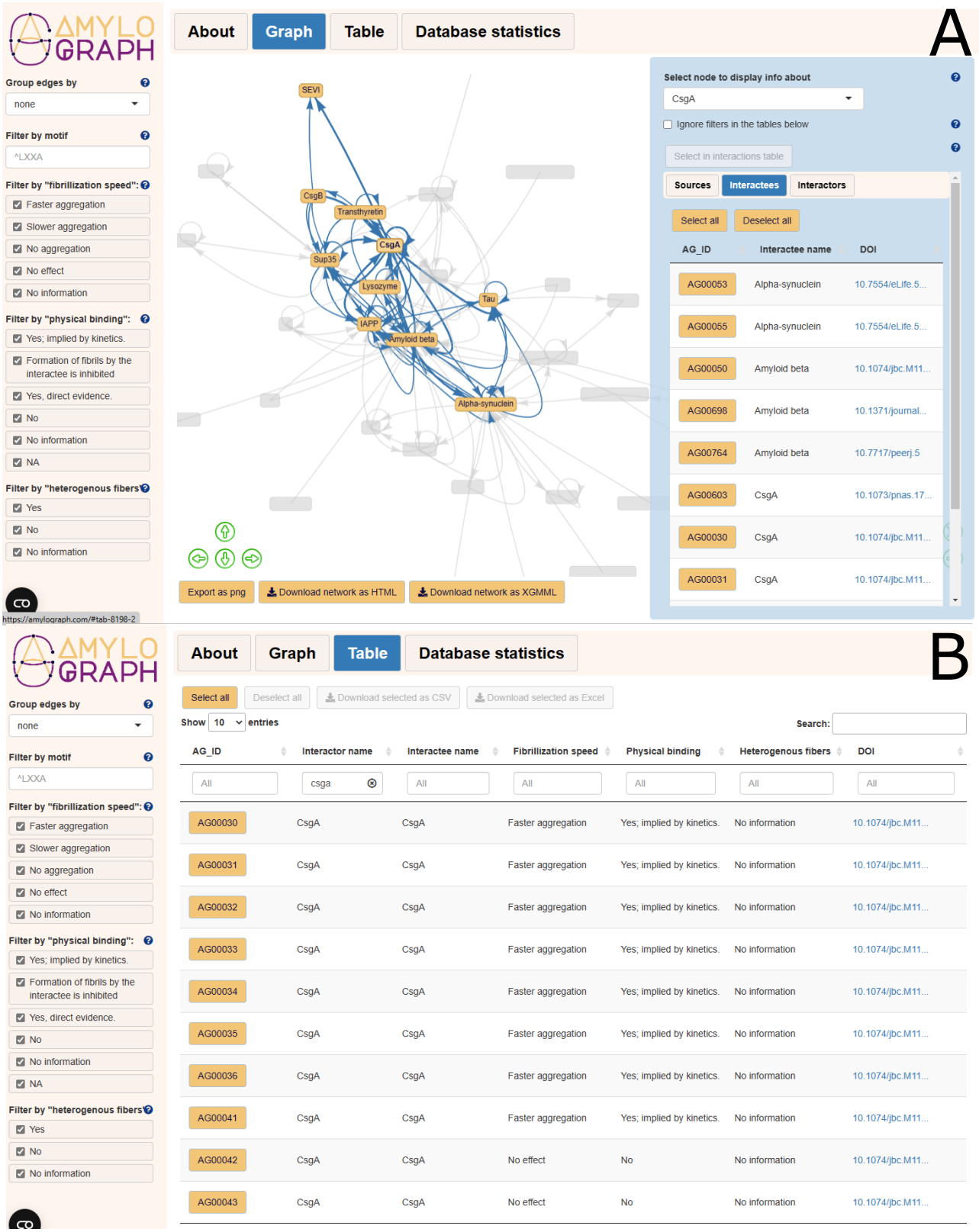

Figure 64: **Record for CsgA in the AmyloGraph database.** A) Graph view of amyloid-amyloid interactions. The interactions (edges) are colored according to the levels of descriptor 2, "physical binding". The right-hand panel represents an overview of the interactions. B) Tabular view of interactions. The top section of this card contains download options allowing to obtain data.

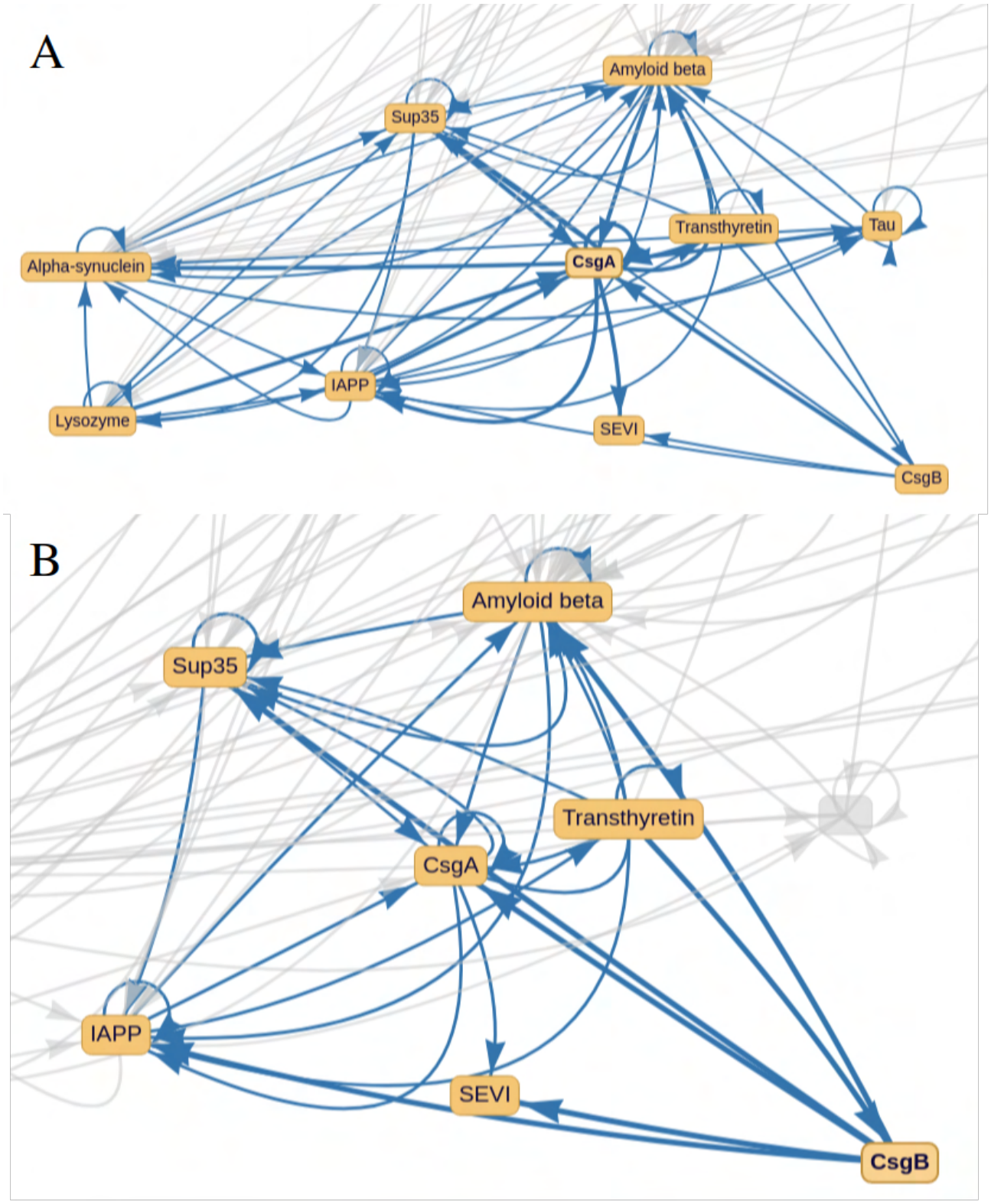

Figure 65: **CsgA and CsgB graph interactions from AmyloGraph.** A) CsgA, B) CsgB interactions with other amyloid proteins.

It was confirmed that CsgB accelerates the aggregation of CsgA. In articles, it can either speed up or slow down the aggregation of amyloid β, despite both proteins having identical sequences. Additionally, CsgB speeds up the aggregation of IAPP and SEVI, but has no effect on Sup35. Considering transthyretin, it either has no effect or slows down the aggregation of CsgB, depending on the sequence length. Specifically, transthyretin with 20 fewer amino acids (27 aa) at the N-terminus has no effect, whereas that with 147 amino acids slows down the process.

9 Discussion

9.1 Experimental validation of amyloid peptides

The first problem undertaken in the dissertation was to validate the performance of our amyloid protein prediction software, AmyloGram [Burdukiewicz et al., 2017]. We selected two datasets for validation, the first, called a reference, contained the 10 peptides (5 amyloid, 5 non-amyloid) that were correctly predicted by this software according to annotations in AmyLoad database [Wozniak and Kotulska, 2015]. The second collection contained 24 peptides, 12 predicted by AmyloGram as false positives, and 12 as false negatives. To experimentally validate the amyloidogenicity properties of these peptides, we used Thioflavin T (ThT) assay and Atomic Force Microscopy (AFM). We used the reference dataset to set the relevant parameters and work out the entire testing protocol. The reference peptides revealed properties in the experiments expected from the computational predictions, which supports the efficiency of this algorithm.

An interesting case showed peptide SWVIIE, which was predicted as non-amyloid matching also the database annotation. This peptide gave unexpectedly a high signal in the ThT assay, despite its indication as a non-amyloid. By using AFM, we determined that it did not form fibrils but amorphous oligomers, which can also be made up of cross-β sheets. The results indicate that oligomeric forms can also bind thioflavin T providing a misleading conclusion that they can create amyloid fibrils. It also shows how important is validation by microscopic methods, e.g., AFM or EM, of the results from ThT assay [Gosal et al., 2006, Fitzpatrick and Saibil, 2019, Martins et al., 2020].

A good performance of AmyloGram was confirmed in the analyses of 24 peptides showing contradictory computational predictions and database annotations. Six peptides predicted as amyloids bounded ThT in the experiments, whereas ten recognized as non-amyloid did not. Thereby, we were able to find 16 out of the 24 sequences that were mislabelled in the database. Even though the data were erroneous, our predictor, AmyloGram, proved resistant to overfitting and was able to identify the mislabelled sequences in its training dataset.

It should also be added that the peptide FTFIQF was initially annotated as non-amyloid in AmyLoad, but AmyloGram and ThT assays concluded that it shows amyloidogenic properties.

AmyloGram has been proven correct by another amyloid database, where they have AFM image of fibrils created by the FTFIQF peptide [Louros et al., 2020a].

The results indicate that it is important to verify experimentally even those peptides that are predicted with high probability and annotated in databases. Due to the variable performance of the peptides and various applied experimental procedures, a standard should be worked out in order as the results are comparable and the weight of sequences used to train computational models to be similar. The obtained results can help to improve the current software and set more precise cut-offs for amyloid prediction using experimental peptide data.

9.2 Bioinformatic and phylogenetic analyses of CsgA and CsgB

One of the main aims of this dissertation was a detailed computational analysis of CsgA and CsgB amyloid proteins, which fulfill a common role in fibril formation and aggregation [Chapman et al., 2002, Wang et al., 2008]. So far only the closest homologs were investigated [Dueholm et al., 2012, Christensen et al., 2019], and the evolutionary history and phylogenetic relationships of these proteins were not known. Therefore, we applied more objective motif searching with statistical evaluation of this finding. We found five repeated regions in sequences of both proteins. The regions are separated by 1 to 2 residues in CsgA, but are adjacent in CsgB. The consensus motif of CsgA regions is twenty-one amino acids long and is characterized by nine or more conserved sites, whereas in CsgB, it has 22 amino acids and includes at least 7 conserved sites. Interestingly, the found motifs differ from those identified by [Hammer et al., 2007], which are shorter and shifted by two residues in the case of CsgA and by three residues in CsgB (Fig. 3).

The motifs in these two proteins share common features. They contain a central glycine surrounded by polar residues (asparagine, glutamine and serine) and hydrophobic residues (valine and isoleucine) occurring alternatively. It seems that such an organization enables the formation of β-sheets by the regions interacting with each other and in consequence fibril creation. The glycine breaks two β-strands created by the halves of the given region. In agreement with that, glycine is commonly known as a secondary structure breaker and is frequent in turns. Asparagine and serine are more frequent in β-sheets than α-helices. Glutamine is generally present in both secondary structure types in similar content, but more frequent than in turns.

Hydrophobic branched amino acids, such as valine and isoleucine) are also favored to be found in β-strands in the middle of β-sheets.

Using a more sensitive internal sequence comparison, we found an additional region located before the others and showing significant similarity to those five already identified, especially in CsgA. This region demonstrated the presence of β-strands such as the five duplicated regions, which supports this hypothesis that the analyzed amyloid proteins could have in the past at least six similar repeating motifs. Five of them stayed more conserved and the other diverged. It is not inconceivable that the region between the signal peptide and region 1 also participates in aggregation.

In spite of the similar organization, sequences of CsgA and CsgB showed similarity only up to 30% [Zhou et al., 2012c]. Therefore, we tried to find out whether the similar structural organization is a result of convergence or whether these proteins are distant homologs and share a common ancestry. That is why we performed sensitive searches for distant homologs and conducted a comparison of HMM profiles based on sequences classified into clusters according to similarity.

We identified 15,703 potential homologous sequences that contained at least one of the three conserved curlin domains characteristic of the reference CsgA and CsgB proteins. The homologs also showed a typical N-terminal signal peptide with a length usually from 20 to 26 residues. It indicates that these proteins are secretory as the reference CsgA and CsgB. The prediction of the secondary structure of the reference sequences from *E. coli* showed that signal peptide in CsgB folds likely into α-helix, whereas that in CsgA can also adopt β-strand.

Clustering analysis, indicated that these proteins are distant homologs not directly related, and their evolution was longer and more complex than expected. More than 98% homologs were found in *Bacteria* and only some sequences in Archaea and viruses. They can represent cases of horizontal gene transfer. Most findings in Eukaryota are probably contamination or false positives because many of these sequences showed the presence of other conserved domains, which can resemble curli motifs due to molecular convergence. Considering the Bacteria domain, these homologs are predominantly present in two phyla *Bacteroidota* and *Proteobacteria*, especially in α-*Proteobacteria* and γ-*Proteobacteria*. It indicates that in these groups the evolution of these proteins have occurred.

Our analyses showed that homologs from *Bacteroidota* are closely related with those from *γ-Proteobacteria*. In global phylogenomics studies, these two groups are distant lineages [Zhu et al., 2019, Hug et al., 2016], so we can assume that *Bacteroidota* gained a curli gene in the way of horizontal gene transfer from *γ-Proteobacteria*. Some sequences identified in other bacteria phyla and classes, present in minority and branched in phylogenetic trees contrary to their taxonomic affiliation, can be also associated with horizontal gene transfer, e.g. between *Proteobacteria* subgroups. Phylogenetic analyses demonstrated that the division of CsgA and CsgB lineages occurred after the divergence of *γ-Proteobacteria* from *α-* and *β-Proteobacteria*, but before the radiation of *γ-Proteobacteria*.

Detailed phylogenetic analyses in closer relatives to *E. coli* CsgA and CsgB, mainly in *Enterobacterales* identified potential case of horizontal gene transfer of CsgA and CsgB from *Ewingella* to *Pseudomonas reactans* as well as CsgA from *Enterobacter* to *Astraeus odoratus* and CsgB from *Enterobacterales* to *Bacteroidales*. The separation of some *Enterobacteriaceae genera*, i.e. *Kluyvera*, *Shimwellia* and *Klebsiella*, from the main clade of this family, can also suggest that CsgA and CsgB genes were transferred to them from representatives of other *Enterobacterales* families. It is also possible that the taxonomic classification of these genera is not correct, and they should not move to other families.

Investigations of five duplicated regions in *Enterobacterales* showed that those in CsgA homologs are characterized by seven conserved sites including in the order: serine, a hydrophobic residue (valine or isoleucine), glutamine, glycine, asparagine, alanine and glutamine. In CsgB homologs, there are six conserved sites across all five regions containing in the order: alanine, a hydrophobic residue (valine or isoleucine), glutamine, asparagine, a hydrophobic residue (valine or isoleucine), and glutamine. The most deviated is the 5th region. When we consider only four more conserved regions, they will share additional three conserved sites including glycine, asparagine and alanine. The conserved residues are important in formation of *β*-strands, which create *β*-sheets and specific fibril organization of these proteins. It should be added that each region is also characterized by specific residues, which distinguish it from the others. These characteristic residues are conserved across compared sequences in various taxa, which suggests that they can be also important in forming appropriate secondary structure and interactions between regions. It also indicates that a similar secondary structure can be formed by different

sequences. The collected sequences of regions can be used to construct HMM profiles individually for CsgA and CsgB or even particular regions. These profiles can help in more sensitive searches of curlin domains.

Generally, CsgA regions showed a larger variation than those in CsgB, which means that CsgA evolves quicker than CsgB. The greatest number of substitutions were accumulated in region 4 and the smallest in 5 in CsgA homologs. The smallest divergence of region 5 can be related with its role in direct interaction with region 1 from CsgB [Dunbar et al., 2019]. In CsgB sequences, region 2 evolved the fastest, and region 5 was also the least changed. It is interesting that the CsgB region 5, the most deviated from the common motif, demonstrated the smallest sequence variation within *Enterobacterales*. The conservation of region 5 in CsgB can be associated with its role in the nucleation of CsgA [Hammer et al., 2012].

Based on the phylogenetic tree of HMM profiles of individual regions we can propose a potential order of duplication of these regions presented in Fig. 35. They were duplicated in a different order in CsgA and CsgB homologs. Definitely, the regions in one protein share a common ancestry and the duplication events occurred at first within a given protein lineage.

Pairwise comparisons of distances calculated for regions, showed that potentially interacting regions evolved in a more correlated manner than those more distant in the structure. Larger correlations were demonstrated by CsgA regions, which can mean that interactions between these regions should be more conserved in this protein than in CsgB. In fact, the interactions of between CsgA regions result in the formation of amyloid fibrils and CsgB initiates this process [Hammer et al., 2007, Shu et al., 2012].

9.3 Experimental analyses of CsgA and CsgB variants without selected regions

The subsequent problem undertaken in this dissertation was to evaluate the importance of five individual regions of CsgA and CsgB proteins in their aggregation process. We successfully constructed plasmids encoded by various variants of CsgA and CsgB proteins, a wild type and five variants with deleted one of the regions. We also were able to effectively purify these proteins. It should be emphasized that this experimental endeavor was not an easy task because

the extraction and purification of amyloid proteins and their use in further studies is difficult due to a high tendency to aggregation. The purification of CsgA variants using Cobalt and Ni-NTA resins was successful but those of CsgB failed, despite many trials in small-scale production, and the visible product on the Western Blot. One reason may be that purifying the CsgB protein does not scale from small to large-scale production as well as it does in the case of CsgA. Another reason could be that CsgB is much more toxic for a cell than CsgA, so when produced on a larger scale, it caused the cell producing it can die. In the future, we will test this by gradually scaling up the production.

The variant after deletion of R1 region turned out the most efficiently aggregating according to our ThT assay analyses. It suggests that this region, although also participating in the fibril polymerization process, regulates and slows down the aggregation in the native CsgA protein. On the other hand, the variant with R5 deletion very poorly bounded thioflavin T and required more time in the aggregation process. The removal of R4 region did not increase substantially the fluorescence in ThT assay either but improved the aggregation better than ΔR5 variant. Other variants and wild type were placed rather between the most extreme variants. CsgA WT showed very poor aggregation in some experiments, although we would like to note that this protein eluted the slowest during purification. This may reflect the fact that it had already started aggregation on the deposit and stayed there, thus we were unable to capture the aggregation kinetics.

The results indicate that R5 and also R4 are more important in the aggregation than others. The importance of R5 very well corresponds to the slowest evolution rate of this region in *Enterobacterales* as found in this thesis. The R5 region interacting with R1 of the other molecule, and R4 interacting with R5 can control the start of the aggregation process, without which amyloid fibrils cannot form. The R2 and R3 regions may have a smaller impact on the aggregation process than other due to the fact that they are placed in the middle of the folded protein. The results are comparable with those obtained by other experiments [Wang et al., 2008], which found that the lag phase for rapid fibril growth is the largest for variants without R5 and the shortest for R1. The difference concerns the WT and ΔR2 variant. They noticed the aggregation efficiency is slightly better for WT than ΔR2, but our studies indicated that it is comparable or greater for the latter.

We observed amyloid fibrils under AFM only for selected variants, including the very well aggregating ΔR1 variant. A problem with obtaining the fibril under AFM, occurred even when we changed the polarity of the micas to make them better adhere to the substrate. We suspect that one week of incubation might have been too short in the case of these proteins to obtain long and visible amyloid fibrils like in the works of Sleutel et al. [2017], or insulin fibrils in Sakalauskas et al. [2019]. Increasing the concentration did not help their observation, either. Alternatively, CsgA fibrils need a more modified protocol for their preparation on mica. Therefore, we plan to refine the purification protocol so that the purified proteins would have more reproducible results when measuring aggregation kinetics before analysis with the HDX-MS technique.

9.4 Comparison of functional and non-functional amyloids in terms of sequence features

Besides CsgA and CsgB, there are also other functional amyloids that fulfill various functions and represent different protein families in prokaryotes and eukaryotes [Maury, 2009, Van Gerven et al., 2018]. Therefore, we decided to compare the functional and non-functional amyloids in terms of sequence features that can be used to elaborate a prediction model. Detailed statistical and discriminant studies showed that sequences in these groups are characterized by specific features, which can be used in their recognition. Sequences of functional amyloids show a high frequency of small hydroxylated amino acids, serine and threonine. In consequence, they are also rich in dipeptides including these residues, i.e. ST, TS, TT, SS, AS, SA, SG, TG, AT, GS, SN, GT and DS. It can be noticed that these amino acids co-occur with other small amino acids, glycine and alanine, and include polar asparagine and aspartic acid. On the other hand, sequences of non-functional amyloids contain more basic amino acids (arginine, lysine and histidine), hydrophobic (leucine, methionine and cysteine) and polar tyrosine. The most common dipeptides in these sequences are LK, AK, KV, LL, RV, FQ, ER, KE, EL, EK, VR, LR and DL. These specific compositions can be associated with the distinct behavior of their structures. The non-functional amyloids have a tendency to adopt other conformations, e.g. β-cross, which causes the loss of functionality and is related with many disorders. In agreement with that our analyses demonstrated that discriminating amino acid indices are related mainly

with loop, turn, β-sheet, β-turn, β region, flexibility, buriability, hydrophobicity, net charge, protein stability, optimized relative partition energies and free energy change.

Interestingly, in each group of these amyloids, it is possible to identify subgroups of sequences that show different composition and sequence descriptors. It indicates that various types of proteins belong to these proteins. The heterogeneity was revealed in plots of methods reducing multidimensionality.

We studied many features and the best discriminating occurred dipeptide composition and amino acid indices in random forest models. Measures used in the assessment of classifiers and predictors, i.e. precision, sensitivity, specificity, accuracy, AUC and MCC achieved values mostly higher than 0.9. Based on that a very good predictor for the functional and non-functional amyloids can be elaborated.

9.5 Database of interactions between amyloids

The final problem undertaken in this dissertation was the creation of an amyloid interaction database, which would include information on interactions between the CsgA and CsgB proteins and other amyloids. Interactions between amyloid proteins were the subject of many experimental studies. Although, there are several databases that collect data on amyloid proteins [Wozniak and Kotulska, 2015, Louros et al., 2020a, Varadi et al., 2018, Rawat et al., 2020, Pawlicki et al., 2008], they do not have data on interactions between them. Ren et al. [2019] has attempted to systematize and organize the data about the interactions between amyloids. Although he included many amyloid proteins in his work, he did not describe functional amyloids, CsgA and CsgB. Moreover, we can find contradictory interaction data in many papers dealing with interamyloid interactions [Tran et al., 2017]. One of the reasons for this is the lack of clear definitions of interactions and the required standardized experiments to determine this. This makes the analysis of interactions between amyloids problematic when we want to compare different experimental results.

Therefore, to create our database of amyloid interactions, we began by designing six scenarios of what such interactions look like. And three descriptors describe interactions' effect on aggregation rates, whether and how they bind, and what type of fibril they form. Thereby, we have organized and systematized the terminology related to amyloid interactions. Thanks to the

development of the database, we were able to find 48 interactions of CsgA and 14 interactions of CsgB with other amyloid proteins. CsgB reacts with five and CsgA with additional three amyloid proteins. After in-depth sequence analyses, this information can give us a preview of which regions of CsgA and why they are able to bind an additional three amyloids than CsgB, even though they have a very similar structure. Information in the database can help to understand how individual proteins affect, inhibit or accelerate the aggregation process. It may also serve in future analyses to identify sequences that will be able to efficiently inhibit or accelerate the aggregation of selected amyloid proteins associated with diseases.

The created interactive database based on the Shiny [Chang et al., 2022] server can be intelligible for many users. Based on the gathered data, we want to elaborate on a predictor of regions that might be more aggregation-prone or responsible for nucleation. At the moment, the database is focused on the whole families of protein homologs or single variants. But in the future, we would like to expand it significantly, including information on the effect of small molecules on aggregation. In addition, we also want to add information on the conditions of the experiment, such as pH, temperature, or protein concentration [Pfefferkorn et al., 2010, Hu et al., 2009].

The analysis of the interactions between CsgA and CsgB and other amyloid proteins provides valuable insights into the complex nature of protein aggregation. The fact that CsgA and CsgB react with essentially the same five proteins suggests that these proteins may play a key role in the aggregation process. However, the exceptions, such as Tau, α-synuclein, and lysozyme, which only react with CsgA, indicate that there are specific interactions between these proteins that may contribute to the differences in their aggregation behavior. The acceleration of CsgA aggregation by other CsgA is a common observation, although there are also some papers that show no interaction at all. This variability may be attributed to differences in experimental conditions or protein concentrations, which can affect the rate of aggregation. It is interesting that a protein sequence that is only two amino acids longer than CsgA and CsgB, i.e. amyloid β, does not affect the acceleration of CsgA aggregation, while those 40 amino acids long do. It highlights the importance of some sequence features in protein-protein interactions and aggregation.

The mutually exclusive results observed with Sup35, where one publication indicates acceler-

ation and the other slowing of the interaction, underscores the complexity of protein aggregation and the need for further investigation to fully understand the mechanisms involved. Similarly, the different effects of α-synuclein and IAPP on the rate of aggregation depending on whether they are the interactee or interactor suggest that the behavior of these proteins is highly context-dependent.

The fact that Tau and lysosome have no effect on aggregation, whereas SEVI accelerates CsgA aggregation indicates that different proteins may have distinct roles in the aggregation process. The findings regarding CsgB interactions with other amyloid proteins further emphasize the complexity of protein aggregation and the need for further research to fully understand the mechanisms involved.

The interactions between proteins and their effects on each other are complex and multi-faceted. This is highlighted by the different effects that CsgA has on various proteins, as well as the effects of other proteins on CsgA. One interesting observation is that the sequence variation in CsgA can lead to different effects on the aggregation of α-synuclein and amyloid β. This suggests that the specific amino acid sequence of a protein can have a significant impact on its interactions with other proteins. It is also noteworthy that CsgA can affect the aggregation of a variety of different proteins, including itself, IAPP, SEVI, and Sup35. This suggests that CsgA may have a broader role in protein aggregation than previously thought. Transthyretin also has a complex relationship with both CsgA and CsgB, with varying effects depending on the specific sequences involved. This highlights the fact that protein-protein interactions can be highly specific and context-dependent or there may be other factors at play beyond just the amino acid sequence of the protein.

Overall, the analysis of the interactions between CsgA and CsgB and other amyloid proteins provides valuable insights into the complex nature of protein aggregation and highlights the need for further investigation to fully understand the mechanisms involved. Moreover, the found interactions between amyloids underscore the need for further research to better understand the mechanisms underlying protein-protein interactions and their effects on protein aggregation. By gaining a deeper understanding of these interactions, we may be able to develop new approaches for preventing or treating protein aggregation-related diseases such as Alzheimer's and Parkinson's.

10 Conclusions

- Our algorithm AmyloGram designed for the prediction of amyloids is resistant to overfitting and occurred efficient in recognition of experimentally validated peptides.
- We identified 16 out of the 24 peptides that were incorrectly annotated in the database AmyLoad.
- Some peptides, e.g. SWVIIE, which are amorphous oligomers and do not form fibrils, can give a high signal in the ThT assay.
- More objective motif searching with statistical evaluation showed five repeated regions in sequences of CsgA and CsgB. CsgA repeating motifs with a length of 21 residues are separated by one or two amino acids, whereas those in CsgB has a length of 22 residues and with no separation.
- The consensus motif in CsgA regions includes nine or more conserved sites, whereas that in CsgB comprises at least seven conserved sites. The composition and distribution of polar and hydrophobic residues corresponds to the presence of two β-strands per region broken by the central glycine.
- A more sensitive internal sequence comparison revealed an extra region located before the others, showing similarity to them and possibly folding into β-strands.
- There are at least 15,703 potential homologs of CsgA and CsgB that comprise conserved curlin domains. Most of them are also equipped with N-terminal signal peptide typical of the reference CsgA and CsgB.
- CsgA and CsgB homologs evolved mainly in Bacteria, especially in *Bacteroidota* as well as α-*Proteobacteria* and γ-*Proteobacteria*.
- CsgA and CsgB are distant homologs that arose by duplication after separation of γ-*Proteobacteria* from α- and β-*Proteobacteria* but before γ-*Proteobacteria* differentiation.

- The CsgA and CsgB homologs were likely subjected to horizontal gene transfer, e.g. between various *Proteobacteria* subgroups, from γ-*Proteobacteria* to *Bacteroidota*, and possibly from *Enterobacterales* to *Pseudomonas reactans* and fungus *Astraeus odoratus*.

- Five duplicated regions of CsgA and CsgB in *Enterobacterales* exhibit seven and six conserved sites respectively, including hydrophobic, polar and glycine residues, which are important in the formation of β-strands. Each region can be distinguished from the others by distinctive conserved residues.

- Regions in CsgA evolve faster than those in CsgB. Region 5 shows the smallest divergence rate in both proteins probably due to the selection on interactions with region 1 of other molecules of the curly proteins.

- CsgA and CsgB regions were duplicated in a different order and the duplication events occurred before the lineages of these proteins separated.

- The evolution of potentially interacting regions in the curli proteins was generally more correlated than those of more remote in the structure. CsgA regions showed stronger correlations in amino acid substitutions, which may indicate that interactions between these regions in this protein should be more conserved than in CsgB.

- CsgA ΔR1 turned out the most efficiently aggregating, suggesting that region 1 is responsible for slowing down the process.

- CsgA ΔR5 needs much more time to start aggregation, which suggests that region 5 plays a crucial role in the polymerization process of amyloid fibrils, e.g. due to interaction with region 1 of other CsgA molecule.

- The other of CsgA variants were placed between the aforementioned protein constructs it terms of aggregation speed.

- Wild type of CsgA showed a very long lag phase and low fluorescence intensity, which may result from the fact that it had already started to aggregate during elution and was left on the resin.

- Although functional and non-functional amyloids are heterogeneous groups, they have distinctive characteristics that can be used in their prediction. Sequences of functional amyloids have a high content of small hydroxylated amino acids, serine and threonine co-occurring with other small glycine and alanine, and polar asparagine and aspartic acid. However, sequences of non-functional amyloids comprise more basic arginine, lysine and histidine, hydrophobic leucine, methionine and cysteine as well as and polar tyrosine.

- The functional and non-functional amyloids can be effectively predicted in random forest models based on dipeptide composition and amino acid indices associated with various secondary structures, flexibility, buriability, hydrophobicity, net charge, protein stability, optimized relative partition energies and free energy change.

- Based on a newly developed database of amyloid interactions AmyloGraph, we identified 48 interactions of CsgA and 14 interactions of CsgB with other amyloid proteins. CsgA interacts with 9 proteins, whereas CsgB with 6.

- The database can be useful in determining how individual proteins affect the aggregation process.

11 References

J. Adamcik and R. Mezzenga. Study of amyloid fibrils via atomic force microscopy. *Current Opinion in Colloid & Interface Science*, 17(6):369–376, Dec. 2012. ISSN 1359-0294. doi: 10.1016/j.cocis.2012.08.001.

A. Aguzzi and A. M. Calella. Prions: Protein aggregation and infectious diseases. *Physiological Reviews*, 89(4):1105–1152, Oct. 2009. ISSN 0031-9333. doi: 10.1152/physrev.00006.2009.

A. Aguzzi and A. K. K. Lakkaraju. Cell Biology of Prions and Prionoids: A Status Report. *Trends in Cell Biology*, 26(1):40–51, Jan. 2016. ISSN 1879-3088. doi: 10.1016/j.tcb.2015.08.007.

S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, Sept. 1997. ISSN 0305-1048. doi: 10.1093/nar/25.17.3389.

M. Andreasen, G. Meisl, J. D. Taylor, T. C. T. Michaels, A. Levin, D. E. Otzen, M. R. Chapman, C. M. Dobson, S. J. Matthews, and T. P. J. Knowles. Physical Determinants of Amyloid Assembly in Biofilm Formation. *mBio*, 10(1), Jan. 2019a. ISSN 2150-7511. doi: 10.1128/mBio.02279-18.

M. Andreasen, G. Meisl, J. D. Taylor, T. C. T. Michaels, A. Levin, D. E. Otzen, M. R. Chapman, C. M. Dobson, S. J. Matthews, and T. P. J. Knowles. Physical Determinants of Amyloid Assembly in Biofilm Formation. *mBio*, 10(1):e02279–18, Feb. 2019b. ISSN 2150-7511. doi: 10.1128/mBio.02279-18.

C. Anfinsen and H. Scheraga. Experimental and Theoretical Aspects of Protein Folding. In *Advances in Protein Chemistry*, volume 29, pages 205–300. Elsevier, 1975. ISBN 978-0-12-034229-7. doi: 10.1016/S0065-3233(08)60413-1.

C. B. Anfinsen. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, July 1973. doi: 10.1126/science.181.4096.223.

K. Annamalai, K.-H. Gührs, R. Koehler, M. Schmidt, H. Michel, C. Loos, P. M. Gaffney, C. J. Sigurdson, U. Hegenbart, S. Schönland, and M. Fändrich. Polymorphism of Amyloid Fibrils In Vivo. *Angewandte Chemie (International Ed. in English)*, 55(15):4822–4825, Apr. 2016. ISSN 1521-3773. doi: 10.1002/anie.201511524.

A. Arnqvist, A. Olsén, J. Pfeifer, D. G. Russell, and S. Normark. The Crl protein activates cryptic genes for curli formation and fibronectin binding in Escherichia coli HB101. *Molecular Microbiology*, 6(17):2443–2452, Sept. 1992. ISSN 0950-382X. doi: 10.1111/j.1365-2958.1992.tb01420.x.

P. Arosio, T. P. J. Knowles, and S. Linse. On the lag phase in amyloid fibril formation. *Physical Chemistry Chemical Physics*, 17(12):7606–7618, 2015. doi: 10.1039/C4CP05563B.

C. Ascoli, F. Dinelli, C. Frediani, D. Petracchi, M. Salerno, M. Labardi, M. Allegrini, and F. Fuso. Normal and lateral forces in scanning force microscopy. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena*, 12(3):1642–1645, May 1994. ISSN 1071-1023. doi: 10.1116/1.587251.

A. E. Badaczewska-Dawid, J. Garcia-Pardo, A. Kuriata, J. Pujols, S. Ventura, and S. Kmiecik. A3D Database: Structure-based Protein Aggregation Predictions for the Human Proteome, Nov. 2021.

G. Bahramali, B. Goliaei, Z. Minuchehr, and A. Salari. Chameleon sequences in neurodegenerative diseases. *Biochemical and Biophysical Research Communications*, 472(1):209–216, Mar. 2016. ISSN 0006-291X. doi: 10.1016/j.bbrc.2016.01.187.

T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, 2:28–36, 1994. ISSN 1553-0833.

T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble. The MEME Suite. *Nucleic Acids Research*, 43(W1):W39–W49, July 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv416.

K. R. Baker and L. Rice. The Amyloidoses: Clinical Features, Diagnosis and Treatment. *Methodist DeBakey Cardiovascular Journal*, 8(3):3–7, 2012. ISSN 1947-6094.

A. Balistreri, E. Goetzler, and M. Chapman. Functional Amyloids Are the Rule Rather Than the Exception in Cellular Biology. *Microorganisms*, 8(12):1951, Dec. 2020. ISSN 2076-2607. doi: 10.3390/microorganisms8121951.

S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen. Biomolecular condensates: Organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology*, 18(5):285–298, May 2017. ISSN 1471-0080. doi: 10.1038/nrm.2017.7.

A. M. Barbosa. fuzzySim: Applying fuzzy logic to binary similarity indices in ecology. *Methods in Ecology and Evolution*, 6(7):853–858, 2015. ISSN 2041-210X. doi: 10.1111/2041-210X.12372.

M. M. Barnhart and M. R. Chapman. Curli Biogenesis and Function. *Annual review of microbiology*, 60:131–147, 2006. ISSN 0066-4227. doi: 10.1146/annurev.micro.60.080805.142106.

M. Belli, M. Ramazzotti, and F. Chiti. Prediction of amyloid aggregation in vivo. *EMBO Reports*, 12(7):657–663, July 2011. ISSN 1469-221X. doi: 10.1038/embor.2011.116.

M. J. Benskey, R. G. Perez, and F. P. Manfredsson. The contribution of alpha synuclein to neuronal survival and function – Implications for Parkinson's disease. *Journal of Neurochemistry*, 137(3):331–359, 2016. ISSN 1471-4159. doi: 10.1111/jnc.13570.

K. Berthelot, S. Lecomte, Y. Estevez, B. Coulary-Salin, A. Bentaleb, C. Cullin, A. Deffieux, and F. Peruch. Rubber Elongation Factor (REF), a Major Allergen Component in Hevea brasiliensis Latex Has Amyloid Properties. *PLoS ONE*, 7(10):e48065, Oct. 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0048065.

M. Biancalana and S. Koide. Molecular Mechanism of Thioflavin-T Binding to Amyloid Fibrils. *Biochimica et biophysica acta*, 1804(7):1405–1412, July 2010. ISSN 0006-3002. doi: 10.1016/j.bbapap.2010.04.001.

M. Biancalana, K. Makabe, A. Koide, and S. Koide. Molecular Mechanism of Thioflavin-T Binding to the Surface of β-Rich Peptide Self-Assemblies. *Journal of Molecular Biology*, 385 (4):1052–1063, Jan. 2009. ISSN 0022-2836. doi: 10.1016/j.jmb.2008.11.006.

L. P. Blanco, M. L. Evans, D. R. Smith, M. P. Badtke, and M. R. Chapman. Diversity, biogenesis and function of microbial amyloids. *Trends in Microbiology*, 20(2):66–73, Feb. 2012. ISSN 0966-842X. doi: 10.1016/j.tim.2011.11.005.

D. C. Bolton, M. P. McKinley, and S. B. Prusiner. Identification of a Protein That Purifies with the Scrapie Prion. *Science*, 218(4579):1309–1311, Dec. 1982. doi: 10.1126/science.6815801.

S. A. Bondarev, O. V. Bondareva, G. A. Zhouravleva, and A. V. Kajava. BetaSerpentine: A bioinformatics tool for reconstruction of amyloid structures. *Bioinformatics*, 34(4):599–608, Feb. 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx629.

A. V. Bryksin and I. Matsumura. Overlap extension PCR cloning: A simple and reliable way to create recombinant plasmids. *BioTechniques*, 48(6):463–465, June 2010. ISSN 0736-6205. doi: 10.2144/000113418.

M. Burdukiewicz, P. Sobczyk, S. Rödiger, A. Duda-Madej, P. Mackiewicz, and M. Kotulska. Amyloidogenic Motifs Revealed by N-Gram Analysis. *Scientific Reports*, 7(1):12961, Oct. 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-13210-9.

M. Burdukiewicz, D. Rafacz, A. Barbach, K. Hubicka, L. Bąkała, A. Lassota, J. Stecko, N. Szymańska, J. W. Wojciechowski, D. Kozakiewicz, N. Szulc, J. Chilimoniuk, I. Jęśkowiak, M. Gąsior-Głogowska, and M. Kotulska. AmyloGraph: A comprehensive database of amyloid–amyloid interactions. *Nucleic Acids Research*, page gkac882, Oct. 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac882.

E. Callaway. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*, 588(7837):203–204, Nov. 2020. doi: 10.1038/d41586-020-03348-4.

C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. BLAST+: Architecture and applications. *BMC Bioinformatics*, 10(1):421, Dec. 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421.

S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, 25(15):1972–1973, Aug. 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp348.

A. Cárdenas, J.-B. Raina, C. Pogoreutz, N. Rädecker, J. Bougoure, P. Guagliardo, M. Pernice, and C. R. Voolstra. Greater functional diversity and redundancy of coral endolithic microbiomes align with lower coral bleaching susceptibility. *The ISME Journal*, 16(10):2406–2420, Oct. 2022. ISSN 1751-7370. doi: 10.1038/s41396-022-01283-y.

A. Carija, S. Navarro, N. S. de Groot, and S. Ventura. Protein aggregation into insoluble deposits protects from oxidative stress. *Redox Biology*, 12:699–711, Aug. 2017. ISSN 2213-2317. doi: 10.1016/j.redox.2017.03.027.

W. Chang, J. Cheng, J. J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, B. Borges, RStudio, j. F. j. l. a. j. U. library), j. c. j. library; authors listed in inst/www/shared/jquery-AUTHORS.txt), j. U. c. j. U. library; authors listed in inst/www/shared/jqueryui/AUTHORS.txt), M. O. B. library), J. T. B. library), B. c. B. library), Twitter, I. B. library), P. N. K. B. accessibility plugin), V. T. B. accessibility plugin), D. L. B. accessibility plugin), S. C. B. accessibility plugin), C. O. B. accessibility plugin), PayPal, I. B. accessibility plugin), S. P. B.-d. library), A. R. B.-d. library), B. R. s. js library), S. B. s.-p.-a. library), D. I. i. rangeSlider library), S. S. J. strftime library), S. L. D. library), J. F. s. js library), J. G. s. js library), I. S. h. js library), and R. C. T. t. implementation from R). Shiny: Web Application Framework for R, Dec. 2022.

M. R. Chapman, L. S. Robinson, J. S. Pinkner, R. Roth, J. Heuser, M. Hammar, S. Normark, and S. J. Hultgren. Role of Escherichia Coli Curli Operons in Directing Amyloid Fiber Formation. *Science (New York, N.Y.)*, 295(5556):851–855, Feb. 2002. ISSN 1095-9203. doi: 10.1126/science.1067484.

D. Charif, O. Clerc, C. Frank, J. R. Lobry, A. Necşulea, L. Palmeira, S. Penel, and G. Perrière. Seqinr: Biological Sequences Retrieval and Analysis, Nov. 2022.

P. Chien, J. S. Weissman, and A. H. DePace. Emerging principles of conformation-based prion inheritance. *Annual Review of Biochemistry*, 73:617–656, 2004. ISSN 0066-4154. doi: 10.1146/annurev.biochem.72.121801.161837.

F. Chiti and C. M. Dobson. Protein Misfolding, Amyloid Formation, and Human Disease: A

Summary of Progress Over the Last Decade. *Annual Review of Biochemistry*, 86:27–68, June 2017. ISSN 1545-4509. doi: 10.1146/annurev-biochem-061516-045115.

L. F. B. Christensen, N. Schafer, A. Wolf-Perez, D. J. Madsen, and D. E. Otzen. Bacterial Amyloids: Biogenesis and Biomaterials. In S. Perrett, A. K. Buell, and T. P. Knowles, editors, *Biological and Bio-inspired Nanomaterials: Properties and Assembly Mechanisms*, Advances in Experimental Medicine and Biology, pages 113–159. Springer, Singapore, 2019. ISBN 9789811397912. doi: 10.1007/978-981-13-9791-2_4.

O. Conchillo-Solé, N. S. de Groot, F. X. Avilés, J. Vendrell, X. Daura, and S. Ventura. AGGRESCAN: A server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*, 8(1):65, Feb. 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-65.

G. J. Cooper, A. C. Willis, A. Clark, R. C. Turner, R. B. Sim, and K. B. Reid. Purification and characterization of a peptide from amyloid-rich pancreases of type 2 diabetic patients. *Proceedings of the National Academy of Sciences of the United States of America*, 84(23): 8628–8632, Dec. 1987. ISSN 0027-8424.

B. Cox, F. Ness, and M. Tuite. Analysis of the generation and segregation of propagons: Entities that propagate the [PSI+] prion in yeast. *Genetics*, 165(1):23–33, Sept. 2003. ISSN 0016-6731. doi: 10.1093/genetics/165.1.23.

F. o. b. L. B. a. A. Cutler and R. p. b. A. L. a. M. Wiener. randomForest: Breiman and Cutler's Random Forests for Classification and Regression, May 2022.

R. K. Das and R. V. Pappu. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences*, 110(33):13392–13397, Aug. 2013. doi: 10.1073/pnas.1304749110.

N. S. de Groot, I. Pallarés, F. X. Avilés, J. Vendrell, and S. Ventura. Prediction of "Hot Spots" of Aggregation in Disease-Linked Polypeptides. *BMC Structural Biology*, 5:18, Sept. 2005. ISSN 1472-6807. doi: 10.1186/1472-6807-5-18.

A. Drozdetskiy, C. Cole, J. Procter, and G. J. Barton. JPred4: A protein secondary structure

prediction server. *Nucleic Acids Research*, 43(W1):W389–W394, July 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv332.

E. Drummond, S. Nayak, A. Faustin, G. Pires, R. Hickman, M. Askenazi, M. Cohen, T. Haldiman, C. Kim, X. Han, Y. Shao, J. G. Safar, B. Ueberheide, and T. Wisniewski. Proteomic Differences in Amyloid Plaques in Rapidly Progressive and Sporadic Alzheimer's Disease. *Acta neuropathologica*, 133(6):933–954, June 2017. ISSN 0001-6322. doi: 10.1007/s00401-017-1691-0.

M. S. Dueholm, S. V. Petersen, M. Sønderkær, P. Larsen, G. Christiansen, K. L. Hein, J. J. Enghild, J. L. Nielsen, K. L. Nielsen, P. H. Nielsen, and D. E. Otzen. Functional Amyloid in Pseudomonas. *Molecular Microbiology*, 77(4):1009–1020, Aug. 2010. ISSN 1365-2958. doi: 10.1111/j.1365-2958.2010.07269.x.

M. S. Dueholm, M. Albertsen, D. Otzen, and P. H. Nielsen. Curli Functional Amyloid Systems Are Phylogenetically Widespread and Display Large Diversity in Operon and Protein Structure. *PloS One*, 7(12):e51274, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0051274.

M. S. Dueholm, M. T. Søndergaard, M. Nilsson, G. Christiansen, A. Stensballe, M. T. Overgaard, M. Givskov, T. Tolker-Nielsen, D. E. Otzen, and P. H. Nielsen. Expression of Fap Amyloids in Pseudomonas Aeruginosa, P. Fluorescens, and P. Putida Results in Aggregation and Increased Biofilm Formation. *MicrobiologyOpen*, 2(3):365–382, June 2013. ISSN 2045-8827. doi: 10.1002/mbo3.81.

M. Dunbar, E. DeBenedictis, and S. Keten. Dimerization energetics of curli fiber subunits CsgA and CsgB. *npj Computational Materials*, 5(1):1–9, Feb. 2019. ISSN 2057-3960. doi: 10.1038/s41524-019-0164-5.

S. R. Durell and A. Ben-Naim. Hydrophobic-hydrophilic forces in protein folding. *Biopolymers*, 107(8):e23020, 2017. ISSN 1097-0282. doi: 10.1002/bip.23020.

S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, Jan. 1998. ISSN 1367-4803. doi: 10.1093/bioinformatics/14.9.755.

S. R. Eddy. Accelerated Profile HMM Searches. *PLOS Computational Biology*, 7(10):e1002195, Oct. 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002195.

D. S. Eisenberg and M. R. Sawaya. Structural Studies of Amyloid Proteins at the Molecular Level. *Annual Review of Biochemistry*, 86(1):69–95, June 2017. ISSN 0066-4154, 1545-4509. doi: 10.1146/annurev-biochem-061516-045104.

M. A. Elliot, N. Karoonuthaisiri, J. Huang, M. J. Bibb, S. N. Cohen, C. M. Kao, and M. J. Buttner. The chaplins: A family of hydrophobic cell-surface proteins involved in aerial mycelium formation in Streptomyces coelicolor. *Genes & Development*, 17(14):1727–1740, July 2003. ISSN 0890-9369. doi: 10.1101/gad.264403.

J. R. Engen. Analysis of Protein Conformation and Dynamics by Hydrogen/Deuterium Exchange MS. *Analytical Chemistry*, 81(19):7870–7875, Oct. 2009. ISSN 0003-2700. doi: 10.1021/ac901154s.

J. J. Englander, J. R. Rogero, and S. W. Englander. Protein Hydrogen Exchange Studied by the Fragment Separation Method. *Analytical biochemistry*, 147(1):234–244, May 1985. ISSN 0003-2697.

E. Erskine, R. J. Morris, M. Schor, C. Earl, R. M. C. Gillespie, K. M. Bromley, T. Sukhodub, L. Clark, P. K. Fyfe, L. C. Serpell, N. R. Stanley-Wall, and C. E. MacPhee. Formation of Functional, Non-amyloidogenic Fibres by Recombinant Bacillus Subtilis TasA. *Molecular Microbiology*, 110(6):897–913, Dec. 2018. ISSN 0950-382X. doi: 10.1111/mmi.13985.

M. L. Evans and M. R. Chapman. Curli Biogenesis: Order out of Disorder. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1843(8):1551–1558, Aug. 2014. ISSN 0167-4889. doi: 10.1016/j.bbamcr.2013.09.010.

C. Família, S. R. Dennison, A. Quintas, and D. A. Phoenix. Prediction of Peptide and Protein Propensity for Amyloid Formation. *PLoS ONE*, 10(8):e0134679, Aug. 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0134679.

A.-M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano. Prediction of

Sequence-Dependent and Mutational Effects on the Aggregation of Peptides and Proteins. *Nature Biotechnology*, 22(10):1302–1306, Oct. 2004. ISSN 1546-1696. doi: 10.1038/nbt1012.

A. W. Fitzpatrick and H. R. Saibil. Cryo-EM of amyloid fibrils and cellular aggregates. *Current Opinion in Structural Biology*, 58:34–42, Oct. 2019. ISSN 0959-440X. doi: 10.1016/j.sbi.2019.05.003.

A. W. P. Fitzpatrick, G. T. Debelouchina, M. J. Bayro, D. K. Clare, M. A. Caporini, V. S. Bajaj, C. P. Jaroniec, L. Wang, V. Ladizhansky, S. A. Müller, C. E. MacPhee, C. A. Waudby, H. R. Mott, A. De Simone, T. P. J. Knowles, H. R. Saibil, M. Vendruscolo, E. V. Orlova, R. G. Griffin, and C. M. Dobson. Atomic structure and hierarchical assembly of a cross-β amyloid fibril. *Proceedings of the National Academy of Sciences of the United States of America*, 110 (14):5468–5473, Apr. 2013. ISSN 1091-6490. doi: 10.1073/pnas.1219476110.

D. M. Fowler, A. V. Koulov, C. Alory-Jost, M. S. Marks, W. E. Balch, and J. W. Kelly. Functional Amyloid Formation within Mammalian Tissue. *PLoS Biology*, 4(1):e6, Jan. 2006. ISSN 1544-9173. doi: 10.1371/journal.pbio.0040006.

T. Frickey and A. Lupas. CLANS: A Java application for visualizing protein families based on pairwise similarity. *Bioinformatics (Oxford, England)*, 20(18):3702–3704, Dec. 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth444.

C. Frieden. Protein aggregation processes: In search of the mechanism. *Protein Science : A Publication of the Protein Society*, 16(11):2334–2344, Nov. 2007. ISSN 0961-8368. doi: 10.1110/ps.073164107.

L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, Dec. 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts565.

R. G. Creasey, C. T. Gibson, and N. H. Voelcker. Characterization of Fiber-Forming Peptides and Proteins by Means of Atomic Force Microscopy. *Current Protein & Peptide Science*, 13 (3):232–257, May 2012. ISSN 13892037. doi: 10.2174/138920312800785058.

R. García and R. Pérez. Dynamic atomic force microscopy methods. *Surface Science Reports*, 47(6):197–301, Sept. 2002. ISSN 0167-5729. doi: 10.1016/S0167-5729(02)00077-8.

C. R. García-Jacas, S. A. Pinacho-Castellanos, L. A. García-González, and C. A. Brizuela. Do deep learning models make a difference in the identification of antimicrobial peptides? *Briefings in Bioinformatics*, 23(3):bbac094, May 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac094.

R. C. Garratt, N. F. Valadares, and J. F. R. Bachega. Oligomeric Proteins. In G. C. K. Roberts, editor, *Encyclopedia of Biophysics*, pages 1781–1789. Springer, Berlin, Heidelberg, 2013. ISBN 978-3-642-16712-6. doi: 10.1007/978-3-642-16712-6_416.

M. Garvey, S. Meehan, S. L. Gras, H. J. Schirra, D. J. Craik, N. L. Van der Weerden, M. A. Anderson, J. A. Gerrard, and J. A. Carver. A radish seed antifungal peptide with a high amyloid fibril-forming propensity. *Biochimica Et Biophysica Acta*, 1834(8):1615–1623, Aug. 2013. ISSN 0006-3002. doi: 10.1016/j.bbapap.2013.04.030.

D. M. Gendoo and P. M. Harrison. Discordant and chameleon sequences: Their distribution and implications for amyloidogenicity. *Protein Science*, 20(3):567–579, 2011. ISSN 1469-896X. doi: 10.1002/pro.590.

C. J. Gibbs, D. C. Gajdusek, D. M. Asher, M. P. Alpers, E. Beck, P. M. Daniel, and W. B. Matthews. Creutzfeldt-Jakob Disease (Spongiform Encephalopathy): Transmission to the Chimpanzee. *Science*, 161(3839):388–389, July 1968. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.161.3839.388.

L. Giehm and D. E. Otzen. Strategies to increase the reproducibility of protein fibrillization in plate reader assays. *Analytical Biochemistry*, 400(2):270–281, May 2010. ISSN 0003-2697. doi: 10.1016/j.ab.2010.02.001.

C. Goldsbury, J. Kistler, U. Aebi, T. Arvinte, and G. J. S. Cooper. Watching amyloid fibrils grow by time-lapse atomic force microscopy11Edited by W. Baumeister. *Journal of Molecular Biology*, 285(1):33–39, Jan. 1999. ISSN 0022-2836. doi: 10.1006/jmbi.1998.2299.

D. J. Gordon, J. J. Balbach, R. Tycko, and S. C. Meredith. Increasing the Amphiphilicity of an Amyloidogenic Peptide Changes the β-Sheet Structure in the Fibrils from Antiparallel to Parallel. *Biophysical Journal*, 86(1):428–434, Jan. 2004. ISSN 0006-3495.

W. S. Gosal, S. L. Myers, S. E. Radford, and N. H. Thomson. Amyloid Under the Atomic Force Microscope. *Protein and Peptide Letters*, 13(3):261–270, Mar. 2006. doi: 10.2174/092986606775338498.

S. Gour, V. Kaushik, V. Kumar, P. Bhat, S. C. Yadav, and J. K. Yadav. Antimicrobial peptide (Cn-AMP2) from liquid endosperm of Cocos nucifera forms amyloid-like fibrillar structure. *Journal of Peptide Science: An Official Publication of the European Peptide Society*, 22(4): 201–207, Apr. 2016. ISSN 1099-1387. doi: 10.1002/psc.2860.

B. J. Grant, L. Skjærven, and X.-Q. Yao. The Bio3D packages for structural bioinformatics. *Protein Science*, 30(1):20–30, 2021. ISSN 1469-896X. doi: 10.1002/pro.3923.

R. Grantham. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science (New York, N.Y.)*, 185(4154):862–4, 1974.

I. Grundke-Iqbal, K. Iqbal, Y. C. Tung, M. Quinlan, H. M. Wisniewski, and L. I. Binder. Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology. *Proceedings of the National Academy of Sciences*, 83(13):4913–4917, July 1986. doi: 10.1073/pnas.83.13.4913.

J.-T. Guo, J. W. Jaromczyk, and Y. Xu. Analysis of chameleon sequences and their implications in biological processes. *Proteins: Structure, Function, and Bioinformatics*, 67(3):548–558, 2007. ISSN 1097-0134. doi: 10.1002/prot.21285.

N. D. Hammer, J. C. Schmidt, and M. R. Chapman. The Curli Nucleator Protein, CsgB, Contains an Amyloidogenic Domain That Directs CsgA Polymerization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(30):12494–12499, July 2007. ISSN 0027-8424. doi: 10.1073/pnas.0703310104.

N. D. Hammer, B. A. McGuffie, Y. Zhou, M. P. Badtke, A. A. Reinke, K. Brännström, J. E. Gestwicki, A. Olofsson, F. Almqvist, and M. R. Chapman. The C-terminal repeating units

of CsgB direct bacterial functional amyloid nucleation. *Journal of molecular biology*, 422(3): 376–389, Sept. 2012. ISSN 0022-2836. doi: 10.1016/j.jmb.2012.05.043.

E. Hellstrand, B. Boland, D. M. Walsh, and S. Linse. Amyloid β-Protein Aggregation Produces Highly Reproducible Kinetic Data and Occurs by a Two-Phase Process. *ACS Chemical Neuroscience*, 1(1):13–18, Oct. 2009. ISSN 1948-7193. doi: 10.1021/cn900015v.

X. Hu, S. L. Crick, G. Bu, C. Frieden, R. V. Pappu, and J.-M. Lee. Amyloid seeds formed by cellular uptake, concentration, and aggregation of the amyloid-beta peptide. *Proceedings of the National Academy of Sciences*, 106(48):20324–20329, Dec. 2009. doi: 10.1073/pnas.0911281106.

L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, and J. F. Banfield. A New View of the Tree of Life. *Nature Microbiology*, 1:16048, Apr. 2016. ISSN 2058-5276. doi: 10.1038/nmicrobiol.2016.48.

R. L. Hull, G. T. Westermark, P. Westermark, and S. E. Kahn. Islet Amyloid: A Critical Entity in the Pathogenesis of Type 2 Diabetes. *The Journal of Clinical Endocrinology & Metabolism*, 89(8):3629–3643, Aug. 2004. ISSN 0021-972X. doi: 10.1210/jc.2004-0405.

L. Huo, H. Zhang, X. Huo, Y. Yang, X. Li, and Y. Yin. pHMM-tree: Phylogeny of profile hidden Markov models. *Bioinformatics*, 33(7):1093–1095, Apr. 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw779.

M. G. Iadanza, M. P. Jackson, E. W. Hewitt, N. A. Ranson, and S. E. Radford. A new era for understanding amyloid structures and disease. *Nature Reviews Molecular Cell Biology*, 19 (12):755–773, Dec. 2018. ISSN 1471-0072, 1471-0080. doi: 10.1038/s41580-018-0060-8.

H. Inouye, P. E. Fraser, and D. A. Kirschner. Structure of beta-crystallite assemblies formed by Alzheimer beta-amyloid protein analogues: Analysis by x-ray diffraction. *Biophysical Journal*, 64(2):502–519, Feb. 1993. ISSN 0006-3495. doi: 10.1016/S0006-3495(93)81393-6.

J. J. Roa, G. Oncins, J. Diaz, F. Sanz, and M. Segarra. Calculation of Young's Modulus Value

by Means of AFM. *Recent Patents on Nanotechnology*, 5(1):27–36, Jan. 2011. ISSN 18722105. doi: 10.2174/187221011794474985.

T. R. Jahn and S. E. Radford. Folding versus aggregation: Polypeptide conformations on competing pathways. *Archives of Biochemistry and Biophysics*, 469(1):100–117, Jan. 2008. ISSN 0003-9861. doi: 10.1016/j.abb.2007.05.015.

M. Jamal, U. Tasneem, T. Hussain, and S. Andleeb. Bacterial Biofilm: Its Composition, Formation and Role in Human Infections. *Research & Reviews: Journal of Microbiology and Biotechnology*, 4(3), July 2015. ISSN E- 2320 - 3528 brP- 2347 - 2286.

P. F. Jensen and K. D. Rand. Hydrogen Exchange. In *Hydrogen Exchange Mass Spectrometry of Proteins*, chapter 1, pages 1–17. John Wiley & Sons, Ltd, 2016. ISBN 978-1-118-70374-8. doi: 10.1002/9781118703748.ch1.

R. Jia, C. Martens, M. Shekhar, S. Pant, G. A. Pellowe, A. M. Lau, H. E. Findlay, N. J. Harris, E. Tajkhorshid, P. J. Booth, and A. Politis. Hydrogen-deuterium exchange mass spectrometry captures distinct dynamics upon substrate and inhibitor binding to a transporter. *Nature Communications*, 11(1):6162, Dec. 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-20032-3.

J. L. Jiménez, E. J. Nettleton, M. Bouchard, C. V. Robinson, C. M. Dobson, and H. R. Saibil. The protofilament structure of insulin amyloid fibrils. *Proceedings of the National Academy of Sciences*, 99(14):9196–9201, July 2002. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.142459399.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug. 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2.

S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6): 587–589, June 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4285.

K. Katoh and D. M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780, Apr. 2013. ISSN 1537-1719. doi: 10.1093/molbev/mst010.

S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa. AAindex: Amino Acid Index Database, Progress Report 2008. *Nucleic acids research*, 36 (Database issue):D202–5, 2008.

P. C. Ke, R. Zhou, L. C. Serpell, R. Riek, T. P. J. Knowles, H. A. Lashuel, E. Gazit, I. W. Hamley, T. P. Davis, M. Fändrich, D. E. Otzen, M. R. Chapman, C. M. Dobson, D. S. Eisenberg, and R. Mezzenga. Half a century of amyloids: Past, present and future. *Chemical Society Reviews*, 49(15):5473–5509, Aug. 2020. ISSN 1460-4744. doi: 10.1039/C9CS00199A.

L. Keresztes, E. Szögi, B. Varga, V. Farkas, A. Perczel, and V. Grolmusz. The Budapest Amyloid Predictor and Its Applications. *Biomolecules*, 11(4):500, Apr. 2021. doi: 10.3390/biom11040500.

H. E. Klock and S. A. Lesley. The Polymerase Incomplete Primer Extension (PIPE) method applied to high-throughput cloning and site-directed mutagenesis. *Methods in Molecular Biology (Clifton, N.J.)*, 498:91–103, 2009. ISSN 1064-3745. doi: 10.1007/978-1-59745-196-3_6.

M. F. Knauer, B. Soreghan, D. Burdick, J. Kosmoski, and C. G. Glabe. Intracellular accumulation and resistance to degradation of the Alzheimer amyloid A4/beta protein. *Proceedings of the National Academy of Sciences*, 89(16):7437–7441, Aug. 1992. doi: 10.1073/pnas.89.16.7437.

T. P. J. Knowles, M. Vendruscolo, and C. M. Dobson. The Amyloid State and Its Association with Protein Misfolding Diseases. *Nature Reviews. Molecular Cell Biology*, 15(6):384–396, June 2014. ISSN 1471-0080. doi: 10.1038/nrm3810.

R. Kodali, A. D. Williams, S. Chemuru, and R. Wetzel. Abeta(1-40) forms five distinct amyloid structures whose beta-sheet contents and fibril stabilities are correlated. *Journal of Molecular Biology*, 401(3):503–517, Aug. 2010. ISSN 1089-8638. doi: 10.1016/j.jmb.2010.06.023.

M. Koliński, R. Dec, and W. Dzwolak. Multiscale Modeling of Amyloid Fibrils Formed by Aggregating Peptides Derived from the Amyloidogenic Fragment of the A-Chain of Insulin. *International Journal of Molecular Sciences*, 22(22):12325, Jan. 2021. ISSN 1422-0067. doi: 10.3390/ijms222212325.

L. Konermann, J. Pan, and Y.-H. Liu. Hydrogen Exchange Mass Spectrometry for Studying Protein Structure and Dynamics. *Chemical Society Reviews*, 40(3):1224–1234, Feb. 2011. ISSN 1460-4744. doi: 10.1039/C0CS00113A.

M. R. H. Krebs, E. H. C. Bromley, and A. M. Donald. The binding of thioflavin-T to amyloid fibrils: Localisation and implications. *Journal of Structural Biology*, 149(1):30–37, Jan. 2005. ISSN 1047-8477. doi: 10.1016/j.jsb.2004.08.002.

A. Kulandaisamy, V. Lathi, K. ViswaPoorani, K. Yugandhar, and M. M. Gromiha. Important amino acid residues involved in folding and binding of protein–protein complexes. *International Journal of Biological Macromolecules*, 94:438–444, Jan. 2017. ISSN 0141-8130. doi: 10.1016/j.ijbiomac.2016.10.045.

M. Kurcinski, M. Pawel Ciemny, T. Oleniecki, A. Kuriata, A. E. Badaczewska-Dawid, A. Kolinski, and S. Kmiecik. CABS-dock standalone: A toolbox for flexible protein–peptide docking. *Bioinformatics*, 35(20):4170–4172, Oct. 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz185.

A. Kuriata, V. Iglesias, J. Pujols, M. Kurcinski, S. Kmiecik, and S. Ventura. Aggrescan3D (A3D) 2.0: Prediction and engineering of protein solubility. *Nucleic Acids Research*, 47(W1): W300–W307, July 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz321.

G. Lamour, R. Nassar, P. H. W. Chan, G. Bozkurt, J. Li, J. M. Bui, C. K. Yip, T. Mayor, H. Li, H. Wu, and J. A. Gsponer. Mapping the Broad Structural and Mechanical Properties

of Amyloid Fibrils. *Biophysical Journal*, 112(4):584–594, Feb. 2017. ISSN 0006-3495. doi: 10.1016/j.bpj.2016.12.036.

S. Lê, J. Josse, and F. Husson. **FactoMineR** : An *R* Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1), 2008. ISSN 1548-7660. doi: 10.18637/jss.v025.i01.

G. Legname. Elucidating the function of the prion protein. *PLoS Pathogens*, 13(8):e1006458, Aug. 2017. ISSN 1553-7366. doi: 10.1371/journal.ppat.1006458.

H. LeVine. Thioflavine T interaction with synthetic Alzheimer's disease beta-amyloid peptides: Detection of amyloid aggregation in solution. *Protein Science : A Publication of the Protein Society*, 2(3):404–410, Mar. 1993. ISSN 0961-8368.

M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–698, Feb. 1975. ISSN 1476-4687. doi: 10.1038/253694a0.

J. Li, T. McQuade, A. B. Siemer, J. Napetschnig, K. Moriwaki, Y.-S. Hsiao, E. Damko, D. Moquin, T. Walz, A. McDermott, F. Ka-Ming Chan, and H. Wu. The RIP1/RIP3 Necrosome Forms a Functional Amyloid Signaling Complex Required for Programmed Necrosis. *Cell*, 150(2):339–350, July 2012. ISSN 0092-8674. doi: 10.1016/j.cell.2012.06.019.

D. J. Lindberg, A. Wenger, E. Sundin, E. Wesén, F. Westerlund, and E. K. Esbjörner. Binding of Thioflavin-T to Amyloid Fibrils Leads to Fluorescence Self-Quenching and Fibril Compaction. *Biochemistry*, 56(16):2170–2174, 2017. ISSN 1520-4995. doi: 10.1021/acs.biochem.7b00035.

Y. Liu, T. Liu, T. Lei, D. Zhang, S. Du, L. Girani, D. Qi, C. Lin, R. Tong, and Y. Wang. RIP1/RIP3-regulated necroptosis as a target for multifaceted disease therapy (Review). *International Journal of Molecular Medicine*, 44(3):771–786, Sept. 2019. ISSN 1107-3756. doi: 10.3892/ijmm.2019.4244.

Y. Liu, I. Sokolov, M. E. Dokukin, Y. Xiong, and P. Peng. Can AFM be used to measure absolute values of Young's modulus of nanocomposite materials down to the nanoscale? *Nanoscale*, 12(23):12432–12443, 2020. ISSN 2040-3364, 2040-3372. doi: 10.1039/D0NR02314K.

M. Lopez de la Paz and L. Serrano. Sequence determinants of amyloid fibril formation. *Proceedings of the National Academy of Sciences*, 101(1):87–92, Jan. 2004. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2634884100.

N. Louros, K. Konstantoulea, M. De Vleeschouwer, M. Ramakers, J. Schymkowitz, and F. Rousseau. WALTZ-DB 2.0: An updated database containing structural information of experimentally determined amyloid-forming peptides. *Nucleic Acids Research*, 48(D1):D389–D393, Jan. 2020a. ISSN 0305-1048. doi: 10.1093/nar/gkz758.

N. Louros, G. Orlando, M. De Vleeschouwer, F. Rousseau, and J. Schymkowitz. Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities. *Nature Communications*, 11(1):3314, July 2020b. ISSN 2041-1723. doi: 10.1038/s41467-020-17207-3.

A. S. Lyon, W. B. Peeples, and M. K. Rosen. A framework for understanding the functions of biomolecular condensates across scales. *Nature Reviews Molecular Cell Biology*, 22(3): 215–235, Mar. 2021. ISSN 1471-0080. doi: 10.1038/s41580-020-00303-z.

M. Madera. Profile Comparer: A program for scoring and aligning profile hidden Markov models. *Bioinformatics (Oxford, England)*, 24(22):2630–2631, Nov. 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn504.

S. K. Maji, M. H. Perrin, M. R. Sawaya, S. Jessberger, K. Vadodaria, R. A. Rissman, P. S. Singru, K. P. R. Nilsson, R. Simon, D. Schubert, D. Eisenberg, J. Rivier, P. Sawchenko, W. Vale, and R. Riek. Functional Amyloids as Natural Storage of Peptide Hormones in Pituitary Secretory Granules. *Science (New York, N.Y.)*, 325(5938):328–332, July 2009. ISSN 0036-8075. doi: 10.1126/science.1173155.

O. S. Makin and L. C. Serpell. Structures for amyloid fibrils. *The FEBS Journal*, 272(23): 5950–5961, 2005. ISSN 1742-4658. doi: 10.1111/j.1742-4658.2005.05025.x.

O. S. Makin, E. Atkins, P. Sikorski, J. Johansson, and L. C. Serpell. Molecular basis for amyloid fibril formation and stability. *Proceedings of the National Academy of Sciences of the*

United States of America, 102(2):315–320, Jan. 2005. ISSN 0027-8424. doi: 10.1073/pnas.0406847102.

M. Malmberg, T. Malm, O. Gustafsson, A. Sturchio, C. Graff, A. J. Espay, A. P. Wright, S. El Andaloussi, A. Lindén, and K. Ezzat. Disentangling the Amyloid Pathways: A Mechanistic Approach to Etiology. *Frontiers in Neuroscience*, 14:256, Apr. 2020. ISSN 1662-4548. doi: 10.3389/fnins.2020.00256.

K. G. Malmos, L. M. Blancas-Mejia, B. Weber, J. Buchner, M. Ramirez-Alvarado, H. Naiki, and D. Otzen. ThT 101: A primer on the use of thioflavin T to investigate amyloid formation. *Amyloid-journal of Protein Folding Disorders*, 24(1):1–16, Jan. 2017a. ISSN 1350-6129. doi: 10.1080/13506129.2017.1304905.

K. G. Malmos, L. M. Blancas-Mejia, B. Weber, J. Buchner, M. Ramirez-Alvarado, H. Naiki, and D. Otzen. ThT 101: A Primer on the Use of Thioflavin T to Investigate Amyloid Formation. *Amyloid : the international journal of experimental and clinical investigation : the official journal of the International Society of Amyloidosis*, 24(1):1–16, Jan. 2017b. ISSN 1350-6129. doi: 10.1080/13506129.2017.1304905.

A. Marchler-Bauer, C. Zheng, F. Chitsaz, M. K. Derbyshire, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, F. Lu, S. Lu, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, and S. H. Bryant. CDD: Conserved domains and protein three-dimensional structure. *Nucleic Acids Research*, 41(D1):D348–D352, Jan. 2013. ISSN 0305-1048. doi: 10.1093/nar/gks1243.

C. Martens, M. Shekhar, A. J. Borysik, A. M. Lau, E. Reading, E. Tajkhorshid, P. J. Booth, and A. Politis. Direct protein-lipid interactions shape the conformational landscape of secondary transporters. *Nature Communications*, 9(1):4151, Oct. 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-06704-1.

P. M. Martins, S. Navarro, A. Silva, M. F. Pinto, Z. Sárkány, F. Figueiredo, P. J. B. Pereira, F. Pinheiro, Z. Bednarikova, M. Burdukiewicz, O. V. Galzitskaya, Z. Gazova, C. M. Gomes, A. Pastore, L. C. Serpell, R. Skrabana, V. Smirnovas, M. Ziaunys, D. E. Otzen, S. Ventura,

and S. Macedo-Ribeiro. MIRRAGGE - Minimum Information Required for Reproducible AGGregation Experiments. *Frontiers in Molecular Neuroscience*, 13:582488, 2020. ISSN 1662-5099. doi: 10.3389/fnmol.2020.582488.

C. P. J. Maury. The emerging concept of functional amyloid. *Journal of Internal Medicine*, 265 (3):329–334, 2009. ISSN 1365-2796. doi: 10.1111/j.1365-2796.2008.02068.x.

J. Meinhardt, C. Sachse, P. Hortschansky, N. Grigorieff, and M. Fändrich. Abeta(1-40) fibril polymorphism implies diverse interaction patterns in amyloid fibrils. *Journal of Molecular Biology*, 386(3):869–877, Feb. 2009. ISSN 1089-8638. doi: 10.1016/j.jmb.2008.11.005.

G. Meisl, X. Yang, E. Hellstrand, B. Frohm, J. B. Kirkegaard, S. I. A. Cohen, C. M. Dobson, S. Linse, and T. P. J. Knowles. Differences in nucleation behavior underlie the contrasting aggregation kinetics of the Aβ40 and Aβ42 peptides. *Proceedings of the National Academy of Sciences of the United States of America*, 111(26):9384–9389, July 2014. ISSN 1091-6490. doi: 10.1073/pnas.1401564111.

B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, May 2020. ISSN 0737-4038. doi: 10.1093/molbev/msaa015.

V. J. Morris, A. R. Kirby, and A. P. Gunning. *Atomic Force Microscopy for Biologists*. IMPERIAL COLLEGE PRESS, second edition, Aug. 2009. ISBN 978-1-84816-467-3 978-1-84816-468-0. doi: 10.1142/p674.

L. R. Murphy, A. Wallqvist, and R. M. Levy. Simplified Amino Acid Alphabets for Protein Fold Recognition and Implications for Folding. *Protein Engineering*, 13(3):149–152, Mar. 2000. ISSN 0269-2139.

H. Naiki, K. Higuchi, M. Hosokawa, and T. Takeda. Fluorometric determination of amyloid fibrils in vitro using the fluorescent dye, thioflavine T. *Analytical Biochemistry*, 177(2):244–249, Mar. 1989. ISSN 0003-2697. doi: 10.1016/0003-2697(89)90046-8.

K. Naka. Monomers, Oligomers, Polymers, and Macromolecules (Overview). In S. Kobayashi and K. Müllen, editors, *Encyclopedia of Polymeric Nanomaterials*, pages 1–6. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-642-36199-9. doi: 10.1007/978-3-642-36199-9_237-1.

P. Narayan, A. Orte, R. W. Clarke, B. Bolognesi, S. Hook, K. A. Ganzinger, S. Meehan, M. R. Wilson, C. M. Dobson, and D. Klenerman. The extracellular chaperone clusterin sequesters oligomeric forms of the amyloid-B1-40 peptide. *Nature Structural & Molecular Biology*, 19 (1):79–83, Jan. 2012. ISSN 1545-9985. doi: 10.1038/nsmb.2191.

D. Nečas and P. Klapetek. Gwyddion: An open-source software for SPM data analysis. *Open Physics*, 10(1):181–188, Feb. 2012. ISSN 2391-5471. doi: 10.2478/s11534-011-0096-2.

J. Oh, J.-G. Kim, E. Jeon, C.-H. Yoo, J. S. Moon, S. Rhee, and I. Hwang. Amyloidogenesis of Type III-dependent Harpins from Plant Pathogenic Bacteria *. *Journal of Biological Chemistry*, 282(18):13601–13609, May 2007. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M602576200.

J. Pan, D. J. Wilson, and L. Konermann. Pulsed Hydrogen Exchange and Electrospray Charge-State Distribution as Complementary Probes of Protein Structure in Kinetic Experiments: Implications for Ubiquitin Folding. *Biochemistry*, 44(24):8627–8633, June 2005. ISSN 0006-2960. doi: 10.1021/bi050575e.

S. Pawlicki, A. Le Béchec, and C. Delamarche. AMYPdb: A database dedicated to amyloid precursor proteins. *BMC Bioinformatics*, 9(1):273, June 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-273.

W. R. Pearson, T. Wood, Z. Zhang, and W. Miller. Comparison of DNA Sequences with Protein Sequences. *Genomics*, 46(1):24–36, Nov. 1997. ISSN 0888-7543. doi: 10.1006/geno.1997.4995.

S. Perov, O. Lidor, N. Salinas, N. Golan, E. Tayeb- Fligelman, M. Deshmukh, D. Willbold, and M. Landau. Structural Insights into Curli CsgA Cross-β Fibril Architecture Inspire Repurposing of Anti-Amyloid Compounds as Anti-Biofilm Agents. *PLoS Pathogens*, 15(8), Aug. 2019. ISSN 1553-7366. doi: 10.1371/journal.ppat.1007978.

R. A. Peterson. The R Journal: Finding Optimal Normalizing Transformations via bestNormalize. *The R Journal*, 13(1):294–313, June 2021. ISSN 2073-4859. doi: 10.32614/RJ-2021-041.

A. T. Petkova, R. D. Leapman, Z. Guo, W.-M. Yau, M. P. Mattson, and R. Tycko. Self-Propagating, Molecular-Level Polymorphism in Alzheimer's ß-Amyloid Fibrils. *Science*, 307 (5707):262–265, Jan. 2005. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1105850.

C. M. Pfefferkorn, R. P. McGlinchey, and J. C. Lee. Effects of pH on aggregation kinetics of the repeat domain of a functional amyloid, Pmel17. *Proceedings of the National Academy of Sciences*, 107(50):21447–21452, Dec. 2010. doi: 10.1073/pnas.1006424107.

F. Pinheiro, J. Santos, and S. Ventura. AlphaFold and the amyloid landscape. *Journal of Molecular Biology*, 433(20):167059, Oct. 2021. ISSN 0022-2836. doi: 10.1016/j.jmb.2021.167059.

S. Prusiner. Molecular biology and pathogenesis of prion diseases. *Trends in Biochemical Sciences*, 21(12):482–487, Dec. 1996. ISSN 09680004. doi: 10.1016/S0968-0004(96)10063-3.

J. Pujols, S. Peña-Díaz, and S. Ventura. AGGRESCAN3D: Toward the Prediction of the Aggregation Propensities of Protein Structures. In M. Gore and U. B. Jagtap, editors, *Computational Drug Discovery and Design*, Methods in Molecular Biology, pages 427–443. Springer, New York, NY, 2018. ISBN 978-1-4939-7756-7. doi: 10.1007/978-1-4939-7756-7_21.

D. Puzzo, L. Privitera, M. Fa', A. Staniszewski, G. Hashimoto, F. Aziz, M. Sakurai, E. M. Ribe, C. M. Troy, M. Mercken, S. S. Jung, A. Palmeri, and O. Arancio. Endogenous amyloid-β is necessary for hippocampal synaptic plasticity and memory. *Annals of Neurology*, 69(5): 819–830, May 2011. ISSN 1531-8249. doi: 10.1002/ana.22313.

A. Rambaut and A. Drummond. FigTree version 1.4. 0. 2012.

P. Rawat, R. Prabakaran, R. Sakthivel, A. Mary Thangakani, S. Kumar, and M. M. Gromiha. CPAD 2.0: A repository of curated experimental data on aggregating proteins and peptides. *Amyloid*, 27(2):128–133, Apr. 2020. ISSN 1350-6129. doi: 10.1080/13506129.2020.1715363.

B. Ren, Y. Zhang, M. Zhang, Y. Liu, D. Zhang, X. Gong, Z. Feng, J. Tang, Y. Chang, and J. Zheng. Fundamentals of cross-seeding of amyloid proteins: An introduction. *Journal of Materials Chemistry B*, 7(46):7267–7282, Nov. 2019. ISSN 2050-7518. doi: 10.1039/C9TB01871A.

P. Rice, I. Longden, and A. Bleasby. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in genetics: TIG*, 16(6):276–277, June 2000. ISSN 0168-9525. doi: 10.1016/s0168-9525(00)02024-2.

R. Riek and D. S. Eisenberg. The activities of amyloids from a structural perspective. *Nature*, 539(7628):227–235, Nov. 2016. ISSN 1476-4687. doi: 10.1038/nature20416.

D. Romero and R. Kolter. Functional Amyloids in Bacteria. *International Microbiology: The Official Journal of the Spanish Society for Microbiology*, 17(2):65–73, June 2014. ISSN 1139-6709. doi: 10.2436/20.1501.01.208.

J. J. Rosa and F. M. Richards. An experimental procedure for increasing the structural resolution of chemical hydrogen-exchange measurements on proteins: Application to ribonuclease S peptide. *Journal of Molecular Biology*, 133(3):399–416, Sept. 1979. ISSN 0022-2836. doi: 10.1016/0022-2836(79)90400-5.

A. N. Round and M. J. Miles. Exploring the Consequences of Attractive and Repulsive Interaction Regimes in Tapping Mode Atomic Force Microscopy of DNA. *Nanotechnology*, 15(4):S176, 2004. ISSN 0957-4484. doi: 10.1088/0957-4484/15/4/011.

RStudio Team. *RStudio: Integrated Development Environment for r*. RStudio, PBC., Boston, MA, 2020.

A. Sakalauskas, M. Ziaunys, and V. Smirnovas. Concentration-dependent polymorphism of insulin amyloid fibrils. *PeerJ*, 7:e8208, Dec. 2019. ISSN 2167-8359. doi: 10.7717/peerj.8208.

K. Sankar, S. R. Krystek Jr, S. M. Carl, T. Day, and J. K. X. Maier. AggScore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches. *Proteins: Structure, Function, and Bioinformatics*, 86(11):1147–1156, 2018. ISSN 1097-0134. doi: 10.1002/prot.25594.

J. Santos and S. Ventura. Functional Amyloids Germinate in Plants. *Trends in Plant Science*, 26(1):7–10, Jan. 2021. ISSN 1360-1385. doi: 10.1016/j.tplants.2020.10.001.

J. Santos, V. Iglesias, and S. Ventura. Computational prediction and redesign of aberrant protein oligomerization. In *Progress in Molecular Biology and Translational Science*, volume 169, pages 43–83. Elsevier, 2020. ISBN 978-0-12-817929-1. doi: 10.1016/bs.pmbts.2019.11.002.

N. C. Santos and M. A. R. B. Castanho. An overview of the biophysical applications of atomic force microscopy. *Biophysical Chemistry*, 107(2):133–149, Feb. 2004. ISSN 0301-4622. doi: 10.1016/j.bpc.2003.09.001.

T. Scheibel, R. Parthasarathy, G. Sawicki, X.-M. Lin, H. Jaeger, and S. L. Lindquist. Conducting nanowires built by controlled self-assembly of amyloid fibers and selective metal deposition. *Proceedings of the National Academy of Sciences*, 100(8):4527–4532, Apr. 2003. doi: 10.1073/pnas.0431081100.

A. Schmidt, K. Annamalai, M. Schmidt, N. Grigorieff, and M. Fändrich. Cryo-EM reveals the steric zipper structure of a light chain-derived amyloid fibril. *Proceedings of the National Academy of Sciences of the United States of America*, 113(22):6200–6205, May 2016. ISSN 1091-6490. doi: 10.1073/pnas.1522282113.

G. Schneider and P. Wrede. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: De novo design of an idealized leader peptidase cleavage site. *Biophysical Journal*, 66(2 Pt 1):335–344, Feb. 1994. ISSN 0006-3495.

K. Schwartz and B. R. Boles. Microbial Amyloids–functions and interactions within the host. *Current opinion in microbiology*, 16(1):93–99, Feb. 2013. ISSN 1369-5274. doi: 10.1016/j.mib.2012.12.001.

A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, Jan. 2020. ISSN 1476-4687. doi: 10.1038/s41586-019-1923-7.

S. L. Shammas, G. A. Garcia, S. Kumar, M. Kjaergaard, M. H. Horrocks, N. Shivji, E. Mandelkow, T. P. Knowles, E. Mandelkow, and D. Klenerman. A Mechanistic Model of Tau Amyloid Aggregation Based on Direct Observation of Oligomers. *Nature Communications*, 6, Apr. 2015. ISSN 2041-1723. doi: 10.1038/ncomms8025.

J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang. Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, 104(11):4337–4341, Mar. 2007. doi: 10.1073/pnas.0607879104.

Q. Shu, S. L. Crick, J. S. Pinkner, B. Ford, S. J. Hultgren, and C. Frieden. The E. coli CsgB nucleator of curli assembles to β-sheet oligomers that alter the CsgA fibrillization mechanism. *Proceedings of the National Academy of Sciences*, 109(17):6502–6507, Apr. 2012. doi: 10.1073/pnas.1204161109.

L. P. Silva. Imaging proteins with atomic force microscopy: An overview. *Current Protein & Peptide Science*, 6(4):387–395, Aug. 2005. ISSN 1389-2037. doi: 10.2174/1389203054546389.

R. Simm, I. Ahmad, M. Rhen, S. Le Guyon, and U. Römling. Regulation of Biofilm Formation in Salmonella Enterica Serovar Typhimurium. *Future Microbiology*, 9(11):1261–1282, 2014. ISSN 1746-0921. doi: 10.2217/fmb.14.88.

J. D. Sipe, M. D. Benson, J. N. Buxbaum, S.-I. Ikeda, G. Merlini, M. J. M. Saraiva, and P. Westermark. Amyloid Fibril Proteins and Amyloidosis: Chemical Identification and Clinical Classification International Society of Amyloidosis 2016 Nomenclature Guidelines. *Amyloid: The International Journal of Experimental and Clinical Investigation: The Official Journal of the International Society of Amyloidosis*, 23(4):209–213, Dec. 2016a. ISSN 1744-2818. doi: 10.1080/13506129.2016.1257986.

J. D. Sipe, M. D. Benson, J. N. Buxbaum, S.-i. Ikeda, G. Merlini, M. J. M. Saraiva, and P. Westermark. Amyloid fibril proteins and amyloidosis: Chemical identification and clinical classification International Society of Amyloidosis 2016 Nomenclature Guidelines. *Amyloid*, 23(4):209–213, Oct. 2016b. ISSN 1350-6129. doi: 10.1080/13506129.2016.1257986.

M. Sleutel, I. Van den Broeck, N. Van Gerven, C. Feuillie, W. Jonckheere, C. Valotteau, Y. F. Dufrêne, and H. Remaut. Nucleation and Growth of a Bacterial Functional Amyloid at Single-Fiber Resolution. *Nature Chemical Biology*, 13(8):902–908, Aug. 2017. ISSN 1552-4469. doi: 10.1038/nchembio.2413.

J. F. Smith, T. P. J. Knowles, C. M. Dobson, C. E. MacPhee, and M. E. Welland. Characterization of the nanoscale properties of individual amyloid fibrils. *Proceedings of the National Academy of Sciences*, 103(43):15806–15811, Oct. 2006. doi: 10.1073/pnas.0604035103.

T. Šneideris, L. Baranauskienė, J. G. Cannon, R. Rutkienė, R. Meškys, and V. Smirnovas. Looking for a Generic Inhibitor of Amyloid-like Fibril Formation among Flavone Derivatives. *PeerJ*, 3:e1271, Sept. 2015. ISSN 2167-8359. doi: 10.7717/peerj.1271.

A. Srivastava, P. K. Singh, M. Kumbhakar, T. Mukherjee, S. Chattopadyay, H. Pal, and S. Nath. Identifying the Bond Responsible for the Fluorescence Modulation in an Amyloid Fibril Sensor. *Chemistry – A European Journal*, 16(30):9257–9263, 2010. ISSN 1521-3765. doi: 10.1002/chem.200902968.

A. I. Sulatskaya, E. A. Volova, Y. Y. Komissarchik, E. S. Snigirevskaya, A. A. Maskevich, E. A. Drobchenko, I. M. Kuznetsova, and K. K. Turoverov. Investigation of the kinetics of insulin amyloid fibrils formation. *Cell and Tissue Biology*, 8(2):186–191, Mar. 2014. ISSN 1990-5203. doi: 10.1134/S1990519X14020114.

D. L. Swofford. Laboratory of Molecular Systematics Smithsonian Institution. 1998.

N. Szulc, M. Burdukiewicz, M. Gąsior-Głogowska, J. W. Wojciechowski, J. Chilimoniuk, P. Mackiewicz, T. Šneideris, V. Smirnovas, and M. Kotulska. Bioinformatics methods for identification of amyloidogenic peptides show robustness to misannotated training data. *Scientific Reports*, 11(1):8934, Apr. 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-86530-6.

K. Tamura, G. Stecher, and S. Kumar. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, 38(7):3022–3027, July 2021. ISSN 1537-1719. doi: 10.1093/molbev/msab120.

G. G. Tartaglia and M. Vendruscolo. The Zyggregator method for predicting protein aggregation propensities. *Chemical Society Reviews*, 37(7):1395–1401, June 2008. ISSN 1460-4744. doi: 10.1039/B706784B.

G. G. Tartaglia, A. P. Pawar, S. Campioni, C. M. Dobson, F. Chiti, and M. Vendruscolo. Prediction of Aggregation-Prone Regions in Structured Proteins. *Journal of Molecular Biology*, 380(2):425–436, July 2008. ISSN 0022-2836. doi: 10.1016/j.jmb.2008.05.013.

R. C. Team, R. C. Team, et al. R: A language and environment for statistical computing. R foundation for statistical computing, vienna, austria, 4.0. 5, 2021.

F. Teufel, J. J. Almagro Armenteros, A. R. Johansen, M. H. Gíslason, S. I. Pihl, K. D. Tsirigos, O. Winther, S. Brunak, G. von Heijne, and H. Nielsen. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, 40(7):1023–1025, July 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01156-3.

J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, Nov. 1994. ISSN 0305-1048. doi: 10.1093/nar/22.22.4673.

P. Tompa. Intrinsically disordered proteins: A 10-year recap. *Trends in Biochemical Sciences*, 37(12):509–516, Dec. 2012. ISSN 0968-0004. doi: 10.1016/j.tibs.2012.08.004.

M. Törnquist, T. C. T. Michaels, K. Sanagavarapu, X. Yang, G. Meisl, S. I. A. Cohen, T. P. J. Knowles, and S. Linse. Secondary nucleation in amyloid formation. *Chemical Communications*, 54(63):8667–8684, Aug. 2018. ISSN 1364-548X. doi: 10.1039/C8CC02204F.

B. H. Toyama and J. S. Weissman. Amyloid Structure: Conformational Diversity and Consequences. *Annual review of biochemistry*, 80:10.1146/annurev–biochem–090908–120656, 2011. ISSN 0066-4154. doi: 10.1146/annurev-biochem-090908-120656.

J. Tran, D. Chang, F. Hsu, H. Wang, and Z. Guo. Cross-seeding between Aβ40 and Aβ42 in Alzheimer's disease. *FEBS Letters*, 591(1):177–185, 2017. ISSN 1873-3468. doi: 10.1002/1873-3468.12526.

B. Turcq, C. Deleu, M. Denayrolles, and J. Bégueret. Two allelic genes responsible for vegetative incompatibility in the fungus Podospora anserina are not essential for cell viability. *Molecular & general genetics*, 228(1-2):265–269, Aug. 1991. ISSN 0026-8925. doi: 10.1007/bf00282475.

T. UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 46(5):2699–2699, Mar. 2018. ISSN 0305-1048. doi: 10.1093/nar/gky092.

O. Vadas and J. E. Burke. Probing the dynamic regulation of peripheral membrane proteins using hydrogen deuterium exchange-MS (HDX-MS). *Biochemical Society Transactions*, 43 (5):773–786, Oct. 2015. ISSN 1470-8752. doi: 10.1042/BST20150065.

N. Van Gerven, S. E. Van der Verren, D. M. Reiter, and H. Remaut. The Role of Functional Amyloids in Bacterial Virulence. *Journal of Molecular Biology*, 430(20):3657–3684, Oct. 2018. ISSN 0022-2836. doi: 10.1016/j.jmb.2018.07.010.

K. Van Roey, B. Uyar, R. J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T. J. Gibson, and N. E. Davey. Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation. *Chemical Reviews*, 114(13):6733–6778, July 2014. ISSN 0009-2665. doi: 10.1021/cr400585q.

M. Varadi, G. De Baets, W. F. Vranken, P. Tompa, and R. Pancsa. AmyPro: A database of proteins with validated amyloidogenic regions. *Nucleic Acids Research*, 46(D1):D387–D392, Jan. 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx950.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Statistics and Computing. Springer, New York, NY, 2002. ISBN 978-1-4419-3008-8 978-0-387-21706-2. doi: 10.1007/978-0-387-21706-2.

M. Vendruscolo and M. Fuxreiter. Sequence Determinants of the Aggregation of Proteins Within Condensates Generated by Liquid-liquid Phase Separation. *Journal of Molecular Biology*, 434 (1):167201, Jan. 2022. ISSN 0022-2836. doi: 10.1016/j.jmb.2021.167201.

S. Ventura. Sequence determinants of protein aggregation: Tools to increase protein solubility. *Microbial Cell Factories*, 4(1):11, Apr. 2005. ISSN 1475-2859. doi: 10.1186/1475-2859-4-11.

S. Ventura, J. Zurdo, S. Narayanan, M. Parreño, R. Mangues, B. Reif, F. Chiti, E. Giannoni, C. M. Dobson, F. X. Aviles, and L. Serrano. Short amino acid stretches can mediate amyloid formation in globular proteins: The Src homology 3 (SH3) case. *Proceedings of the National Academy of Sciences*, 101(19):7258–7263, May 2004. doi: 10.1073/pnas.0308249101.

S. Veretnik, J. L. Fink, and P. E. Bourne. Computational Biology Resources Lack Persistence and Usability. *PLOS Computational Biology*, 4(7):e1000136, July 2008. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000136.

R. M. Vernon and J. D. Forman-Kay. First-generation predictors of biological protein phase separation. *Current Opinion in Structural Biology*, 58:88–96, Oct. 2019. ISSN 0959-440X. doi: 10.1016/j.sbi.2019.05.016.

R. Vidal and B. Ghetti. Characterization of Amyloid Deposits in Neurodegenerative Diseases. *Methods in Molecular Biology (Clifton, N.J.)*, 793:241–258, 2011. ISSN 1940-6029. doi: 10.1007/978-1-61779-328-8_16.

E. S. Voropai, M. P. Samtsov, K. N. Kaplevskii, A. A. Maskevich, V. I. Stepuro, O. I. Povarova, I. M. Kuznetsova, K. K. Turoverov, A. L. Fink, and V. N. Uverskii. Spectral Properties of Thioflavin T and Its Complexes with Amyloid Fibrils. *Journal of Applied Spectroscopy*, 70 (6):868–874, Nov. 2003. ISSN 1573-8647. doi: 10.1023/B:JAPS.0000016303.37573.7e.

J. M. Walker, editor. *The Proteomics Protocols Handbook*. Humana Press, 2005. ISBN 978-1-58829-343-5.

I. Walsh, F. Seno, S. C. E. Tosatto, and A. Trovato. PASTA 2.0: An Improved Server for Protein Aggregation Prediction. *Nucleic Acids Research*, 42(W1):W301–W307, July 2014. ISSN 0305-1048. doi: 10.1093/nar/gku399.

H. Wang, Q. Shu, D. L. Rempel, C. Frieden, and M. L. Gross. Continuous and Pulsed Hydrogen–Deuterium Exchange and Mass Spectrometry Characterize CsgE Oligomerization. *Biochemistry*, 54(42):6475–6481, Oct. 2015. ISSN 0006-2960, 1520-4995. doi: 10.1021/acs.biochem.5b00871.

X. Wang and M. R. Chapman. Sequence Determinants of Bacterial Amyloid Formation. *Journal of Molecular Biology*, 380(3):570–580, July 2008. ISSN 0022-2836. doi: 10.1016/j.jmb.2008.05.019.

X. Wang, N. D. Hammer, and M. R. Chapman. The Molecular Basis of Functional Bacterial Amyloid Polymerization and Nucleation. *The Journal of Biological Chemistry*, 283(31):21530–21539, Aug. 2008. ISSN 0021-9258. doi: 10.1074/jbc.M800466200.

X. Wang, Y. Zhou, J.-J. Ren, N. D. Hammer, and M. R. Chapman. Gatekeeper residues in the major curlin subunit modulate bacterial amyloid fiber biogenesis. *Proceedings of the National Academy of Sciences*, 107(1):163–168, Jan. 2010. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0908714107.

C. Wasmer, A. Lange, H. Van Melckebeke, A. B. Siemer, R. Riek, and B. H. Meier. Amyloid fibrils of the HET-s(218-289) prion form a beta solenoid with a triangular hydrophobic core. *Science (New York, N.Y.)*, 319(5869):1523–1526, Mar. 2008. ISSN 1095-9203. doi: 10.1126/science.1151839.

A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton. Jalview Version 2—a Multiple Sequence Alignment Editor and Analysis Workbench. *Bioinformatics (Oxford, England)*, 25(9):1189–1191, May 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp033.

G. A. Wells, A. C. Scott, C. T. Johnson, R. F. Gunning, R. D. Hancock, M. Jeffrey, M. Dawson, and R. Bradley. A novel progressive spongiform encephalopathy in cattle. *The Veterinary Record*, 121(18):419–420, Oct. 1987. ISSN 0042-4900. doi: 10.1136/vr.121.18.419.

T. J. Wheeler, J. Clements, and R. D. Finn. Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics*, 15(1):7, Jan. 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-7.

J. Wood, L. Lund, and S. Done. The natural occurrence of scrapie in moufflon. *The Veterinary record*, 130(2):25–27, 1992. ISSN 0042-4900. doi: 10.1136/vr.130.2.25.

P. P. Wozniak and M. Kotulska. AmyLoad: Website Dedicated to Amyloidogenic Protein

Fragments. *Bioinformatics (Oxford, England)*, 31(20):3395–3397, Oct. 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv375.

C. Wu, M. Biancalana, S. Koide, and J.-E. Shea. Binding Modes of Thioflavin-T to the Single-Layer β-Sheet of the Peptide Self-Assembly Mimics. *Journal of Molecular Biology*, 394(4): 627–633, Dec. 2009. ISSN 0022-2836. doi: 10.1016/j.jmb.2009.09.056.

N. Xiao, D.-S. Cao, M.-F. Zhu, and Q.-S. Xu. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics (Oxford, England)*, 31(11):1857–1859, June 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv042.

C. Xue, T. Y. Lin, D. Chang, and Z. Guo. Thioflavin T as an amyloid dye: Fibril quantification, optimal concentration and effect on aggregation. *Royal Society Open Science*, 2017. doi: 10.1098/rsos.160696.

G. Yu, T. T.-Y. Lam, S. Xu, L. Li, B. Jones, J. Silverman, W. M. Iwasaki, Y. Xia, and R. Huang. Ggtree: An R package for visualization of tree and annotation data. Bioconductor version: Release (3.16), 2023.

H. Zhang, L.-l. Li, F.-y. Dai, H.-h. Zhang, B. Ni, W. Zhou, X. Yang, and Y.-z. Wu. Preparation and characterization of silk fibroin as a biomaterial with potential for drug delivery. *Journal of Translational Medicine*, 10(1):117, June 2012. ISSN 1479-5876. doi: 10.1186/1479-5876-10-117.

Y. Zhang, D. L. Rempel, J. Zhang, A. K. Sharma, L. M. Mirica, and M. L. Gross. Pulsed hydrogen–deuterium exchange mass spectrometry probes conformational changes in amyloid beta (Aβ) peptide aggregation. *Proceedings of the National Academy of Sciences*, 110(36): 14604–14609, Sept. 2013. doi: 10.1073/pnas.1309175110.

Z. Zhang and D. L. Smith. Determination of amide hydrogen exchange by mass spectrometry: A new tool for protein structure elucidation. *Protein Science : A Publication of the Protein Society*, 2(4):522–531, Apr. 1993. ISSN 0961-8368.

Y. Zhou, L. P. Blanco, D. R. Smith, and M. R. Chapman. Bacterial Amyloids. In E. M. Sigurdsson, M. Calero, and M. Gasset, editors, *Amyloid Proteins: Methods and Protocols*, Methods in Molecular Biology, pages 303–320. Humana Press, Totowa, NJ, 2012a. ISBN 978-1-61779-551-0. doi: 10.1007/978-1-61779-551-0_21.

Y. Zhou, L. P. Blanco, D. R. Smith, and M. R. Chapman. Bacterial Amyloids. *Methods in Molecular Biology (Clifton, N.J.)*, 849:303–320, 2012b. ISSN 1940-6029. doi: 10.1007/978-1-61779-551-0_21.

Y. Zhou, D. Smith, B. J. Leong, K. Brännström, F. Almqvist, and M. R. Chapman. Promiscuous Cross-Seeding between Bacterial Amyloids Promotes Interspecies Biofilms. *The Journal of Biological Chemistry*, 287(42):35092–35103, Oct. 2012c. ISSN 1083-351X. doi: 10.1074/jbc.M112.383737.

Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab, and R. Knight. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nature Communications*, 10(1):5477, Dec. 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-13443-4.

12 Achievements

12.1 Grants

- Deutscher Akademischer Austauschdienst (DAAD)
  Research Grants - One-Year Grants for Doctoral Candidates, 2020/21 (57507870)
  Proteinaceous scaffolds of biofilms produced by gram-negative bacteria

12.2 Publications

- AmyloGraph: a comprehensive database of amyloid-amyloid interactions.
  Michał Burdukiewicz, Dominik Rafacz, Agnieszka Barbach, Katarzyna Hubicka, Laura Bąkała, Anna Lassota, Jakub Stecko, Natalia Szymańska, Jakub W Wojciechowski, Dominika Kozakiewicz, Natalia Szulc, Jarosław Chilimoniuk, Izabela Jęśkowiak, Marlena Gąsior-Głogowska, Małgorzata Kotulska.
  Nucleic Acids Research 2022 Oct 16;gkac882. doi: 10.1093/nar/gkac882.
  IF = 16.971, PM = 200

- Adhesion of Enteropathogenic, Enterotoxigenic, and Commensal Escherichia coli to the Major Zymogen Granule Membrane Glycoprotein 2.
  Christin Bartlitz, Rafał Kolenda, Jarosław Chilimoniuk, Krzysztof Grzymajło, Stefan Rödiger, Rolf Bauerfeind, Aamir Ali, Veronika Tchesnokova, Dirk Roggenbuck, Peter Schierack.
  Applied and Environmental Microbiology 2022 Mar 8;88(5):e0227921. doi: 10.1128/aem.02279-21.
  IF = 4.792, PM = 100

- Bioinformatics methods for identification of amyloidogenic peptides show robustness to misannotated training data.
  Natalia Szulc, Michał Burdukiewicz, Marlena Gąsior-Głogowska, Jakub W. Wojciechowski, Jarosław Chilimoniuk, Paweł Mackiewicz, Tomas Šneideris, Vytautas Smirnovas Malgorzata Kotulska.

Scientific Reports (2021) 11:8934
IF = 4.379, PM = 140

- countfitteR: efficient selection of count distributions to assess DNA damage.
  Jarosław Chilimoniuk, Alicja Gosiewska, Jadwiga Słowik, Romano Weiss, P. Markus Deckert, Stefan Rödiger, Michał Burdukiewicz.
  Annals of Translational Medicine 2021;9(7):528
  IF = 3.932, PM = 40

- Proteomic Screening for Prediction and Design of Antimicrobial Peptides with AmpGram.
  Michał Burdukiewicz, Katarzyna Sidorczuk, Dominik Rafacz, Filip Pietluch, Jarosław Chilimoniuk, Stefan Rödiger, Przemysław Gagat.
  International Journal of Molecular Sciences, 21:12, 2020. 10.3390/ijms21124310
  IF = 5.923, PM = 140

- Prediction of Signal Peptides in Proteins from Malaria Parasites.
  Burdukiewicz M., Sobczyk P., Chilimoniuk J., Gagat P., Mackiewicz P.
  International Journal of Molecular Sciences 19(12), 3709, 2018.
  IF = 5.923, PM = 140

- PhyMet2: a database and toolkit for phylogenetic and metabolic analyses of methanogens.
  Michał Burdukiewicz, Przemysław Gagat, Sławomir Jabłoński, Jarosław Chilimoniuk, Michał Gaworski, Paweł Mackiewicz, Marcin Łukaszewicz.
  Environmental microbiology reports 10(3):378-382, 2018
  IF = 3.541, PM = 100

Total IF = 45.47, Total PM = 860

12.3 Internships

- Amyloid Research Group, Institute of Biotechnology, Vilnius University. Vilnius, Lithuania.
  05.12.2022-24.12.2022

- Multiparameter Dianostic group, Institute of Biotechnology, Brandenburg University of Technology Cottbus - Senftenberg. Senftenberg, Germany.
  01.11.2021-26.12.2021

- Multiparameter Dianostic group, Institute of Biotechnology, Brandenburg University of Technology Cottbus - Senftenberg. Senftenberg, Germany.
  01.10.2020-30.09.2021

- Image Based Assays group, Institute of Biotechnology, Brandenburg University of Technology Cottbus - Senftenberg. Senftenberg, Germany.
  01.03.2020-29.05.2020

- Image Based Assays group, Institute of Biotechnology, Brandenburg University of Technology Cottbus - Senftenberg. Senftenberg, Germany.
  01.10.2019-16.12.2019

- Image Based Assays group, Institute of Biotechnology, Brandenburg University of Technology Cottbus - Senftenberg. Senftenberg, Germany.
  01.07.2019-30.11.2019 - Erasmus+

- Image Based Assays group, Institute of Biotechnology, Brandenburg University of Technology Cottbus - Senftenberg. Senftenberg, Germany.
  15.04.2019-31.05.2019

- Image Based Assays group, Institute of Biotechnology, Brandenburg University of Technology Cottbus - Senftenberg. Senftenberg, Germany.
  04.02.2019-15.02.2019

- Department of Biothermodynamics and Drug Design, Institute of Biotechnology, Vilnius University. Vilnius, Lithuania.
  02.11.2018-09.11.2018

- Multiparameter Diagnostics group, Institute of Biotechnology, Brandenburg University of Technology Cottbus - Senftenberg. Senftenberg, Germany.
  01.04-30.09.2018

- Department of Biothermodynamics and Drug Design, Institute of Biotechnology, Vilnius University. Vilnius, Lithuania.
  07.01-26.01.2018

12.4 Conference Talks

- Imputomics: Imputation of missing values for "Omics" data.
  Metabolomics Circle, Wrocław, Poland.
  Jarosław Chilimoniuk, Krystyna Grzesiak, Dominik Nowakowski, Adam Krętowski, Michał Ciborowski, and Michał Burdukiewicz
  27.01-28.01-2023

- countfitteR: count data analysis for precision medicine.
  International Biotech Innovation Days 2020 (IBID), Senftenberg, Germany.
  Jarosław Chilimoniuk, Alicja Gosiewska, Jadwiga Słowik, Romano Weiss, P. Markus Deckert, Stefan Rödiger and Michał Burdukiewicz
  28.10-29.10.2020

- Count data analysis with countfitteR.
  Why R? 2020, Warszawa, Poland.
  Jaroslaw Chilimoniuk, Alicja Gosiewska, Jadwiga Slowik, Romano Weiss, P. Markus Deckert, Stefan Rödiger, Michal Burdukiewicz
  24.09-27.09-2020

- AmyloGram: prediction of amyloid sequences in R.
  satRday, 2019, Gdańsk, Poland.
  Jarosław Chilimoniuk, Michał Burdukiewicz, Piotr Sobczyk, Stefan Rödiger, Małgorzata Kotulska and Paweł Mackiewicz.
  17.05-18.05.2019

- AmyloGram: the R package and a Shiny server for amyloid prediction.
  Why R? 2019, Warszawa, Poland.
  Jaroslaw Chilimoniuk, Michał Burdukiewicz, Piotr Sobczyk, Stefan Rödiger, Malgorzata

Kotulska and Pawel Mackiewicz.

26.09-29.09.2019

- PhyMet2: a database and toolkit for phylogenetic and metabolic analyses of methanogens.
  PTBI: Polish Bioinformatics Society, 2018, Wrocław, Poland.
  Jarosław Chilimoniuk, Michał Burdukiewicz, Przemysław Gagat, Sławomir Jabłoński, Michał Gaworski, Paweł Mackiewicz, Marcin Łukaszewicz.
  05.09-07.09.2018

12.5 Conference posters

- AmpGram – a novel tool for prediction of antimicrobial peptides.
  6th Joint Conference of the DGHM VAAM, 2020, Leipzig, Germany.
  J. Chilimoniuk, M. Burdukiewicz, K. Sidorczuk, F. Pietluch, D. Rafacz, S. Rödiger, P. Gagat. 08.03-11.03.2020

- AmyloGram: prediction of amyloid sequences in R.
  EuPA: XIII. Annual Congress of the European Proteomics Association: From Genes via Proteins and their Interactions to Functions, 2019, Potsdam, Germany.
  Jarosław Chilimoniuk, Michał Burdukiewicz, Piotr Sobczyk, Stefan Rödiger, Małgorzata Kotulska and Paweł Mackiewicz.
  24.03-28.03.2019

- Co-evolution of curli components CsgA and CsgB.
  VAAM: Jahrestagung 2019 der Vereinigung für Allgemeine und Angewandte Mikrobiologie, 2019, Mainz, Germany.
  Jaroslaw Chilimoniuk, Michał Burdukiewicz, Paweł Mackiewicz.
  17.03-20.03.2019

- AmyloGram: prediction of amyloid sequences in R.
  PL in ML: Polish View on Machine Learning, 2018, Warsaw, Poland.
  Jarosław Chilimoniuk, Michał Burdukiewicz, Piotr Sobczyk, Stefan Rödiger, Anna Duda-Madej, Marlena Gąsior-Głogowska, Małgorzata Kotulska and Paweł Mackiewicz.

14.12-17.12.2018

- CsgA and CsgC - evolutionary interplay in curli biogenesis.
  8th ASM Conference on Biofilms, 2018, Washington, DC, USA.
  Jaroslaw Chilimoniuk, Michał Burdukiewicz, Paweł Mackiewicz.
  07.10-11.10.2018

- PhyMet2: database and algorithm predicting culturing conditions of methanogens.
  IBID: International Biotech Innovation Days, 2018, Senftenberg, Germany.
  Jarosław Chilimoniuk, Michał Burdukiewicz, Przemysław Gagat, Sławomir Jabłoński, Michał Gaworski, Paweł Mackiewicz, Marcin Łukaszewicz.
  23.05-25.05.2018

- PhyMet2: complex database containing records on methanogens with unique feature (MethanoGram) allowing prediction of culture conditions based on 16S rRNA.
  VAAM: Jahrestagung 2018 der Vereinigung für Allgemeine und Angewandte Mikrobiologie, 2018, Wolfsburg, Germany.
  Michał Burdukiewicz, Przemysław Gagat, Sławomir Jabłoński, Jarosław Chilimoniuk, Michał Gaworski, Paweł Mackiewicz, Marcin Łukaszewicz.
  15.04-18.04.2018