



Wrocław University
of Science and Technology

Computational studies of amyloids and their interactions

Jakub W. Wojciechowski MSc, Engr

Supervisor: Prof. Małgorzata Kotulska

Department of Biomedical Engineering
Wrocław University of Science and Technology
Wrocław, Poland

Co-supervisor: Dr Johannes Söding

Max Planck Institute for Multidisciplinary Sciences
Gottingen, Germany

Department of Biomedical Engineering
Wrocław University of Science and Technology
Wrocław, Poland

Contents

| | | |
|----------|--|-----------|
| 1 | List of Articles | 5 |
| 1.1 | Articles that are the basis for the PhD degree | 5 |
| 1.2 | Other articles related to PhD degree | 6 |
| 1.3 | Other articles not related to PhD degree | 6 |
| 2 | Streszczenie pracy doktorskiej (Summary in Polish) | 7 |
| 2.1 | Wstęp | 7 |
| 2.2 | Wyniki | 8 |
| 2.3 | Podsumowanie | 12 |
| 3 | General Introduction | 13 |
| 3.1 | Amyloids | 14 |
| 3.2 | Amyloids and diseases | 14 |
| 3.3 | Amyloid structures | 14 |
| 3.4 | Functional Amyloids | 16 |
| 3.5 | Amyloid cross-interactions | 17 |
| 3.6 | Computational methods for studying amyloids | 18 |
| 3.7 | Amyloid databases | 19 |
| 3.8 | Aggregation-prone region predictors | 20 |
| 3.9 | Prediction of amyloidogenicity of an entire protein | 22 |
| 3.10 | Prediction of the mutation effects and aggregation kinetics | 22 |
| 3.11 | Other computational methods | 23 |
| 4 | Thesis of this work | 24 |
| 5 | Results | 25 |
| 5.1 | Path - prediction of amyloidogenicity by threading and machine learning. | 26 |
| 5.2 | Bioinformatics methods for identification of amyloidogenic peptides show robustness to misannotated training data | 28 |
| 5.3 | Variability of Amyloid Propensity in Imperfect Repeats of CsgA Protein of Salmonella enterica and Escherichia coli | 29 |
| 5.4 | A spatiotemporal reconstruction of the C. elegans pharyngeal cuticle reveals a structure rich in phase-separating proteins | 30 |
| 5.5 | Exploring a diverse world of effector domains and amyloid signaling motifs in fungal NLR proteins | 32 |
| 5.6 | AmyloGraph: a comprehensive database of amyloid–amyloid interactions | 34 |
| 5.7 | PACT - Prediction of Amyloid Cross-interaction by Threading | 35 |

5.8 Summary and Discussion 38

Acknowledgements

Powstanie tej pracy nie byłoby możliwe bez wsparcia bardzo wielu osób, którym w tym miejscu chciałbym podziękować. Przede wszystkim mojej rodzinie za cierpliwość i okazywane mi wsparcie. Mojej ukochanej Ali, za wyrozumiałość i za to że zawsze we mnie wierzyła, ale także za wiele merytorycznych dyskusji i pomysłów. Natalii Szulc i dr Marlenie Gąsior-Głogowskiej za długie godziny dyskusji i nieocenioną pomoc w sprawach eksperymentów. Dr Witoldowi Dyrce, za wprowadzenie mnie w fascynujący świat amyloidów grzybowych. Prof. Peterowi Royowi z Uniwersytetu w Toronto za otwartość i ciekawe wyzwanie jakim były badania proteomu *C. elegans*. Dr Johannesowi Soedingowi za liczne cenne rozmowy oraz uwagi, a także bardzo ciepłe przyjęcie w Getyndze. Wreszcie chciałbym podziękować Prof. Małgorzacie Kotulskiej za wieloletnią współpracę, duże pokłady cierpliwości oraz wytrwałe wspieranie mnie na drodze, której zwyciężeniem jest niniejsza praca.

Abstract

Amyloids are insoluble, fibrillar protein aggregates, best known for their role in the development of neurodegenerative disorders. However, more recent studies show that such structures can be utilized by a variety of organisms to perform physiological functions including biofilm formation, hormone storage and signaling. It was shown that both functional and pathological amyloids can interact in a number of ways. Such interactions can lead to a significant increase in aggregation rates or inhibition of fibril formation. Several computational methods have been proposed. Unfortunately, their accuracy is still limited. This is especially true in the case of functional amyloids which are severely underrepresented in amyloid databases. Furthermore, there are no tools dedicated to the prediction of amyloid cross-interactions. This relatively recently discovered phenomenon can play a pivotal role in our understanding of the comorbidity of amyloid-related disorders.

During my PhD, I extended my research on the modeling of amyloid aggregates and developed a new method for the identification of aggregation-prone regions in proteins - PATH (Prediction of Amyloidogenicity by THreading). The method combines structural modeling with machine learning. The proposed method allows for the accurate identification of amyloidogenic fragments and enables the user to infer the most probable structural class of the resulting amyloid core. Our results showed that PATH, as well as some other bioinformatics methods, is robust against misannotated training data. All tested methods have problems with the identification of functional amyloids which are severely underrepresented in available databases. To better understand functional amyloids, a detailed characterization of CsgA proteins from *Escherichia coli* and *Salmonella enterica* was performed. We also investigated amyloids in the *Caenorhabditis elegans* proteome. We scanned the whole proteome for possible amyloidogenic proteins, using PATH and AmyloiGram tools. Despite using two different methods, our analysis showed many false positives, which shows potential problems with applying computational methods on a proteome scale.

Difficulties with large-scale identification of amyloids encouraged us to try a different method in the next project. We aimed to better understand the fungal NLR system. It uses amyloid aggregation to propagate the signal between the receptor and effector protein by so-called Amyloid Signaling Motifs (ASM). Using de novo motif detection tools as well as natural language processing models we identified a new type of ASM - PUASM which stands for Pnp_Udp Amyloid Signaling Motif. We also significantly improve the annotation of NLR-related protein domains. Finally, I took part in the development of AmyloGraph database and I developed PACT - the first method for the prediction of amyloid cross-interactions. PACT not only achieved good accuracy on the task of interaction prediction but can also be used to identify novel amyloid-prone regions. The method was then used to identify which region is most likely involved in interactions of CsgA with hIAPP.

Chapter 1

List of Articles

1.1 Articles that are the basis for the PhD degree

1. **Wojciechowski, J. W.**, & Kotulska, M. (2020). Path-prediction of amyloidogenicity by threading and machine learning. *Scientific Reports*, 10(1), 1-9.
2. Szulc, N., Burdukiewicz, M., Gąsior-Głogowska, M., **Wojciechowski, J. W.**, Chilimoniuk, J., Mackiewicz, P., ... & Kotulska, M. (2021). Bioinformatics methods for identification of amyloidogenic peptides show robustness to misannotated training data. *Scientific Reports*, 11(1), 1-11.
3. Szulc, N., Gąsior-Głogowska, M., **Wojciechowski, J. W.**, Szeferczyk, M., Żak, A. M., Burdukiewicz, M., & Kotulska, M. (2021). Variability of amyloid propensity in imperfect repeats of csgA protein of salmonella enterica and escherichia coli. *International Journal of Molecular Sciences*, 22(10), 5127.
4. Kamal, M., Tokmakjian, L., Knox, J., Mastrangelo, P., Ji, J., Cai, H., **Wojciechowski, J. W.**, ... & Roy, P. J. (2022). A spatiotemporal reconstruction of the *C. elegans* pharyngeal cuticle reveals a structure rich in phase-separating proteins. *Elife*, 11, e79396.
5. **Wojciechowski, J. W.**, Tekoglu, E., Gąsior-Głogowska, M., Coustou, V., Szulc, N., Szeferczyk, M., ... & Dyrka, W. (2022). Exploring a diverse world of effector domains and amyloid signaling motifs in fungal NLR proteins. *PLOS Computational Biology*, 18(12), e1010787.
6. Burdukiewicz, M., Rafacz, D., Barbach, A., Hubicka, K., Bąkała, L., Lassota, A., Stecko, J., Szymańska, N., **Wojciechowski, J. W.**, ... & Kotulska, M. (2022). AmyloGraph: a comprehensive database of amyloid-amyloid interactions. *Nucleic Acids Research*.
7. Wojciechowski, J. W., Szczurek W., Szulc, N., Szeferczyk, M., & Kotulska, M. (2022). PACT-Prediction of Amyloid Cross-interaction by Threading. *bioRxiv*, 2022-07. (Published in *bioRxiv*, under revision in *Scientific Reports*)

1.2 Other articles related to PhD degree

Kotulska, M., & **Wojciechowski, J. W.** (2022). Bioinformatics Methods in Predicting Amyloid Propensity of Peptides and Proteins. In *Computer Simulations of Aggregation of Proteins and Peptides* (pp. 1-15). Humana, New York, NY.

1.3 Other articles not related to PhD degree

Radosinski, L., Labus, K., Zemojtel, P., & **Wojciechowski, J. W.** (2019). Development and Validation of a Virtual Gelatin Model Using Molecular Modeling Computational Tools. *Molecules*, 24(18), 3365.

Chapter 2

Streszczenie pracy doktorskiej (Summary in Polish)

2.1 Wstęp

Białka, obok lipidów oraz kwasów nukleinowych, są jednym z fundamentalnych elementów każdego żywego organizmu. Ze względu na swoją modułową budowę mogą one pełnić rozmaite role, począwszy od strukturalnych kończąc na katalizowaniu złożonych reakcji chemicznych. Źródłem tak szerokiej możliwości tej klasy makromolekuł jest ich trójwymiarowa struktura, zakodowana w sekwencji aminokwasowej. Gdy jednak z jakiegoś powodu białko zmienia lub traci swoją strukturę, wiąże się to zwykle z utratą jego prawidłowej funkcji, co skutkuje rozwojem wielu chorób. Jedną z klas takich często nieprawidłowo sfałdowanych białek są amyloidy.

Amyloidy zostały po raz pierwsze zidentyfikowane w preparatach histologicznych wyizolowanych z centralnego układu nerwowego. Początkowo ze względu na swoją włóknistą strukturę zostały omyłkowo sklasyfikowano jako zbudowane ze skrobi ("amylum"), skąd wzięły one swą obecną nazwę. Wkrótce jednak stało się jasne, że zbudowane są z białek. Z tego powodu przez wiele lat były one głównie kojarzone z ich rolą w rozwoju chorób neurodegeneracyjnych, takich jak choroba Alzheimera czy Parkinsona. Potencjalna rola w patologii wspomnianych chorób przełożyła się na duże zainteresowanie tymi strukturami w środowisku biologów i biochemików. Bardziej szczegółowe badania pokazały szczególne właściwości włókien amyloidowych takie jak zdolność do wiązania niektórych barwników, w tym czerwieni kongo oraz tioflawiny T, które wciąż stanowią jedną z podstawowych metod identyfikacji amyloidów. Wysokorozdzielcze metody mikroskopowe oraz krystalograficzne pozwoliły zbadać szczegóły ich struktury oraz morfologii. Odkryto, że kluczem do wyjaśnienia niespotykanej stabilności agregatów amyloidowych jest tak zwana struktura zamka błyskawicznego (ang. steric zipper) utworzona przez dwie ściśle przylegające beta-kartki utrzymywane razem przez oddziaływania pomiędzy zazębiającymi się resztami aminokwasowymi. Niezwykła stabilność włókien amyloidowych została również kreatywnie wykorzystana przez naturę. W ciągu ostatnich dwóch dekad tak zwane amyloidy funkcjonalne zostały zidentyfikowane w bardzo wielu organizmach przynależących do wszystkich królestw życia wliczając w to człowieka.

Kolejnym przełomem było odkrycie, że obecność agregatów amyloidowych jednego białka może drastycznie przyspieszać agregację innych białek. Proces ten nazwano

krzyżową inicjacją agregacji i zaczęto w nim upatrywać molekularnych podstaw współwystępowania chorób amyloidowych. Wkrótce potem pokazano również, że w niektórych przypadkach podobny mechanizm może prowadzić również do spowolnienia agregacji. Okazało się, że proces ten jest również wykorzystywany przez niektóre organizmy, czego przykładem może być ekspresja białka CsgB inicjującego agregację białka CsgA przez bakterie *E. coli*. CsgA jest funkcjonalnym amyloidem bakteryjnym budującym rusztowanie dla biofilmu tworzonego przez wybrane szczepy tej bakterii. Nawet bardziej spektakularnym przykładem takich interakcji mogą być białka NLR produkowane przez szereg gatunków grzybów. Są to białka stanowiące swego rodzaju układ odpornościowy grzybów chroniące je poprzez uruchamianie szeregu reakcji prowadzących do śmierci zainfekowanych komórek. Pokazano że jeden z etapów przekazywania sygnału jest tutaj realizowany właśnie za pomocą krzyżowych interakcji amyloidów.

Pomimo dużego znaczenia amyloidów nasza wiedza na ich temat jest wciąż ograniczona. Jednym z największych ograniczeń w badaniu tych struktur jest konieczność przeprowadzania skomplikowanych i czasochłonnych eksperymentów. Z tego względu udało się dobrze scharakteryzować stosunkowo niewiele białek przejawiających skłonności do agregacji amyloidowej. Dużym wyzwaniem pozostaje identyfikacja tak regionów odpowiedzialnych za agregację (hot-spotów) amyloidowych będących relatywnie krótkimi fragmentami ich sekwencji, których obecność jest wystarczająca do utworzenia agregatu przez białko.

Aby rozwiązać ten problem, zaproponowano wiele obliczeniowych metod identyfikacji hot-spotów amyloidowych w białkach. Jedne z pierwszych metod opierały się na modelach fizykochemicznych zbudowanych w oparciu o nieliczne wówczas zbiory sekwencji amyloidowych. Pomimo ich ograniczeń modele te okazały się istotnym wsparciem prac eksperymentalnych, pozwalając w sposób bardziej racjonalny planować badania. Wraz ze wzrostem liczby doświadczalnie scharakteryzowanych sekwencji pojawiły się pierwsze bazy danych, takie jak AmyLoad czy Waltz, zbierające sekwencje amyloidów. Rosnąca ilość danych pozwoliła na budowę bardziej złożonych modeli statystycznych, a wreszcie także modeli uczenia maszynowego. Choć w ostatnich latach skuteczność metod istotnie wzrosła, to wciąż jest ona niewystarczająca do wykorzystania ich na skalę całych proteomów. Dużym problemem jest tutaj niewystarczająca specyficzność, która przekłada się na wiele fałszywie pozytywnych wyników. Co więcej, większość obecnie dostępnych metod nie została zaprojektowana do pracy z amyloidami funkcjonalnymi, które często różnią się składem aminokwasowym od swoich patologicznych odpowiedników. W tym przypadku jednym z głównych ograniczeń jest niewystarczająca ilość dobrze przebadanych sekwencji amyloidów funkcjonalnych. Wreszcie, na chwilę obecną nie ma dostępnych narzędzi pozwalających na przewidywanie interakcji krzyżowych. Celem pracy doktorskiej było zatem opracowanie lepszych metod identyfikacji regionów amyloidowych oraz badania ich interakcji.

2.2 Wyniki

Ze względu na opisane wcześniej ograniczenia dostępnych metod predykcji regionów amyloidowych, zdecydowano się na zaproponowanie nowej metody obliczeniowej. Założeniem było stworzenie narzędzia o wyższej skuteczności, a przede wszystkim wyższej specyficzności. W tym celu zaproponowano model łączący modelowanie

strukturalne z metodami uczenia maszynowego. Stworzona metoda przyjmuje podobne założenia jak jedno z pierwszych dostępnych narzędzi, a mianowicie metoda profili 3D. Wspomniana metoda wykorzystuje nawlekanie sekwencji badanego fragmentu na strukturę włókna amyloidowego uzyskaną z badań krystalograficznych i obliczanie energii uzyskanego w ten sposób modelu. Niestety metoda ta zakłada tylko jeden możliwy sposób upakowania peptydów w strukturze agregatu. Wraz z pojawieniem się większej liczby dostępnych struktur włókien amyloidowych pokazano, że możliwych jest kilka różnych sposobów upakowania. Opierając się na analizie grup symetrii zaproponowano dziesięć możliwych klas strukturalnych z czego siedem zostało potwierdzonych doświadczalnie. Pierwszym krokiem było zatem uwzględnienie tej różnorodności w mojej procedurze modelowania. Zaproponowano metodę korzystającą z narzędzia Modeller do wykonywania nawlekania a następnie ocenę uzyskanych modeli przy pomocy potencjału statystycznego DOPE zaimplementowanego we wspomnianym programie oraz szeregu innych z pakietu Rosetta. Sprawdzone również możliwość zbudowania modelu uczenia maszynowego wykorzystującego obliczone parametry do predykcji regionów amyloidowych. Metoda ta została następnie istotnie rozwinięta. Sprawdzone dodatkowe rodzaje modeli uczenia maszynowego oraz przeprowadzono dodatkowe dostrojenie ich parametrów. Przeprowadzona została szczegółowa analiza skuteczności metody na większych zbiorach danych. Algorytm modelowania został zoptymalizowany i zrównoleglony w celu możliwości wykorzystania go do analizy dużych zbiorów danych. Na tej podstawie zostało stworzone narzędzie PATH (Prediction of Amyloidogenicity by THreading), które jest publicznie dostępne na stronie:

<https://github.com/KubaWojciechowski/PATH>. Wyniki tych badań zostały opublikowane w pracy (Wojciechowski i Kotulska 2020).

Podczas prac nad przygotowaniem narzędzia PATH natrafiłszy na problem, który wcześniej zauważyli również autorzy metody AmyloGram. Zidentyfikowali oni kilka peptydów dla których ich narzędzie konsekwentnie, z dużą pewnością dawało przeciwną klasyfikację fragmentów amyloidowych (agregujące jako nie agregujące i odwrotnie). Dokładniejsza analiza zbioru danych pokazała również, że kilka z sekwencji dostępnych w bazie danych WaltzDB, występuje tam dwa razy zarówno jako fragmenty agregujące jak i nieagregujące. Aby sprawdzić takie przypadki, wybrane zostały 24 peptydy, zaklasyfikowane przez AmyloGram inaczej niż w bazie WaltzDB. Peptydy te zostały zsyntezowane a następnie przebadane eksperymentalnie z wykorzystaniem technik spektroskopowych oraz mikroskopii sił atomowych. Jak się okazało, większość z tych niejednoznacznych 24 sekwencji została źle zaetykietowana w bazie danych. Wyniki te pokazują dość dużą odporność narzędzi takich jak PATH czy AmyloGram na źle zaetykietowane dane, pomimo, że narzędzia te były na nich uczone. Wyniki te zostały opublikowane w pracy (Szulc i inni 2021a).

Obie prace udowadniają pierwszą z postawionych w tej pracy hipotez a mianowicie, że **Modelowanie strukturalne w połączeniu z metodami uczenia maszynowego poprawia skuteczność przewidywania fragmentów amyloidowych.**

Na kolejnym etapie naszych badań zwróciliśmy uwagę na problem identyfikacji funkcjonalnych amyloidów bakteryjnych. W tym celu przeprowadziliśmy analizę bioinformatyczną oraz eksperymentalną białka CsgA z dwóch gatunków bakterii: *Escherichia coli* i *Salmonella enterica*. Białko to tworzy amyloidy funkcjonalne stanowiące jeden z głównych elementów biofilmu. Ciekawą własnością CsgA jest

jego modułowa budowa, składa się ono z pięciu fragmentów powtórzonych R1-R5. Badania pokazały, że pomimo dużego podobieństwa wszystkich fragmentów, tylko fragmenty R1, R3 i R5 w białku z *E. coli* są zdolne do tworzenia agregatów. Analiza tych różnic może zatem rzucić światło na cechy sekwencji amyloidów funkcjonalnych decydujące o ich zdolnościach agregacyjnych. Ze względu na niedużą liczbę fragmentów występujących w CsgA, zdecydowaliśmy się rozszerzyć nasze badania o analizę jego dużo słabiej poznanego homologa CsgA z bakterii *S. enterica*. W ten sposób uzyskaliśmy zbiór 10 dość podobnych fragmentów, różniących się zdolnością do tworzenia agregatów. Przeprowadzono szczegółową charakteryzację każdego z fragmentów z wykorzystaniem metod spektroskopii oscylacyjnej oraz wysokorozdzielczych technik obrazowania, w tym transmisyjnej mikroskopii elektronowej. Na potrzeby przeprowadzonych badań opracowano metodologię badania amyloidów przy pomocy spektroskopii ramanowskiej z transformatą Fouriera (FT-Raman). Pokazano możliwość wykorzystania tej metody jako techniki komplementarnej do szeroko stosowanych technik spektroskopii podczerwieni takich jak ATR-FTIR czy mikro-IR w kontekście badania agregatów peptydowych. Przeprowadzona analiza bioinformatyczna pokazała, że obecnie dostępne metody identyfikacji regionów amyloidowych działają dużo słabiej na sekwencjach amyloidów funkcjonalnych. Wyniki te zostały opublikowane w pracy (Szulc i inni 2021b). Badania w tym zakresie są dalej kontynuowane.

Równolegle, we współpracy z zespołem Prof. Petera Roya z uniwersytetu w Toronto rozpoczęliśmy poszukiwania nowych amyloidów w proteomie modelowego organizmu *Caenorhabditis elegans*. Naukowcy z Kanady prowadząc badania rozwoju tego organizmu zaobserwowali w okolicach jego otworu gębowego struktury wiążące Czerwień Kongo - barwnik tradycyjnie wykorzystywany do identyfikacji włókien amyloidowych. Co więcej, analiza ekspresji genów w trakcie jednej z faz rozwojowych (linienia) pokazała zwiększoną ekspresję enzymów rozkładających amyloidy, oraz inhibitorów agregacji amyloidowej. Niejasnym pozostawało które z białek w proteomie *C. elegans* mają charakterystykę amyloidów. W tym celu przeszukaliśmy cały proteom za pomocą dwóch narzędzi: AmyloGram i opracowanym w ramach tej pracy PATH. W 37% badanych białek zidentyfikowaliśmy przynajmniej jeden hot-spot amyloidowy. Zaobserwowano, że w większości nie są to białka wydzielane na zewnątrz komórki, lecz nie stwierdzono ich nadreprezentacji w strukturach tworzących gardziel. Wyniki te zostały opublikowane w pracy (Kamal i inni 2022). Tak duża liczba znalezionych potencjalnych amyloidów powinna skutkować tworzeniem dużej liczby włókien amyloidowych w różnych częściach tego organizmu. Jednak w rzeczywistości sytuacja taka nie ma miejsca. Istnieją dwa potencjalne wytłumaczenia tej obserwacji. Po pierwsze obecnie dostępne narzędzia działają na zasadzie skanowania sekwencji nakładającym się oknem przesuwным, najczęściej o długości 6 aminokwasów. Oznacza to że przykładowo dla białka o długości 300 aminokwasów, wykonywane jest $300 - 6 = 294$ sprawdzenia, co nawet przy bardzo restrykcyjnych parametrach dających specyficzność na poziomie 0.99 skutkuje średnio trzema fałszywie pozytywnymi wynikami na białko. Nawet jeśli stosujemy dwie różne metody, jak to miało miejsce w tej pracy, przy badaniach w skali całego proteomu wciąż istnieje ryzyko wystąpienia wielu fałszywie pozytywnych wyników. Niemniej jednak przy tak dużej liczbie trafień jest niemal niemożliwe aby wszystkie one były wynikami fałszywie pozytywnymi. Szczególnie, że wówczas nie powinniśmy obserwować statystycznie istotnych różnic pomiędzy ich rozmieszczeniem w różnych

tkankach czy typach białek. Możliwe, że występujące w proteomie *C. elegans* fragmenty amyloidowe zlokalizowane są w większości w hydrofobowych rdzeniach białek, które w normalnych warunkach nie są eksponowane do środowiska, przez co nie powodują agregacji. Hipoteza ta wydaje się tym bardziej prawdopodobna, że przed rozpoczęciem linienia, oprócz ekspresji szeregu enzymów katabolicznych ekspresjonowane są również białka chaperonowe oraz enzymy i inhibitory powstrzymujące agregację. Zatem istnieje mechanizm zdolny do unieszkodliwiania rozkładanych w tym procesie białek które mogą zawierać fragmenty amyloidowe.

Ze względu na opisane trudności z bezpośrednim wykorzystaniem narzędzi do predykcji regionów amyloidowych, w następnym projekcie przyjęliśmy zupełnie inne podejście. We współpracy z Prof. Witoldem Dyrką przyjrzelśmy się bliżej grzybowym białkom NLR. Ciekawą cechą tych białek jest obecność amyloidowych motywów sygnałowych, które biorą udział w przekazywaniu sygnału pomiędzy białkiem receptorowym a efektorowym. Aby lepiej zrozumieć działanie tego systemu przeprowadziliśmy szczegółową anotację domen białkowych występujących w grzybowych białkach NLR. Aby wykryć potencjalne amyloidowe motywy sygnałowe, przeszukaliśmy krótkie (<150 aa) N oraz C-końce wykorzystując narzędzia do identyfikacji *de novo* motywów oraz dedykowanych modeli językowych takich jak gramatyki bezkontekstowe. Przeanalizowaliśmy również współwystępowanie znalezionych motywów w parach receptor-efektor. Badania te doprowadziły między innymi do zidentyfikowania nowego amyloidowego motywu sygnałowego PUASM związanego z domeną PNP_UDP (Pnp Udp Amyloid Signaling Motif). Wyniki tych badań zostały opublikowane w pracy (Wojciechowski i inni 2022).

Opisane powyżej wyniki pokazują problemy z bezpośrednim zastosowaniem metod bioinformatycznych do analizy dużych ilości danych biologicznych. Mniejsza skuteczność metod obliczeniowych dla amyloidów funkcjonalnych oraz problem dużej ilości potencjalnie fałszywie pozytywnych wyników sugeruje konieczność opracowania bardziej specyficznych metod. Stąd druga hipoteza tej pracy potwierdzona wynikami wspomnianych badań mówi, że **Poszukiwanie amyloidów w skali genomowej wymaga wyspecjalizowanych metod.**

Ostatnim etapem badań było stworzenie narzędzia do przewidywania krzyżowych interakcji amyloidów. W tym celu niezbędne było zebranie rozszanych po literaturze danych i utworzenie pierwszej na świecie bazy danych interakcji amyloidowych. Zdecydowano się na budowę bazy grafowej, w której każdy amyloid prezentowany jest jako węzeł, natomiast krawędzie oznaczają interakcje pomiędzy białkami. Na chwilę obecną baza zbiera blisko 900 przypadków interakcji krzyżowych zebranych z prawie 200 artykułów naukowych podzielonych ze względu na charakter interakcji. Baza jest publicznie dostępna pod adresem: <http://AmyloGraph.com>. Oprócz utworzenia bazy, przy okazji jej tworzenia zaproponowano ustandaryzowaną terminologię. Wyniki te zostały opublikowane w pracy (Burdukiewicz i inni 2023).

Zwieńczeniem prac było opracowanie pierwszej metody do przewidywania amyloidowych interakcji krzyżowych PACT (Prediction of Amyloid Cross-interactions by Threading). Ze względu na ograniczoną liczbę danych oraz dużą nadreprezentację interakcji kilku dobrze przebadanych białek, jednym z najważniejszych założeń było zbudowanie jak najbardziej odpornego modelu. Z tego względu zdecydowano się na wykorzystanie modelowania strukturalnego. Podobnie jak w przypadku PATH, PACT wykorzystuje nawlekanie obu sekwencji na znana strukturę włókna amyloidowego. Końcowy model w tym przypadku składa się z czterech łańcuchów, po dwóch

każdego ze sprawdzanych fragmentów. Ponieważ wchodzące w interakcje fragmenty mogą być różnej długości, zdecydowano się na wykorzystanie modelu długiego peptydu jako szablonu do modelowania i korzystania tylko z części z niego w przypadku krótszych peptydów. W takich wypadkach konieczne było również znormalizowanie energii modelu poprzez podzielenie jej przez średnią długość badanych fragmentów. Zaproponowana metoda uzyskała bardzo wysoką skuteczność predykcji interakcji krzyżowych oceniana na podstawie wartości parametrów AUC (0.88) i F1 (0.82). Pokazano również, że zaproponowana metoda może być z powodzeniem wykorzystana do predykcji homoagregacji, i jest skuteczna również w przypadku amyloidów funkcjonalnych. PACT został następnie wykorzystany do przewidywania interakcji pomiędzy różnymi wariantami białka CsgA i ludzkim białkiem odgrywającym kluczową rolę w rozwoju choroby parkinsona - alpha-synucleiną, oraz do ustalenia które z fragmentów CsgA są kluczowe z punktu widzenia interakcji z ludzką amyliną. Wyniki te zostały opublikowane w pracy (Wojciechowski i inni 2023) i potwierdzają trzecią hipotezę mówiącą, że **Modelowanie strukturalne pozwala przewidzieć krzyżowe interakcje amyloidów**. PACT jest dostępny pod adresem <https://github.com/KubaWojciechowski/PACT> oraz jako webserwer pod adresem <https://pact.e-science.pl/pact/>

2.3 Podsumowanie

Wszystkie postawione w tej pracy tezy zostały potwierdzone. Zaproponowano nową metodę przewidywania fragmentów amyloidowych i pokazano jej odporność na błędy w danych uczących. Pokazano ograniczenia możliwości stosowania predyktorów amyloidowych w przypadku danych w skali całych proteomów oraz problemy związane z identyfikacją amyloidów funkcjonalnych. Wreszcie zaproponowano metodę przewidywania krzyżowych interakcji amyloidów, która może być również z powodzeniem wykorzystana do identyfikacji regionów amyloidowych również w przypadku amyloidów funkcjonalnych. Stworzono dwa narzędzia których kod źródłowy został publicznie udostępniony oraz web serwer pozwalający na przewidywanie interakcji. Zaproponowane rozwiązania w istotny sposób zwiększają wachlarz oraz użyteczność metod obliczeniowych w badaniach amyloidów.

Chapter 3

General Introduction

Proteins, lipids, and nucleic acids are the most fundamental building blocks of life. These nano-scale machines encoded in genomes of all living organisms perform an astonishing variety of functions. An orchestrated action of around 20 000 [1] of them in a human organism enables us to survive, thrive, think, and occasionally write a Ph.D. thesis. The fascinating palette of protein functions is a result of a multitude of their possible structures, cleverly encoded in their amino acid sequences [2]. twenty different amino acids can be combined in an astronomical number of ways producing a sequence also known as a primary protein structure. However, this is just the tip of the iceberg in terms of complexity, since in an aqueous environment it folds into well-organized structures. These so-called secondary structures are then arranged in three dimensions to form a tertiary structure. Finally, some proteins are formed by more than one polypeptide chain. The mutual orientation of different folded chains makes the quaternary structure [3].

Although perfected through eons of evolution, this mechanism is not always working as intended. Sometimes even a single mutation can drastically change the protein structure and therefore disturbed its functioning, leading to a disease. One of the striking examples of protein misfolding is the formation of amyloid fibers playing an important role in the onset of neurodegenerative diseases [4].

The history of amyloids is rather convoluted and full of unexpected turns. This term was coined in 1854 by Rudolf Virchow, a German physician, and biologist from eastern Pomerania present-day Świdwin Poland, who used it to describe microscopic tissue abnormalities. Based on iodine staining, he mistakenly concluded that they consist of starch (*amylum*) and therefore called them amyloids [5]. It took a few years before scientists realized that they are in fact protein aggregates. In 2006 an article published in *Nature* showed the connection between amyloid beta and Alzheimer's disease [6]. This discovery directed the attention of many researchers to this protein, making it one of the most important drug targets. However, recently shreds of evidence of image manipulation undermined the credibility of this work [7]. For more than 100 years amyloids were almost exclusively associated with disorders [8], yet in the last two decades it became clear that they can be creatively used by a variety of organisms [9].

History continues, and we learn new things about this class of proteins every day. In the following sections, I will discuss in more detail the current state of knowledge.

3.1 Amyloids

Amyloids are insoluble, fibrillar protein aggregates composed of multiple proteins stacked together along the fiber axis. The core of the fiber consists of tightly packed polypeptides in beta-sheet conformation. The interdigitating sidechains of such fragments form a structure resembling a zipper, hence such an arrangement is often referred to as a “steric zipper structure” or cross-beta structure. This packing results in a characteristic diffraction pattern that was first described in 1935 by British biophysicist William Astbury [10]. Further structural research involving Nuclear Magnetic Resonance (NMR) spectroscopy [11] and X-ray diffraction techniques revealed molecular details of amyloid cores. The tight packing of monomers, the formation of salt bridges, and the regular network of hydrogen bonds play a pivotal role in stabilizing the structure of an aggregate [12]. The resulting fiber has extraordinary material properties comparable to that of steel [13]. Furthermore, they are resistant to proteolytic cleavage, and many denaturing factors [14].

With the growing number of available molecular structures of amyloid fibers, it became clear that there is more than one way of packing peptides into a steric zipper structure. In fact, based on analysis of symmetry groups, structural biologists proposed ten different possible arrangements out of which seven were experimentally confirmed [15]. Surprisingly, it was also shown that the same amyloidogenic proteins can form differently arranged aggregates. This phenomenon called polymorphism can profoundly affect the onset of diseases [16].

3.2 Amyloids and diseases

Since the very discovery of amyloid fibers, they were associated with diseases. In 1907 Alois Alzheimer described amyloid plaques found in the brains of patients suffering from dementia [17]. Later these plaques were found to be formed by two proteins amyloid beta and tau [18]. A few years later, Alzheimer described similar structures in patients suffering from Pick’s disease [19], and a year later in Parkinson’s disease [20]. This time, however, plaques turned out to be formed by Alpha-synuclein [21].

Amyloids were found to be involved in a number of other neurodegenerative disorders including Huntington’s disease [22], amyotrophic lateral sclerosis [23], or Creutzfeldt–Jakob disease [24]. But they are also involved in a range of other disorders. For example, aggregates of Islet Amyloid Polypeptide (IAPP) were found in Type II diabetes patients [25]. Further studies showed that IAPP aggregates show cytotoxic effects against pancreatic cells [26]. Interestingly, in this case, the most toxic were oligomers formed during the early stages of the aggregation process rather than mature fibers. Higher toxicity of oligomers was also shown for Alpha-synuclein [27] and amyloid beta [28]. Later, a number of other amyloids were shown to be the most toxic in oligomeric form [29].

3.3 Amyloid structures

The molecular structure is crucial for understanding amyloid properties. As briefly mentioned in the previous paragraph, one of the defining features of amyloid fibers is

the cross-beta structure of their core. In fact, usually the core of the fiber is formed by a relatively short fragment of the protein. The presence of such fragments, called amyloid hot spots, turned out to be sufficient to drive the aggregation of the whole proteins [30]. Therefore, the first molecular structures of amyloid fibers were solved for very short fragments, usually around six amino acids long. The first structures confirmed the previously proposed model of the cross-beta structure, consisting of two beta-sheets with interdigitating amino acid residues. It was shown that the monomers are stacked on top of each other along the fiber axis. Also, the distance between layers was found to be around 4.8 Å [31]. The first structures shed light on the origins of the incredible stability of amyloids. First of all, a very regular network of hydrogen bonds was observed between subsequent layers [32], but not between sheets within the same layer, which are held together by hydrophobic effect and weak interactions between tightly fitting side chains. It was also shown, that some amino acids, like glutamine, asparagine and tyrosine, are usually aligned to form so-called ladders which introduce additional interactions, such as hydrogen bonds and pi-pi interactions between layers [33, 34].

In the beginning, the only techniques that were able to provide molecular structures were X-ray diffraction and NMR spectroscopy. However, in recent years, Cryo-Electron Microscopy (Cryo-EM) was introduced and applied to the study of amyloid structures [35]. This technique can overcome many problems of previously utilized techniques. Unlike X-ray diffraction, Cryo-EM does not require crystallization of the sample, and unlike solution NMR, it can be used for the study of insoluble structures. It is also possible to study much larger fragments or even whole proteins [36, 37].

With the growing number of available structures, it became apparent that there exists more than one possible packing of peptides into the cross-beta structure. The most obvious possibilities are parallel and antiparallel arrangements of beta sheets, but the variability does not end here. Based on symmetry operations, ten different structural classes were proposed, and seven of them were experimentally confirmed [15] (Fig. 3.1).

Unlike in the case of most globular proteins, the same amyloidogenic sequence can form different structures. Such structures are usually energetically very similar [38], yet can lead to visible differences in fibril morphology [39]. It was also shown that different polymorphs can be specific for different diseases. For example, different polymorphs of tau protein from Alzheimer's disease, Pick's disease, chronic traumatic encephalopathy, and corticobasal degeneration show significantly different structures [40]. Despite years of research and a rapidly growing number of available structures, amyloids can still surprise. For almost a century, amyloids were thought to be composed of peptides in beta-sheet conformation. However, very recently a team of researchers led by Meytal Landau published the structure of Uperin - an amphibian antimicrobial peptide capable of forming amyloid fibers. To everyone's surprise, the fiber core consisted mostly of alpha-helical fragments. What was even more astonishing, the described structure was able to switch conformation between newly discovered cross-alpha and classical cross-beta [41]. Later, similar structures and switching behavior were described for plenty of other antimicrobial amyloids [42].

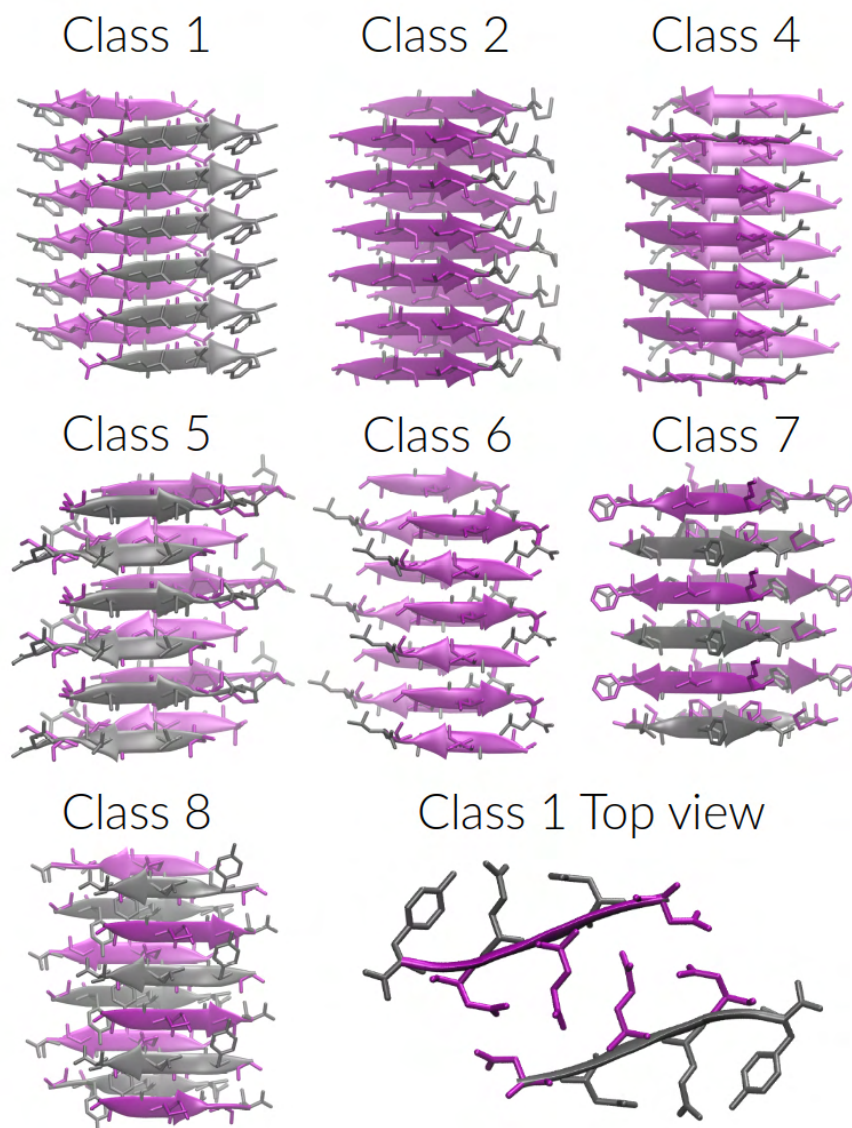


Figure 3.1: seven experimentally confirmed structural classes of amyloid fibers. Each class shows different packing of peptides in steric zipper structure.

3.4 Functional Amyloids

Although amyloids are mostly known for their involvement in neurodegenerative diseases, they can also be utilized by a wide variety of organisms to perform physiological functions. For the last two decades, numerous such functional amyloids have been described. Most of the early discoveries were made by studying microorganisms, the term functional amyloids was first used in the year 2000 to describe Hydrophobins. These are fungal proteins whose main role is an alteration of cell surface properties and formation of hydrophobic aerial structures like aerial hyphae and spores [43]. Two years later amyloids were found to be the main components of curli fibers of *Escherichia coli*, which creates a scaffold for biofilm formation. The more detailed study revealed the presence of Curli Specific Genes (Csg) operon, which turned out to include genes encoding functional amyloid CsgA [44]. Later, functional amyloids were found to be responsible for biofilm formation in a number

of other microorganisms, including FapC from *Pseudomonas* sp. [45], TasA from *Bacillus subtilis* [46] or SuhB [47] and PSM [48] from *Staphylococcus aureus*. Also, some adhesive proteins like Bap [49] or Aap [50], responsible for cell-to-cell and cell-to-surface adhesion, show amyloid aggregation propensity.

However, the role of functional amyloids in microorganisms is not limited to adhesion or biofilm formation. The ability to form reversible amyloids was shown that Cdc19 protein from yeasts plays an important role in stress response. This protein functions as a pyruvate kinase, which under hostile conditions aggregates, which leads to stopping of a cell cycle. After the stressor is removed, the fibers decompose and the cell cycle is restored [51]. On the other hand, the Rim4 protein from the same organism is responsible for translational repression and works in the aggregated state [52]. Furthermore, amyloids can be used for signaling. Nod Like Receptors (NLR) proteins play an important role in the immune responses of fungi [53] and filamentous bacteria [54]. In both cases, signal transmission between receptor and effector proteins can be realized by aggregation and propagation of amyloid fold.

Functional amyloids can also be found in much more complicated organisms, including humans. For example, PMEL 17 is a protein found in human melanocytes that was shown to play an important role in melanin synthesis [55]. Amyloids are also involved in programmed necrosis. They mediate the formation of the RIP1/RIP3 complex, which initiates processes leading to cell death [56].

Although both pathological and functional amyloids share a lot of features, there are some important differences. The most apparent difference is in amino acid composition. Sequences of functional amyloids contain more polar amino acids, especially asparagine, and glutamine. On the other hand, in pathological amyloids more positively charged residues were observed [35].

3.5 Amyloid cross-interactions

Another interesting feature of amyloids is their ability to interact with each other. This was first demonstrated in the work by Lundmark and coworkers. The authors showed that mice treated with pre-formed amyloid fibers of several different proteins developed amyloid plaques in their spleens within 16 days [57]. Soon, as more examples of such interactions were shown in the literature, it became clear that this might explain the comorbidities of different amyloid-related diseases. For example, such interactions were shown between amyloid beta and alpha-synuclein [58] or between amyloid beta and IAPP [59]. However, amyloid cross-interactions do not always lead to increased aggregation rate. In some cases, the opposite effect can be observed. A classic example can be the interaction between IAPP from rats and humans. In this case, the presence of rodent IAPP leads to the inhibition of human IAPP aggregation [60]. This mechanism was even used to design drugs against diabetes that mimic rodent IAPP [61].

Despite the importance of amyloid cross-interactions, their mechanisms still remain unclear. In general, amyloid aggregation follows a sigmoidal characteristic with three distinctive phases: nucleation, elongation, and the equilibrium phase. In the nucleation phase monomers change their structure upon binding with other monomers and form oligomers. Oligomers then act in a similar manner as crystallization seeds and seed the growth of an amyloid fiber (Fig 3.2) [62]. Cross-seeding

seems to follow the same general steps, but in this case the pre-formed oligomers of one protein can seed the aggregation of another. This results in a reduced nucleation phase, and therefore, a faster formation of amyloid fibers. This model is usually referred to as template-assisted cross-seeding. Another model called the conformational selection model assumes the formation of heterogeneous seeds which consist of both interacting peptides [63]. In both cases, such interaction requires that both interacting partners should be able to adopt a similar structure [64]. It was shown that even highly dissimilar proteins like amyloid beta and PrP can interact in that way [65]. The case of interactions leading to slower aggregation is much less understood. A computational study performed by Zhang and coworkers suggested that hetero aggregates formed by IAPP from humans and rats are more energetically favorable compared to homo aggregates [66]. This might suggest the formation of complexes which are less likely to rearrange into mature fibrils [64].

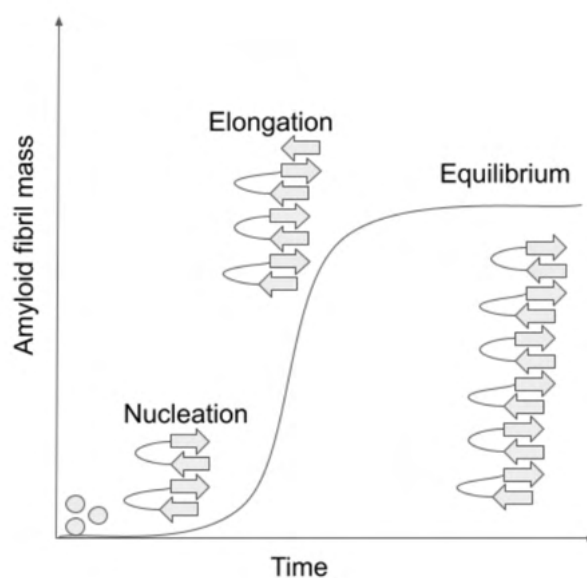


Figure 3.2: Sigmoidal characteristics of amyloid aggregation.

3.6 Computational methods for studying amyloids

Experimental research laid the foundation for our knowledge about amyloid proteins. X-ray diffraction [31], NMR spectroscopy [11], and cryo-EM [36] provided insight into the molecular structure of amyloid fibers. The morphology of amyloid assemblies can be observed using high-resolution imaging techniques such as electron microscopy [67] or atomic force microscopy [68]. Biochemical assays including Congo red [69] and Thioflavin T [70] staining proved to be useful in identifying amyloid plaques in tissue samples and provided a way of monitoring the aggregation process. Unfortunately, all experimental techniques require obtaining a sample of protein or peptide which, due to aggregation propensity, is by itself a challenging task [71]. As a result, experimental studies are expensive and time-consuming. Therefore, their

use on a genome-wide scale is still limited. To overcome this limitation several computational methods have been developed. In the following paragraphs, important computational methods and resources will be discussed.

3.7 Amyloid databases

The development of computational methods for studying amyloids requires data to build and test models. In the early days, such information was scattered across a large body of literature. One of the earliest approaches to collecting this information was the `fibril_one` database. It gathered data about 250 mutations of 22 amyloid proteins in different conditions [72]. However, as the database relied only on disease-related proteins and their mutations, it was unsuitable for developing computational methods. One of the first datasets widely used for the development of amyloid-prone regions was published in 2004. The authors aimed to better understand aggregation driving patterns by mutating synthetic amyloidogenic peptide STVIIIE [30]. This provided an important insight into the properties of amyloid hot spots, but also introduced a strong bias due to the high similarity of tested sequences. Soon after, the AMYPdb database was published. This is the oldest still active database which contains full-length sequences, of amyloids and structures of many precursor proteins [73].

With the growing number of experimentally studied amyloid hot spots, two dedicated databases appeared in 2015. The first of them, Waltz-DB focuses on amyloidogenic hexapeptides. The authors not only gathered the data from existing databases and literature but also tested a number of new sequences [74]. In 2020 it received a major update and now contains more than 1400 amyloidogenic and non-amyloidogenic hexapeptides [75]. In the same 2015 year, the AmyLoad database was published, which also focuses on aggregation-prone regions, but unlike Waltz-DB it collects regions of all lengths. Currently, it contains almost 1500 hot spots of lengths between 4 and 89 amino acids [76].

In 2016 CPAD database was released. This was one of the first approaches to provide a more unified database that will not only contain hot-spot and full-length sequences, but also structures and information about aggregation kinetics [77]. A significant update for this database was released in 2020 [78].

With the rapidly growing number of functional amyloids, and prion sequences, it became apparent that such sequences are an important part of an amyloid landscape. Developed in 2018 AmyPro database, apart from containing information about sequences and aggregation-prone regions, distinguishes between functional and pathological amyloids and prions [79].

Recent advances in structural studies of amyloid fibers provided a large collection of molecular structures of amyloid structures. The PDB_Amyloid database, developed in 2019, aimed at collecting these structures [80]. As of June of 2023, it contains over 770 records.

This year, the first database of amyloid cross-interactions AmyloGraph was developed. It contains information about almost 900 pairs of amyloid interactions extracted from almost 200 articles. Each record reports the sequence of interacting proteins, as well as information about observed changes in aggregation kinetics, including faster and slower aggregation or inhibition of fibril formation. It also reports evidence of the formation of heterogenous aggregates formed by two different

amyloids. In this work, we also proposed a standardized terminology for describing amyloid interactions [81].

3.8 Aggregation-prone region predictors

One of the most important tasks in characterizing an amyloid protein is the identification of its amyloid hot spots. This usually requires the testing of numerous peptides and is a labor-intensive task. Therefore, many computational methods were developed to identify such regions. The first attempt at the identification of aggregating peptides was Tango. This method is grounded in statistical physics. To estimate the aggregation propensity of a peptide, it calculates the partition function of the conformational space and allows for taking into consideration temperature, pH, and ionic strength [82]. Although it can predict aggregation, the authors of the method emphasize that this is not necessarily equivalent to amyloidogenicity [83].

One of the first methods dedicated to the identification of amyloidogenic fragments was the 3D profiles method. It is based on molecular threading of the fragment of interest on the first solved molecular structure of an amyloid fiber. The main idea behind this method is the amyloidogenic fragment threaded on the steric zipper structure will be more energetically favorable than non-amyloidogenic fragments [84]. This method was used for the first large-scale identification of amyloid-prone regions. The authors scanned genomes of more than 100 organisms and found that aggregation-prone regions are in fact very common in proteins. However, in most cases, such fragments are buried in the hydrophobic core and therefore are unlikely to drive the aggregation [85].

In the same year as the 3D profiles method was released, the PASTA predictor was introduced [86]. This method was then significantly updated in 2014 [87]. It is another physicochemical method that uses statistical potential to estimate the pairwise interactions between amino acids in the query sequence. An even simpler approach was chosen by the authors of Aggrescan software [88]. This method uses an experimentally derived aggregation propensity score based on the mutational analysis of the amyloid beta core region [89].

The growing number of known amyloid sequences led to the development of machine learning-based methods. In 2009 two such methods were proposed - NetCSSP [90] and Pafig [91]. The first method uses an artificial neural network, while the other a supported vector machines classifier. The same increase in data availability opened the way for statistics-based models like FoldAmyloid [92] and Waltz [93]. Waltz is based on Position Specific Scoring Matrices and was one of the first methods that emphasized the importance of specific sequences of amino acids.

As the number of available predictors increased, meta-predictors were introduced. These tools combine a range of available methods and use them to perform consensus classification of the sequence of interest. The first such method was AmylPred developed in 2009 and used a relatively small number of tools [94]. Later on, more and more complex methods were built, including AmylPred2 [95] and MetAmyl [96]. Interestingly, MetAmyl not only calculates the consensus of available methods, but it uses their outputs as input for the logistic regression model. The authors of this method precomputed the classification for all possible 64 million hexapeptides and stored the results. This not only reduced the time of prediction, but also made the method independent of the availability of other tools and therefore more

reproducible. This is an important feature since many amyloidogenicity prediction methods are not further supported or even not available.

The development of amyloid predictors leads also to the development of a new machine-learning method. FISH Amyloid is a method of identifying patterns in the co-occurrence of pairs of amino acids [97].

Up to this point, all the discussed methods scanned the protein sequence with a short, usually six amino acids long, sliding window. This was caused by the available datasets, which mostly consisted of hexapeptides. The first method that broke the hexapeptide sliding window paradigm was ArchCandy [98]. The main assumption of this method is that aggregation-prone region in amyloids often forms beta-arch structure. This structural motif is formed by two beta strands connected by a short loop. To overcome the problem of a constant-length sliding window. First, ArchCandy scans the sequence for potential candidate fragments. If such a fragment is found, modeling of beta-arch begins, starting from the middle loop and extending towards the ends as long as the extending beta-strands produce high-scoring models. In 2018 this idea was extended to model longer and more complex structures called beta-serpentine, which consist of multiple beta-arches [99].

One of the important milestones in the amyloid hot spots predictions was the release of Waltz-DB and AmyLoad databases in 2015. For the first time, a large collection of experimentally studied peptides were easily available. This led to the development of a number of more sophisticated models. One such model was APPNN - Amyloidogenicity Propensity Prediction Neural Network [100]. It used many physicochemical properties of amino acids, which were used as an input for the artificial neural network. Soon after, the AmyloGram was developed. This method translates the query sequence to a reduced alphabet representation, in which every amino acid is assigned to one of six groups. Using a smaller alphabet a n-gram enabled the use of n-gram analysis. The frequency vector of n-gram occurrence is used as an input for the random forest classifier [101].

Currently, many new methods combine statistical or machine learning methods with physicochemical or structural information. One such technique is AgMata [102] developed in 2020. This approach combines custom-made statistical potential with unsupervised machine learning. Another example can be developed as a part of this thesis PATH. PATH is a method inspired by a classical 3D profiles method, but it uses a collection of templates representing different possible structural classes of amyloid fiber cores. Unlike 3D profiles, PATH operates on a number of energy functions that are used as an input for the machine learning model [103]. A similar approach was proposed by the authors of Cordax [104]. Nevertheless, also methods based on classical machine learning algorithms are still being developed, taking advantage of a growing number of available data. For example, the Budapest [105] method is based on supported vector machines while ENTAIL [106] uses Naive Bayes Classifier.

Although most of the available predictors use sequences as an input, there is also a group of methods that operates on protein structures. The first of such methods was Aggrescan3D, introduced in 2015 as a web server [107] and later as a standalone package [108]. Recently, a new version of this software has been published [109]. The main goal of this method is to identify aggregation-prone regions that are exposed to the environment, and therefore have a good chance of driving the aggregation. This is done by identifying surface patches with high aggregation

propensity according to the Aggrescan scale. To account for protein flexibility, Aggrescan3D probes the possible conformations using CABSflex [110]. Another example of a structure-based method can be AggScore [111] introduced in 2018 as a part of Schrödinger’s BioLuminate Suite. It also scans the surface of the protein and identifies hydrophobic and charged patches and scores them.

3.9 Prediction of amyloidogenicity of an entire protein

All previously described sequence-based predictors aimed at the identification of amyloid hot spots by using a sliding window approach. The main drawback of this approach is that even for a medium-sized protein hundreds of short fragments need to be tested. Therefore, there is a high probability of obtaining false positive results. This poses challenges in utilizing such methods on a genome-wide scale. To overcome this problem, a number of predictors of amyloidogenicity of entire proteins were proposed. The first such method was RFAmyloid [112]. This tool extracts features of a query sequence and uses them as an input for the random forest classifier. Almost all later methods followed a similar path of feature extraction and machine learning. For example, iAMY-SCM [113] uses dipeptide composition combined with Scoring Card Method (SCM), AMYPred-FRL [114] applies a feature representation learning framework, and ECAMYloid [115] is based on ensemble learning.

3.10 Prediction of the mutation effects and aggregation kinetics

From the clinical point of view, a very important task is the prediction of the effects of mutations in amyloid proteins. Sometimes even small changes to the amino acid sequence can have profound effects on protein aggregation rate. A classical example can be amyloid beta, whose mutations are linked to numerous variants of familial Alzheimer’s disease [116].

Prediction of the mutation effects started from rather simple empirical models. In 2003 the first of such models was proposed by Chiti and coworkers [117]. It takes into consideration the changes in the physicochemical properties of a peptide, like hydrophobicity or charge, caused by a mutation. This model was further developed not only to account for the effects of mutations but also to predict the absolute aggregation kinetics by DuBay and coworkers [118]. As the field matured more sophisticated methods were proposed, including AmyloidMutants [119]. Using a statistical potential to model the effects of mutations on the stability and conformation of an amyloid fiber. However, this method does not provide information about the aggregation kinetics, which can be obtained using methods such as AggreRATE-Disc [120]. AggreRATE-Disc uses a machine-learning model to predict which mutations can enhance or inhibit the aggregation of a sequence of interest. In 2020 this model was significantly improved by including structure-based features and proposing different strategies for short peptides and proteins. The resulting method was called AggreRATE-Pred [78].

3.11 Other computational methods

The multitude of computational problems related to amyloid research led to the development of many unique methods that are difficult to classify. An example can be the Fibrilizer [121, 122] suite containing four modules CreateFibril, Fibril Mutant, MAPOR, and SEMBA, each dedicated to a different task. CreateFibril builds models of amyloid fibrils based on monomer or oligomer structure provided by the user. Fibril Mutant assesses the stability of structural models of amyloids. MAPOR can be used to analyze the effects of mutations, using energy functions approximating Lennard-Jones and Coulomb interactions, as well as solvation energy. A similar approach is utilized by SEMBA to analyze the binding affinity of amyloids.

Another interesting tool related to amyloid research is the probabilistic context-free grammar model for amyloid signaling motifs [123]. Probabilistic context-free grammars are natural language processing models which were previously applied to a number of biological problems [124]. Here, such a model was used to identify amyloid signaling motifs which play an important role in fungal immune response [53]. Such motifs are usually modeled using Hidden Markov Models (HMM). Unlike HMM, probabilistic context-free grammar makes no assumption about the evolutionary relationship of the sequences and can capture dependencies between distant positions in the sequence. This enables them to better generalize even over different motifs [123].

Finally, as a part of this thesis, I have developed PACT the first available method for the prediction of amyloid cross-interactions. PACT models the structure of an amyloid hetero-aggregate formed by two query peptides and assesses if such aggregate is energetically favorable.

Chapter 4


Thesis of this work

This dissertation is based on three theses:

1. Structural modeling combined with machine learning can improve the identification of amyloid-prone regions.
(Modelowanie strukturalne w połączeniu z metodami uczenia maszynowego poprawia skuteczność przewidywania fragmentów amyloidowych)
2. Searching for new amyloids on a genome-wide scale requires more specialized methods.
(Poszukiwanie amyloidów w skali genomowej wymaga wyspecjalizowanych metod.)
3. Structural modeling can be used for the prediction of amyloid cross-interaction.
(Modelowanie strukturalne pozwala przewidzieć krzyżowe interakcje amyloidów)

Chapter 5

Results

his dissertation is based on a set of six scientific articles and one preprint that is currently after the round of revisions (Major revision). Each of the following sections discusses the results published in the article with the same title:

1. **Path - prediction of amyloidogenicity by threading and machine learning.**

In this work I developed the methodology, implemented the algorithm, and tested the resulting method. Together with my supervisor, I analyzed the data and prepared the manuscript.

2. **Bioinformatics methods for identification of amyloidogenic peptides show robustness to misannotated training data**

In this work I performed bioinformatics analysis of sequences using a range of amyloid predictors. Also, I performed a dimensionality reduction on the set of IR spectra and took part in the analysis of results and writing the manuscript.

3. **Variability of Amyloid Propensity in Imperfect Repeats of CsgA Protein of Salmonella enterica and Escherichia coli**

In this work I performed bioinformatics analysis of CsgA fragments using a range of amyloid predictors, performed a dimensionality reduction on the set of IR spectra. Also, I performed FT-Raman measurements of fragments and took part in the analysis of results and writing the manuscript.

4. **A spatiotemporal reconstruction of the C. elegans pharyngeal cuticle reveals a structure rich in phase-separating proteins**

In this work I performed bioinformatics analysis of *C. elegans* proteome using a combination of AmyloGram and PATH. I also took part in the analysis of results and writing the manuscript

5. **Exploring a diverse world of effector domains and amyloid signaling motifs in fungal NLR proteins**

In this work I performed bioinformatics analysis including iterative homology search using HMM models. I filtered the results, extracted N and C-terminals, and performed a *denovo* motif search. I took part in the structural analysis of HeLo, Goodbye and MLKL-like domains. I also took part in the analysis of results and writing the manuscript.

6. **AmyloGraph: a comprehensive database of amyloid–amyloid interactions**

I took part in the preparation of curator’s form and performed data collection and data curation. I also took part in the testing of the database, analysis of results, and writing the manuscript.

7. **PACT - Prediction of Amyloid Cross-interaction by Threading** (Preprint published in bioRxiv, under revision in Scientific Reports)

In this work I developed the methodology, implemented the algorithm, and tested the resulting method. I took part in the development of a web server version of the tool. Together with my supervisor, I analyzed the data and prepared the manuscript.

5.1 **Path - prediction of amyloidogenicity by threading and machine learning.**

The currently available methods for the prediction of aggregation-prone regions still have many limitations. Although their accuracy significantly improved over the past two decades their specificity is still insufficient to use them in large-scale studies. Many such methods heavily rely on limited and biased training data, such models often lacking reliability and interpretability. Furthermore, almost all of them provide only a binary classification of whether the peptide of interest is likely to form amyloid fibers or not. However, more recent studies revealed a diversity of amyloid structures. In order to account for this variability and overcome the limitations of current methods I developed PATH - a new method for the identification of amyloid-prone regions was proposed. These results were published in (Wojciechowski and Kotulska 2020). There were three main goals I have in mind for developing PATH. First, to improve the accuracy and specificity of available methods. The second reason was to propose a method that will be capable of providing more detailed information about possible molecular structures formed by putative amyloidogenic fragments. Finally, we aimed to provide a highly interpretable model, which could provide new knowledge about molecular features of aggregation-prone regions. To achieve these goals we decided to build a physicochemical model involving template-based modeling. A similar approach was previously used by the authors of the 3D profiles method [84]. I extended the method by utilizing multiple modeling templates based on recently available structural data. Furthermore, the procedure was extended by introducing machine learning classification operating on features of obtained models.

In the first step of the proposed method, a putative amyloidogenic fragment is threaded on seven structural templates using Modeller software [125]. Each template represents different possible packing of peptides in the amyloid core [15] and consists of 12 polypeptide chains forming a steric zipper structure. Then, each generated model was scored using DOPE [126] statistical potential implemented in Modeller. This by itself allowed reasonable distinction between amyloids and non-amyloids but to further improve the performance of the method I decided to extend the procedure. For the model with the best score, or in other words, the most energetically favorable one, additional scoring functions were calculated using PyRosetta [127]. Finally, all these scoring functions were used as input to a machine learning classifier. I decided to test several classical machine learning algorithms including logistic regression,

support vector machine, or random forest algorithm implemented in Scikit-Learn [128] Python package.

To train and test the model I used 1080 hexapeptides from the WaltzDB database [74] which were split into training and test sets. Additional benchmarks were performed on the pep424 dataset used by authors of PASTA 2.0 [87], which is more balanced with respect to the number of positive and negative examples. Finally, to test the performance of structural class prediction, 24 amyloid fibers structures from PDB, assigned to specific structural classes were used.

Most of the tested models achieved comparable results, however the best, and the most robust classification was achieved using logistic regression. The procedure is summarized in Fig 5.1. To better understand the proposed model, I analyzed the regression coefficients and investigated which features are the most important for classification using the Boruta algorithm [129]. In general, both methods of analysis highlighted the importance of van der Waals interactions and interactions with solvents. PATH showed comparable performance to state-of-the-art methods like AmyloGram [101], PASTA 2.0 [87], or FoldAmyloid [92] (Table 5.1). The method achieved high values of the Area Under ROC (AUC) parameter of 0.88 and a very high specificity of 0.94. The high specificity of the method can limit the number of false positive results and therefore may enable the use of the method in larger-scale studies.

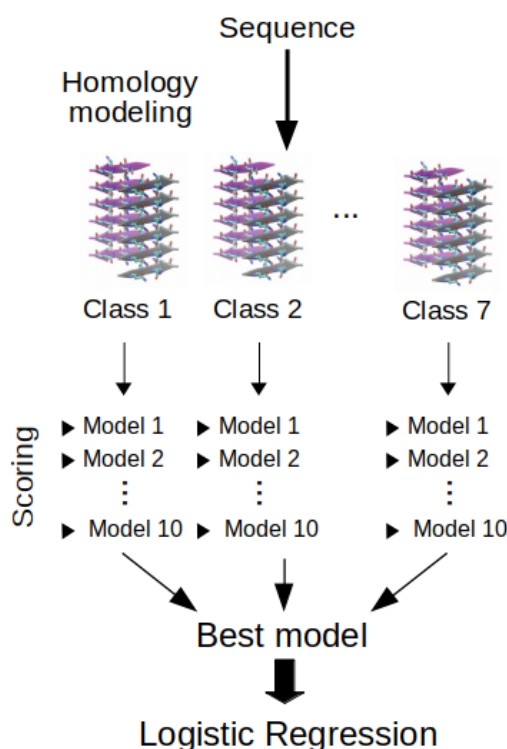


Figure 5.1: Prediction procedure. The query sequence is threaded on seven templates and for each of them ten models are made. The model with the lowest DOPE score is selected. For this model, additional energy terms are calculated and used as input for a machine learning classifier.

Unfortunately, we were not able to reliably classify structural classes of fibers.

Table 5.1: Comparison of PATH with several state of the art predictors of amyloidogenicity. PATH showed one of the best performances among tested methods.

| Method | AUC | Sensitivity | Specificity | MCC |
|-------------|------|-------------|-------------|------|
| PATH | 0.88 | 0.42 | 0.94 | 0.44 |
| PASTA 2.0 | 0.86 | 0.38 | 0.95 | 0.43 |
| AmyloGram | 0.89 | 0.68 | 0.90 | 0.61 |
| FoldAmyloid | 0.76 | 0.75 | 0.72 | 0.45 |

In this case, the one vs all accuracy of the method was around 0.46 which is below our expectation, but still much better than random. This is likely caused by the ability to form polymorphic structures by most amyloidogenic sequences. PATH provides all the structural models created during the modeling procedure, which can then be used for other structural analyses or molecular dynamics simulations. Our model not only achieved good results in identifying aggregation-prone sequences but also highlighted the role of van der Waals and solvent interactions in stabilizing aggregates.

PATH is available at <https://github.com/KubaWojciechowski/PATH>

5.2 Bioinformatics methods for identification of amyloidogenic peptides show robustness to misannotated training data

During the development of PATH we have encountered the same problem as the authors of AmyloGram [101]. Some of the peptides from the WaltzDB database [74, 75] repeatedly obtained very confident yet opposite classifications. 24 of these peptides along with 10 reference peptides with strong and correct classification were selected for more in-depth investigation. The procedure is summarized in Fig 5.2. The aggregation propensity of chosen peptides was assessed using a combination of IR spectroscopy (ATR-FTIR and IR microscope) and Atomic Force Microscopy (AFM) performed for selected peptides. For the reference peptides, obtained classification was in very good agreement with the experimental results and database annotation. For each of them, I performed amyloidogenicity predictions using among others PATH (Wojciechowski and Kotulska 2020), AmyloGram [101], PASTA 2.0 [87] and FoldAmyloid [92]. For the majority of the 24 remaining peptides, bioinformatic predictions matched experimental evidence but not database annotation. Although ambiguous cases were also present. I performed Principal Component Analysis (PCA) on IR spectra to provide a better classification of IR spectroscopy results.

Our study showed that indeed most of the 24 studied peptides were misannotated in the database. Although discovered peptides make up only around 1% of the whole Waltz database, this raised the question about the quality of available data and its effect on the performance of the bioinformatics software. Interestingly all of the mentioned bioinformatic tools classified them correctly despite being trained on them. This study shows the robustness of computational tools including PATH.

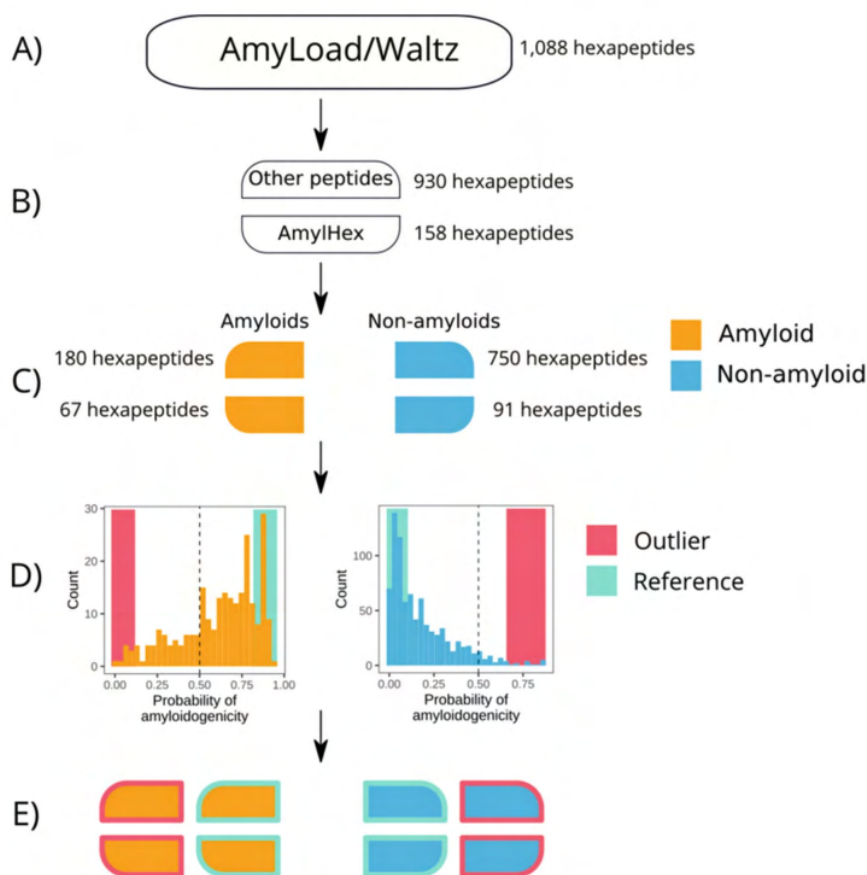


Figure 5.2: Procedure of selecting reference and outliers peptides. A) Hexapeptides are extracted from AmyLoad and Waltz databases. B) Peptides are divided into subsets and into C) aggregating and non-aggregating. D) For each sequence predictions are performed and E) peptides with extreme scores are chosen for experimental study.

The presence of misannotated training data can also explain why in the case of PATH the best classification was achieved using the simplest of tested classifiers - the logistic regression model. Logistic regression is a relatively simple model with a small number of parameters compared to other, more advanced machine learning classifiers. The results were published in (Szulc et al 2021a).

Both articles presented so far support the first thesis of this work which states that **Structural modeling combined with machine learning can improve the identification of amyloid-prone regions.**

5.3 Variability of Amyloid Propensity in Imperfect Repeats of CsgA Protein of *Salmonella enterica* and *Escherichia coli*

One of the important problems regarding the prediction of amyloid-prone regions is the identification of functional amyloids. This is a major limitation of all available methods due to the small number of well-studied sequences. Also, functional amy-

loids usually have different amino acid compositions than their pathological counterparts. Therefore, it is much more difficult to predict their aggregation propensity. To address this problem we decided to study CsgA proteins from two organisms *Escherichia coli* and *Salmonella enterica*. CsgA protein is a major component of curli fibers [44]. It consists of five imperfect repeats R1-R5 which, despite their sequence similarity, differ in their aggregation propensity [130, 131]. Understanding how these small sequential differences can change the behavior of repeated fragments can provide important insight into the mechanisms governing the aggregation of functional amyloids.

In the first step, I performed bioinformatics analysis of R1-R5 fragments of CsgA proteins from *E. coli* using PATH (Wojciechowski and Kotulska 2020), AmyloGram [101], PASTA 2.0 [87], FoldAmyloid [92], Waltz [93], AmylPred2 [95], MetAmyl [96] and ArchCandy [98]. Unfortunately, none of the available methods was able to correctly classify all fragments. In the next step, we extended our analysis to include CsgA from *S. enterica*. Fragments from both proteins were studied using a combination of vibrational spectroscopy and high-resolution imaging techniques. I developed the methodology of measurements and registered spectra using Fourier Transform Raman spectroscopy (FT-Raman) which to our best knowledge, has never been used to study functional amyloids before. This technique can provide complementary information to more widely used Infrared spectroscopy. For example, it allowed overcoming the problem of strong water absorbance in the Amide I region. This region is often used to study a secondary structure of peptides, but its use is limited when proteins and peptides are studied in aqueous solutions. A combination of IR and FT-Raman spectroscopy enabled us to better assign secondary structures, and get more insight into the structures of aggregates. Finally, I performed Principal Component Analysis (PCA) on the set of IR spectra. By analyzing which wavelengths contributed to the first three principal components, I was able to identify bands important for distinguishing between different groups of spectra.

In this work, not only we characterized CsgA protein from *S. enterica*, but we also highlighted the problem of prediction of functional amyloids. In general, CsgA produced by *S. enterica* was shown to be much more aggregation-prone than its counterpart from *E. coli*. We also discussed the possible role of selected residues changing among different repeated fragments and between organisms. Based on these results we proposed the role of several amino acids differing among sequences. Especially the location of charged residues seems to be an important factor. The obtained results lead to the formulation of new hypotheses on the mechanisms of aggregation of this class of proteins, which were later studied in more detail by analysis of mutated versions of the peptides. These results are submitted for publication.

5.4 A spatiotemporal reconstruction of the *C. elegans* pharyngeal cuticle reveals a structure rich in phase-separating proteins

At the next step of my research, I started a collaboration with Prof. Peter Roy from the University of Toronto. His team was studying the development of pharyngeal cuticle of *Caenorhabditis elegans*, which is a flatworm used as a model organism in neuroscience [132] neurodegeneration [133, 134], immunity [135], toxicology [136], as

well as microbiome research [137]. During their research, they observed structures that bind Congo Red and Thioflavin T, dyes commonly used for the identification of amyloids [138, 139]. Furthermore, their analysis of gene expression through the molting stage revealed the presence of proteins capable of regulating amyloid aggregation including Neprilysin, and ITM-2. The first protein was shown to play an important role in the degradation of amyloid-beta [140] while the second is a molecular chaperon containing the BRICHOS domain which can inhibit the secondary nucleation of amyloids [141]. This naturally leads to a question about the presence and potential role of amyloids in *C. elegans* development. To address this question I scanned the whole proteome using AmyloGram and PATH. I decided to use two different methods to minimize the false positive rate. In the first step, all proteins were scanned using AmyloGram, and the identified amyloid-prone regions were checked using PATH. This approach enables the utilization of advantages of both methods - the high performance of AmyloGram and the high specificity of PATH.

After applying only an AmyloGram software, amyloid-prone regions were identified in around 35% of proteins (Fig 5.3A). The additional filtration with PATH decreased this number to about 26% (Fig 5.3B). Despite the identification of hot spots in such a large number of sequences, there was no enrichment of amyloids in pharynx proteins compared to other tissues. The most significant differences were observed between secreted and non-secreted proteins. The second group on average contained more amyloidogenic fragments.

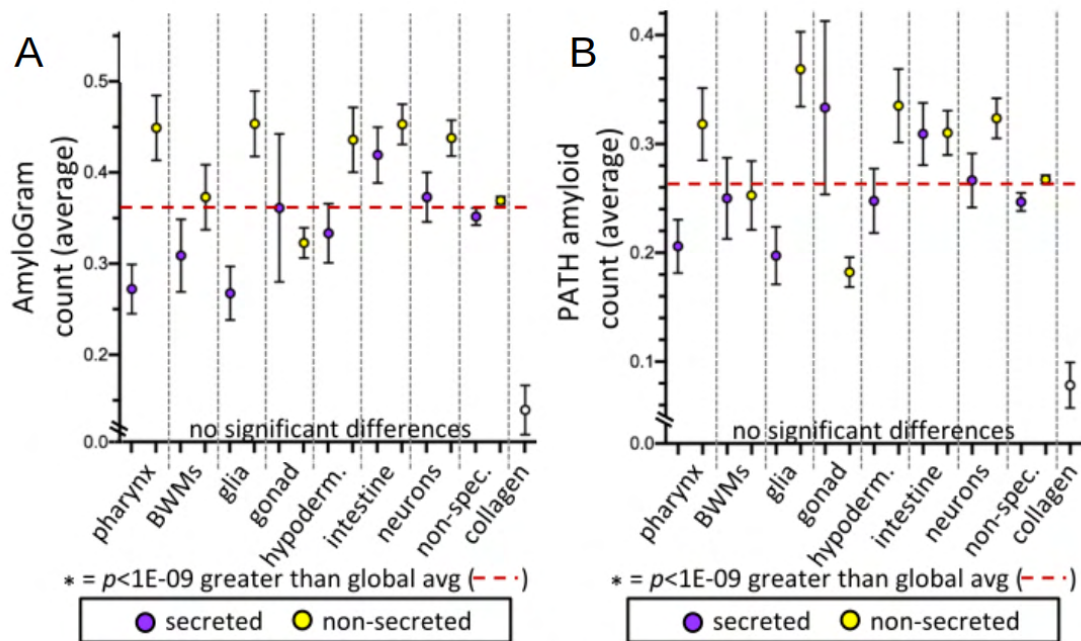


Figure 5.3: Fractions of proteins with identified amyloid hot spots in different tissues.

Despite many clues suggesting the involvement, or at least the presence of amyloids in *C. elegans* pharynx, no fibrillar structures were observed. There are several possible explanations for those seemingly contradictory results. First of all, amyloid-prone regions are very common in many proteomes but their presence is

not necessarily sufficient to drive the aggregation of the whole protein [142]. This is especially the case when they are buried in the hydrophobic core of the protein [85]. The smaller abundance of amyloid hot spots in secreted proteins might be related to the fact that such proteins need to operate outside the well-controlled environment of the cell. Therefore, any possible aggregation-prone region is more likely to be exposed due to external factors such as varying pH, ionic strength, or the presence of other denaturing factors. Hence, there might exist an evolutionary pressure to reduce the number of amyloidogenic regions. This however does not explain the expression of amyloid-regulating proteins. It is therefore possible, that amyloids are produced, but their aggregation can be inhibited by the presence of a large number of intrinsically disordered proteins, which were shown to inhibit amyloid aggregation [143].

These results show that the identification of amyloid-prone regions is not sufficient for the reliable identification of potential novel amyloids. Especially, considering that most of the currently available computational methods scan the query sequence with a short, typically six amino acids long sliding window. This means that for 300 amino acids long proteins $300 - 6 = 294$ tests are performed. Even when such software is capable of achieving a very high specificity of 0.99 we can still expect around three false positive hits.

Although there was no enrichment of amyloids in the pharynx, the enrichment of Low-complexity, Aromatic-Rich, Kinked Segments (LARKS) was detected. These structures closely resemble amyloids. They form reversible aggregates, which were shown to play a role in the formation of membrane-less organelles [144].

5.5 Exploring a diverse world of effector domains and amyloid signaling motifs in fungal NLR proteins

Considering the limitations of available predictors discussed in previous works, I aimed to explore other methods of amyloid identification. I started collaborating with Dr. Witold Dyrka, and we aimed to identify novel functional amyloids in fungi. We decided to focus on NLR proteins that are a vital component of animal, plant, and fungal immune systems. These are immune receptors that trigger a number of host responses including cell death [145, 146]. A very peculiar characteristic of fungal NLR architectures is their ability to use amyloid aggregation to propagate the signal. This is realized through the presence of Amyloid Signaling Motifs (ASM) [147]. One of the best-studied implementations of this system is HET-S system of a model fungi *Podospora anserina*. Here NLR protein - NWD2 contains an amyloidogenic region in its N-terminal domain that can induce aggregation of the C-terminal of HET-S. This initiates a series of events leading to cell death [148]. Considering the crucial role of amyloids in this system, it is reasonable to assume that their aggregation properties should be fine-tuned through the eons of evolution. This makes them a very promising model for studying functional amyloids. Unfortunately, despite intensive research, only a few of them are well described.

To better understand ASM we first need to better understand the context in which they operate. In the first step, we aimed at improving annotations of N-terminals from proteins containing known NLR domains (NACHT or NB-ARC do-

mains). To do so we clustered terminals longer than 20aa using MMseqs2 [149] and for clusters containing at least 20 representants we searched for homologs using HHblitz [150]. This procedure improved the annotation coverage from 57% to 66%. However, this procedure worked poorly for relatively short N-termini (below 150 amino acids). As in this set of sequences, we identified some of known ASM we decided to focus on this group to search for new motifs.

In the next step, I aimed to identify a set of motifs that appears both in the previously mentioned N-terminals and C-terminals of proteins containing effector domains. The main assumption behind this approach is that in this way I should be able to identify only functional motifs involved in signaling and not just any aggregation-prone regions. To obtain a set of N and C-terminals I started from a set of proteins with (NACHT or NB-ARC) domains and known effector domains. I searched NCBI's "nr" sequences database [151] using HMMER 3.2.1 and for both sets, I extracted N and C-terminals respectively. I filter the sets by length to include only fragments between 10 and 150 amino acids and cluster them at 70% of identity using CD-HIT [152] software. In both sets, I searched for sequential motifs using MEME software [153] and built HMM profiles, which were then used to check for the presence of motifs in the opposite set (C-terminal motifs in a set of N-terminals and vice versa). Furthermore, it was verified if the matched motifs were located in genomic proximity. This procedure led to the identification of 22 motifs which were then clustered and compared with previously known ASM (Fig ??). One of them was found to be significantly different from all previously identified motifs. It exclusively occurred in proteins containing the PNP_UDP domain, therefore we called it PUASM for Pnp_Udp-associated Amyloid Signaling Motif.

The newly characterized motif was then studied in more detail using vibrational spectroscopy, ThT assay, and Atomic Force Microscopy. Experiments performed on a selected pair of representative sequences confirmed that this motif is capable of forming amyloid aggregates in vitro. Furthermore, detailed spectroscopic characterization revealed that it is likely to adopt a beta helical structure, characteristic of many previously characterized signaling amyloids including HET-S. Our French collaborators confirmed the prion-like behavior of this motif in vivo.

This work resulted not only in the improved annotation of fungal NLR proteins by identification of functional domains and novel ASM but most importantly proposes a new approach for the identification of signaling amyloids.

Three described above articles (Szulc et al 2021, Kamal et al. 2022, and Wojciechowski et al. 2022) among others explore different aspects of amyloidogenicity prediction. The first of them explore the subtle details that can dramatically affect the aggregation propensity of functional amyloids, as well as provides examples of sequences that are poorly classified by existing amyloidogenicity predictors. The second of them highlights the difficulties and potential pitfalls of directly applying bioinformatic predictors on large datasets. It also shows that even the predicted presence of amyloid-prone regions not always will result in the formation of amyloid fibrils. Finally, the third one shows the benefits of using a dedicated procedure for the identification of functional amyloids. The results published in those three articles support the second thesis of this work which states that **Searching for new amyloids on a genome-wide scale requires more specialized methods.**

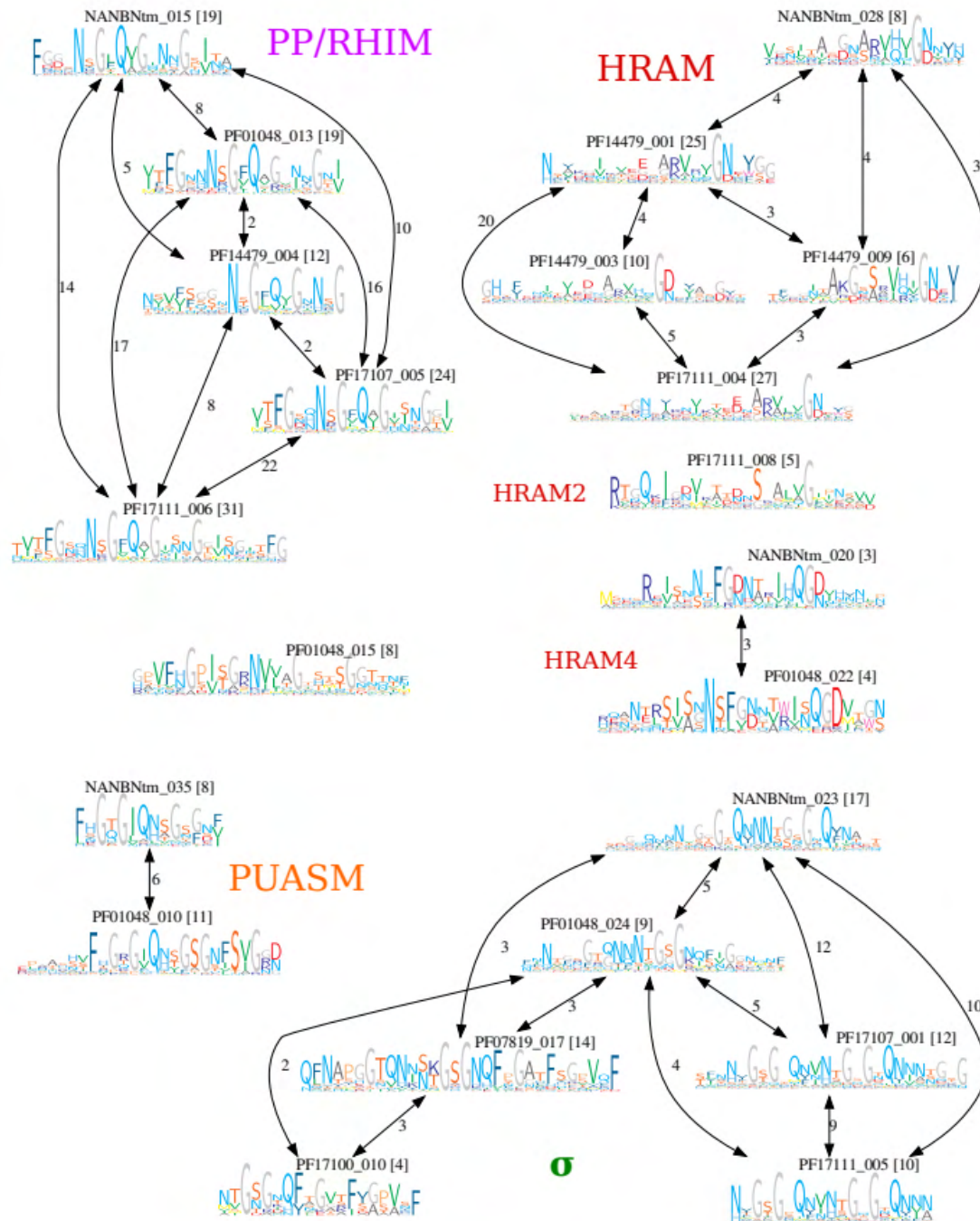


Figure 5.4: Clustering amyloid identified signaling motifs. The number in brackets represents the number of occurrences of the motif. If two motifs fitted the same sequence (significant e – value) they were connected with an edge. The number on the edge indicates the number of such sequences for which two different motifs were aligned.

5.6 AmyloGraph: a comprehensive database of amyloid–amyloid interactions

The last aim of my PhD thesis was to propose the first computational method of amyloid cross-interactions. This was by far the most ambitious part of this work

since there was no previous research on this topic and the available data were quite limited and scattered across a large number of research articles. Therefore to even start thinking about building a tool it was necessary to build a reliable database of amyloid-amyloid interactions. We assembled a team of curators under the lead of Dr. Michał Burdukiewicz. Very quickly we realized that there is no consensus regarding the terminology used to describe amyloid-cross interactions. We considered different possible modes of interaction and proposed a standardized terminology. Based on that we came up with detailed forms containing a number of questions that helped us systematically describe almost 900 pairs of interactions scattered across almost 200 articles.

Data are presented in the form of a graph where nodes represent amyloid proteins and edges interactions between them (Fig 5.5). Among others, users can filter visible interactions by the protein of interest or type of interaction including slower and faster aggregation, formation of heterogeneous fibrils, etc. These results were published in [81].

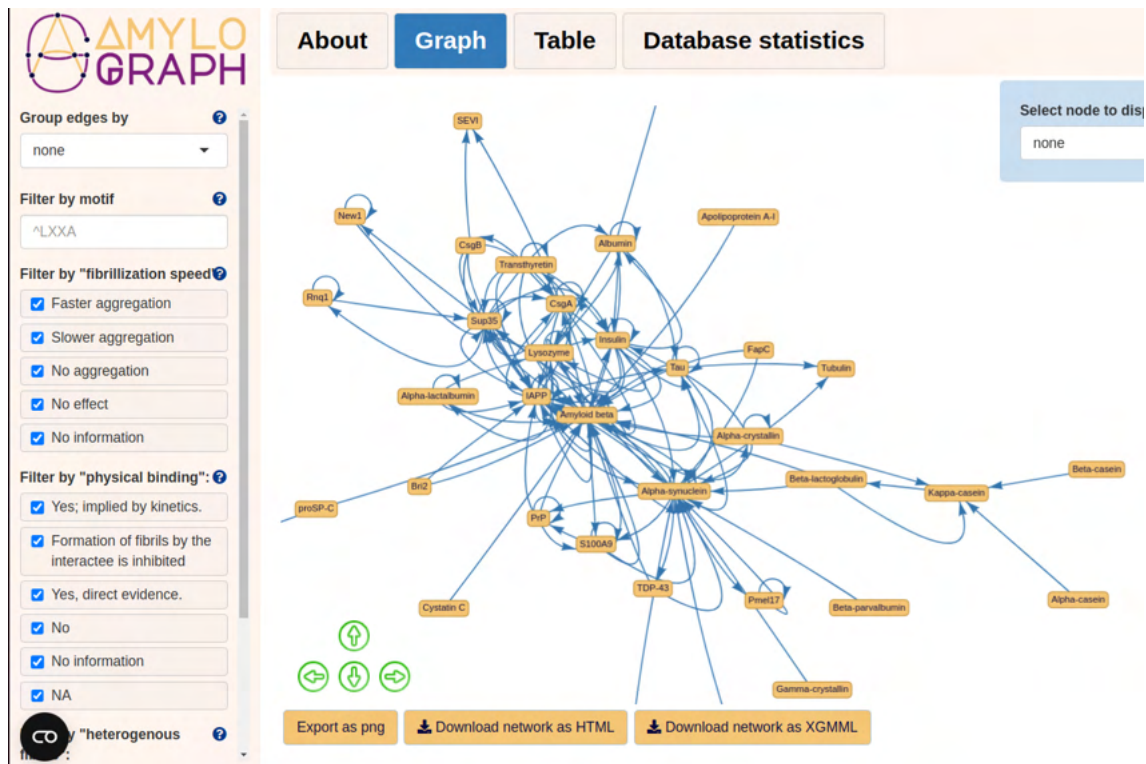


Figure 5.5: Fragment of an AmyloGraph interface. The edges of the graph represent amyloid proteins and the edges - interactions between them.

AmyloGraph is publically available as an online database at <http://AmyloGraph.com/> and R package at <https://github.com/KotulskaLab/AmyloGraph>.

5.7 PACT - Prediction of Amyloid Cross-interaction by Threading

After building AmyloGraph I was ready to pursue the last goal of this project. I aimed to develop a computational method for amyloid-amyloid interaction predic-

tion. After analyzing the content of AmyloGraph I quickly realized that there was a relatively small number of unique proteins available in the database and a few of them are heavily overrepresented. These issues significantly restricted the use of statistical or machine learning models, therefore I decided to build a physicochemical model involving template-based modeling. One of the first problems was that most sequences from AmyloGraph were much longer than typical amyloid hot spots. Furthermore, the prediction of interactions between amyloids requires modeling interaction between two different sequences, that often differ in length.

I decided to use a structure of hIAPP as a template, which is a 37 amino acids long peptide, and at that time was one of the longest amyloidogenic fragments with known detailed molecular structure. To allow for different lengths of interacting peptides, both sequences were aligned to the center of the template (Fig 5.6A). In such a way if any of the sequences is shorter than a template, only a part of the template will be used for modeling. To model cross-interactions both sequences are threaded onto the same template at once. In such a way structure of a hetero aggregate is obtained (Fig 5.6B). Same as in the case of PATH I decided to use Modeller software [125] for modeling. I tested a version using a single template structure as well as multiple templates.

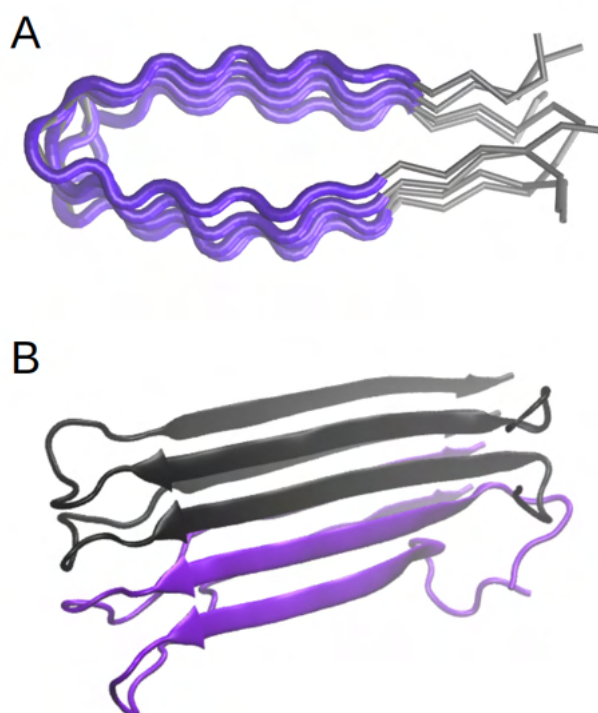


Figure 5.6: A) Values of ndope score for amyloidogenic and non-amyloidogenic peptides from AmyLoad database. B) ROC for amyloid non-amyloid classification.

Since the use of multiple templates does not improve the classification performance significantly, yet dramatically increased the computational cost, I decided to use a single template variant. Another problem that I faced was that DOPE statistical potential provided a different range of values for different sequence lengths.

Therefore, I decided to normalize the score by dividing it by the average length of modeled sequences.

First, I tested if the proposed method is capable of distinguishing between amyloidogenic and non-amyloidogenic peptides. Figure 5.7A shows a difference in obtained normalized DOPE score (ndope). In general, amyloidogenic sequences obtained lower scores than non-aggregating ones. This allows setting a score threshold for classification. To do so I plotted a ROC curve and chose a point closest to 0 False Positive Rate (FPR) and 1 True Positive Rate (TPR) (Fig 5.7 B). Such a classifier achieved an Accuracy of 0.77 and high values of Sensitivity (0.73) and Specificity (0.86). This proves that such a method can distinguish amyloids from non-amyloids.

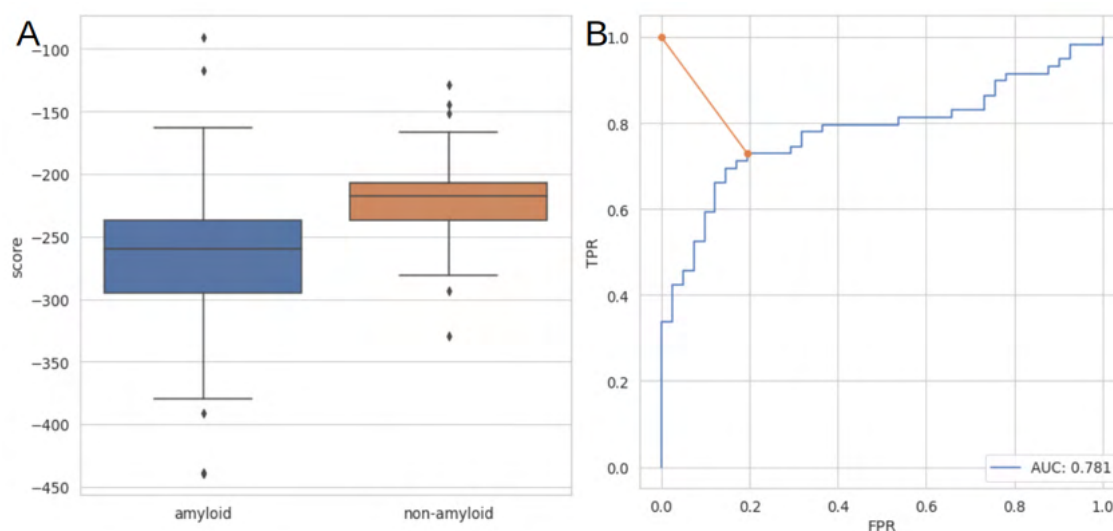


Figure 5.7: A) Values of ndope score for amyloidogenic and non-amyloidogenic peptides from AmyLoad database. B) ROC for amyloid non-amyloid classification.

Considering previously identified problems with the identification of functional amyloids I decided to try this method on a set of CsgA fragments from *E. coli* and *S. enterica* studied experimentally in our previous project. On this set very small set method achieved an accuracy of 0.9 and to my best knowledge, it is currently the only method that worked well on this dataset.

However, the main aim of this project was to predict cross-interactions. In order to do so the same methodology was applied, the only difference being that now pairs of interacting sequences from AmyloGraph were used to train the model. The greatest problem I faced in this part of the project was the lack of a good negative dataset. It was almost impossible to find any reported cases of non-interacting pairs of amyloids as such findings are rarely proven and reported in the literature. Although there are pairs of amyloids labeled as non-interacting it turns out that the same pairs in most cases have other labels such as faster or slower aggregation. These cases predominantly refer to situations where no interactions were detected in specific conditions for example a very low concentration of one of the proteins. Therefore I needed to build a different negative set. To do so, I decided to use non-amyloidogenic sequences from the AmyLoad database. A closer look at this dataset reveals that it is mostly composed of peptides with strong beta propensity used by authors of the Tango method [82].

I have trained the method using pairs of interactions resulting in faster aggregation vs negative set and pairs of interactions slowing down aggregation vs negative set. Figure 5.8 shows the resulting scores. Same as in the previous case, a threshold-based classifier was built. The resulting predictor achieved an accuracy of 0.83 along with good sensitivity (0.78) and specificity (0.88).

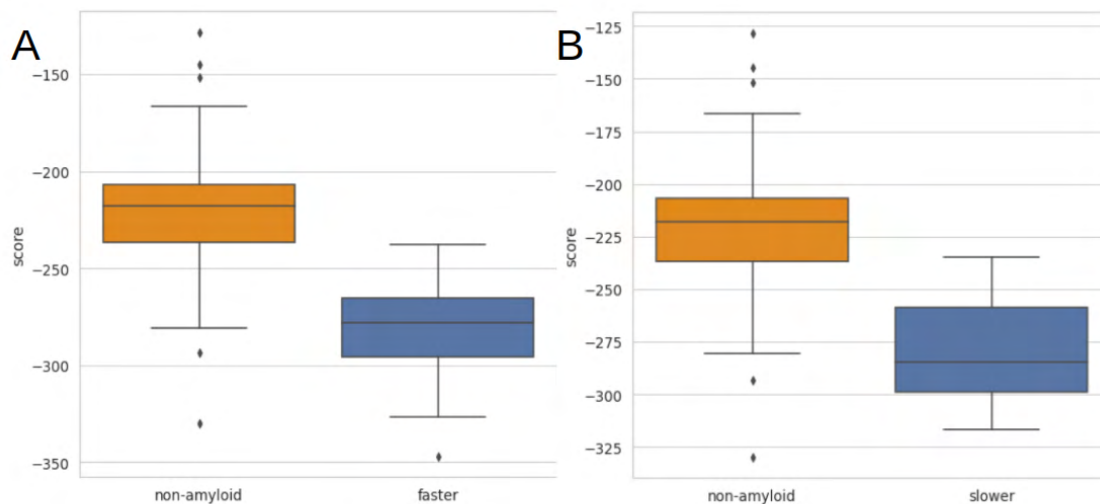


Figure 5.8: Values of ndope score for non-amyloids and A) pairs of peptides resulting in faster and B) slower aggregation.

The resulting method was called PACT (Prediction of Amyloid Cross-interactions by Threading) and it is publicly available as a standalone Python program at the GitHub repository at: <https://github.com/KubaWojciechowski/PACT> and as a web server at: <https://pact.e-science.pl/pact/>

As an additional test, and the showcase of the use of PACT I used it to study interactions between CsgA proteins from several microbial species inhabiting the human gut, which were recently studied experimentally [154]. Results obtained from PACT corresponded very well with the experiment. Finally, we decided to use PACT to predict the interactions of CsgA fragments with hIAPP. Interactions between these two proteins were previously reported in the literature, but the detailed mechanism of the interactions still remains unclear. We decided to shed some light on it by identifying which region of CsgA is most likely to interact with hIAPP. To do so I first performed a prediction of cross-interaction between R1-R5 fragments and hIAPP. Based on the results we chose two of them (R1 and R5) for seeding experiments, and experimentally confirmed that the R1 fragment can enhance the aggregation of hIAPP.

The results of the use of the PACT method prove the final thesis of this work which states that **Structural modeling can be used for the prediction of amyloid cross-interaction.**

5.8 Summary and Discussion

Computational tools have become a vital part of modern biological sciences, supporting experimental efforts to understand life at different scales. The main goal of

this work was to develop bioinformatics methods and protocols for studying amyloids, their interactions and explore their potential role in living organisms. First, a new method of identifying aggregation-prone regions, called PATH, was developed. The proposed method achieved good classification performance, comparable with other state-of-the-art techniques. Unlike the previous attempts at this task PATH aimed not only at identifying new fragments but also at providing more structural and physicochemical information about amyloid structures. The true test of the robustness of this tool was the study of mislabeled peptides found in publicly available databases. PATH, as well as many other computational tools, was able to correctly identify such problematic sequences and classify them correctly despite being trained on them. These results show that the currently available methods can be reliably used to identify amyloidogenic peptides. However, there are still some areas where the performance of such tools can be improved.

All of the tested software failed to correctly identify aggregation-prone regions of CsgA bacterial functional amyloid. To better understand the aggregation of this protein and why it is such a difficult example for amyloid predictors, fragments R1-R5 from two bacterial species were synthesized and characterized using a range of experimental techniques. Obtained results showed that even very few mutations can drastically alter the aggregation properties of functional amyloids. This highlights the incredible precision developed by evolution and might explain why such patterns are so difficult to capture by prediction software. Another area that can be improved is the use of amyloid predictors on a genome-wide scale. Currently, all of the methods using the sliding window approach will most likely produce a significant number of false positive results. Recently a number of predictors aiming at predicting the aggregation propensity of whole proteins have been proposed however their accuracy is comparable to sliding window-based methods. As a result, the use of such techniques in large-scale studies is limited. In recent years our understanding of amyloids drastically changed. Discoveries of LARKS and alpha-amyloids opened whole new areas of research and posed new challenges on amyloid predictors. The study of amyloids in *C. elegans* proteome, apart from providing an important insight into the development of pharyngeal cuticle, highlighted the role of LARKS and phase separation. At the same time, it showed the limitations of currently available amyloid predictors.

Therefore we are in need of more specialized approaches for identifying new amyloids. An example of such a procedure was proposed in the study regarding functional amyloids in NLR proteins. In this study, the possible search space was drastically reduced by focusing on a single group of proteins that were previously known to contain amyloid signaling motifs. We were able to take advantage of this knowledge to identify a number of amyloid signaling motifs in short terminals of NLR including a new one found exclusively in proteins containing the PNP_UDP domain. These results not only lead to a better understanding of the fungal immune system but also showed a complementary approach to identifying novel functional amyloids.

The growing body of research highlights the importance of amyloid interactions. From amyloid signaling in fungi and filamentous bacteria to comorbidities of human disease-related amyloids such interactions seems to be an important factor in the aggregation process. To paint the bigger picture of such interactions for the first time the literature investigating cross-interactions gathered into the AmylGraph

database. This enabled the building of the first amyloid interaction prediction tool - PACT. The new method was shown not only to be capable of accurately identifying cross-interactions but also was able to identify functional amyloids.

New ways of identifying amyloids and their cross-interactions can shed new light on the origin of amyloid-related disorders as well as explain numerous biological processes where amyloids can play a pivotal role. A better understanding of pathological and functional aggregation may lead to the development of new drugs and strategies for combating biofilm formation by pathogenic microorganisms. The incredible properties of amyloids can be also utilized by scientists and engineers to produce novel biomaterials and self-assembling nanostructures. However, to achieve these goals, a deep understanding of mechanisms governing aggregation and interactions of this class of protein will be necessary.

Bibliography

- [1] Mathias Uhlén et al. “Tissue-based map of the human proteome”. In: *Science* 347.6220 (2015), p. 1260419.
- [2] Christian B Anfinsen. “Principles that govern the folding of protein chains”. In: *Science* 181.4096 (1973), pp. 223–230.
- [3] Jeremy M Berg, John L Tymoczko, and Lubert Stryer. *Biochemistry (Loose-Leaf)*. Macmillan, 2007.
- [4] Giovanni B Frisoni et al. “The probabilistic model of Alzheimer disease: the amyloid hypothesis revised”. In: *Nature Reviews Neuroscience* 23.1 (2022), pp. 53–66.
- [5] Jean D Sipe and Alan S Cohen. “History of the amyloid fibril”. In: *Journal of structural biology* 130.2-3 (2000), pp. 88–98.
- [6] Sylvain Lesné et al. “A specific amyloid- β protein assembly in the brain impairs memory”. In: *Nature* 440.7082 (2006), pp. 352–357.
- [7] Charles Piller. “Blots on a field?” In: *Science (New York, NY)* 377.6604 (2022), pp. 358–363.
- [8] Merrill D Benson et al. “Amyloid nomenclature 2020: update and recommendations by the International Society of Amyloidosis (ISA) nomenclature committee”. In: *Amyloid* 27.4 (2020), pp. 217–222.
- [9] Shon A Levkovich, Ehud Gazit, and Dana Laor Bar-Yosef. “Two decades of studying functional amyloids in microorganisms”. In: *Trends in Microbiology* 29.3 (2021), pp. 251–265.
- [10] William Thomas Astbury, Sylvia Dickinson, and Kenneth Bailey. “The X-ray interpretation of denaturation and the structure of the seed globulins”. In: *Biochemical Journal* 29.10 (1935), p. 2351.
- [11] Sorin Luca et al. “Peptide conformation and supramolecular organization in amylin fibrils: constraints from solid-state NMR”. In: *Biochemistry* 46.47 (2007), pp. 13505–13522.
- [12] O Sumner Makin et al. “Molecular basis for amyloid fibril formation and stability”. In: *Proceedings of the National Academy of Sciences* 102.2 (2005), pp. 315–320.
- [13] Jeffrey F Smith et al. “Characterization of the nanoscale properties of individual amyloid fibrils”. In: *Proceedings of the National Academy of Sciences* 103.43 (2006), pp. 15806–15811.

- [14] Melinda Balbirnie, Robert Grothe, and David S Eisenberg. “An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated β -sheet structure for amyloid”. In: *Proceedings of the National Academy of Sciences* 98.5 (2001), pp. 2375–2380.
- [15] David S Eisenberg and Michael R Sawaya. “Structural studies of amyloid proteins at the molecular level”. In: *Annual review of biochemistry* 86 (2017), pp. 69–95.
- [16] Matthew G Iadanza et al. “A new era for understanding amyloid structures and disease”. In: *Nature Reviews Molecular Cell Biology* 19.12 (2018), pp. 755–773.
- [17] Alois Alzheimer. “Über eigenartige Erkrankung der Hirnrinde”. In: *All Z Psychiatr* 64 (1907), pp. 146–148.
- [18] Tara L Spires-Jones and Bradley T Hyman. “The intersection of amyloid beta and tau at synapses in Alzheimer’s disease”. In: *Neuron* 82.4 (2014), pp. 756–771.
- [19] Alois Alzheimer. “Über eigenartige Krankheitsfalle des späteren Alters”. In: *Psychiatr Nervenkr Z Gesamte Neurol Psychiatr* 4 (1911), pp. 356–85.
- [20] FH Lewy and M Lewandowsky. “Handbuch der Neurologie”. In: *Berlin: Julius Springer* (1912).
- [21] Maria Grazia Spillantini et al. “ α -Synuclein in Lewy bodies”. In: *Nature* 388.6645 (1997), pp. 839–840.
- [22] Gillian Bates. “Huntingtin aggregation and toxicity in Huntington’s disease”. In: *The Lancet* 361.9369 (2003), pp. 1642–1644.
- [23] Jennifer Stine Elam et al. “Amyloid-like filaments and water-filled nanotubes formed by SOD1 mutant proteins linked to familial ALS”. In: *Nature Structural & Molecular Biology* 10.6 (2003), pp. 461–467.
- [24] Markus Glatzel et al. “Extraneural pathologic prion protein in sporadic Creutzfeldt–Jakob disease”. In: *New England Journal of Medicine* 349.19 (2003), pp. 1812–1820.
- [25] Kenneth H Johnson et al. “Islet amyloid, islet-amyloid polypeptide, and diabetes mellitus”. In: *New England Journal of Medicine* 321.8 (1989), pp. 513–518.
- [26] Juliette Janson et al. “The mechanism of islet amyloid polypeptide toxicity is membrane disruption by intermediate-sized toxic amyloid particles.” In: *Diabetes* 48.3 (1999), pp. 491–498.
- [27] CB Lücking and A Brice*. “Alpha-synuclein and Parkinson’s disease”. In: *Cellular and Molecular Life Sciences CMLS* 57 (2000), pp. 1894–1908.
- [28] Robert M Koffie et al. “Oligomeric amyloid β associates with postsynaptic densities and correlates with excitatory synapse loss near senile plaques”. In: *Proceedings of the National Academy of Sciences* 106.10 (2009), pp. 4012–4017.
- [29] Cameron Wells et al. “The role of amyloid oligomers in neurodegenerative pathologies”. In: *International Journal of Biological Macromolecules* 181 (2021), pp. 582–604.

- [30] Manuela López de la Paz and Luis Serrano. “Sequence determinants of amyloid fibril formation”. In: *Proceedings of the National Academy of Sciences* 101.1 (2004), pp. 87–92.
- [31] Rebecca Nelson et al. “Structure of the cross- β spine of amyloid-like fibrils”. In: *Nature* 435.7043 (2005), pp. 773–778.
- [32] Kiril Tsemekhman et al. “Cooperative hydrogen bonding in amyloid formation”. In: *Protein science* 16.4 (2007), pp. 761–764.
- [33] David Eisenberg and Mathias Jucker. “The amyloid state of proteins in human diseases”. In: *Cell* 148.6 (2012), pp. 1188–1203.
- [34] Ehud Gazit. “A possible role for π -stacking in the self-assembly of amyloid fibrils”. In: *The FASEB Journal* 16.1 (2002), pp. 77–83.
- [35] Peleg Ragonis-Bachar and Meytal Landau. “Functional and pathological amyloid structures in the eyes of 2020 cryo-EM”. In: *Current Opinion in Structural Biology* 68 (2021), pp. 184–193.
- [36] Lothar Gremer et al. “Fibril structure of amyloid- β (1–42) by cryo-electron microscopy”. In: *Science* 358.6359 (2017), pp. 116–119.
- [37] Yaowang Li et al. “Amyloid fibril structure of α -synuclein determined by cryo-electron microscopy”. In: *Cell research* 28.9 (2018), pp. 897–903.
- [38] Liisa Lutter, Liam D Aubrey, and Wei-Feng Xue. “On the structural diversity and individuality of polymorphic amyloid protein assemblies”. In: *Journal of Molecular Biology* 433.20 (2021), p. 167124.
- [39] Liam D Aubrey et al. “Quantification of amyloid fibril polymorphism by nano-morphometry reveals the individuality of filament assembly”. In: *Communications Chemistry* 3.1 (2020), p. 125.
- [40] Yang Shi et al. “Structure-based classification of tauopathies”. In: *Nature* 598.7880 (2021), pp. 359–363.
- [41] Nir Salinas et al. “The amphibian antimicrobial peptide uperin 3.5 is a cross- α /cross- β chameleon functional amyloid”. In: *Proceedings of the National Academy of Sciences* 118.3 (2021), e2014442118.
- [42] Peleg Ragonis-Bachar et al. “Natural antimicrobial peptides self-assemble as α/β chameleon amyloids”. In: *Biomacromolecules* 23.9 (2022), pp. 3713–3727.
- [43] Han AB Wösten and Marcel L de Vocht. “Hydrophobins, the fungal coat unraveled”. In: *Biochimica et Biophysica Acta (BBA)-Reviews on Biomembranes* 1469.2 (2000), pp. 79–86.
- [44] Matthew R Chapman et al. “Role of Escherichia coli curli operons in directing amyloid fiber formation”. In: *Science* 295.5556 (2002), pp. 851–855.
- [45] Morten S Dueholm et al. “Functional amyloid in Pseudomonas”. In: *Molecular microbiology* 77.4 (2010), pp. 1009–1020.
- [46] Diego Romero et al. “Amyloid fibers provide structural integrity to Bacillus subtilis biofilms”. In: *Proceedings of the National Academy of Sciences* 107.5 (2010), pp. 2230–2234.

- [47] Anirudha Dutta et al. “Macroscopic amyloid fiber formation by staphylococcal biofilm associated SuhB protein”. In: *Biophysical chemistry* 217 (2016), pp. 32–41.
- [48] Kelly Schwartz et al. “Functional amyloids composed of phenol soluble modulins stabilize *Staphylococcus aureus* biofilms”. In: *PLoS pathogens* 8.6 (2012), e1002744.
- [49] Agustina Taglialegna et al. “Staphylococcal Bap proteins build amyloid scaffold biofilm matrices in response to environmental signals”. In: *PLoS pathogens* 12.6 (2016), e1005711.
- [50] Alexander E Yarawsky et al. “The biofilm adhesion protein Aap from *Staphylococcus epidermidis* forms zinc-dependent amyloid fibers”. In: *Journal of Biological Chemistry* 295.14 (2020), pp. 4411–4427.
- [51] Shady Saad et al. “Reversible protein aggregation is a protective mechanism to ensure cell cycle restart after stress”. In: *Nature cell biology* 19.10 (2017), pp. 1202–1213.
- [52] Luke E Berchowitz et al. “Regulated formation of an amyloid-like translational repressor governs gametogenesis”. In: *Cell* 163.2 (2015), pp. 406–418.
- [53] Jakub W Wojciechowski et al. “Exploring a diverse world of effector domains and amyloid signaling motifs in fungal NLR proteins”. In: *PLOS Computational Biology* 18.12 (2022), e1010787.
- [54] Witold Dyrka et al. “Identification of NLR-associated amyloid signaling motifs in bacterial genomes”. In: *Journal of molecular biology* 432.23 (2020), pp. 6005–6027.
- [55] Douglas M Fowler et al. “Functional amyloid formation within mammalian tissue”. In: *PLoS biology* 4.1 (2006), e6.
- [56] Jixi Li et al. “The RIP1/RIP3 necrosome forms a functional amyloid signaling complex required for programmed necrosis”. In: *Cell* 150.2 (2012), pp. 339–350.
- [57] Katarzyna Lundmark et al. “Protein fibrils in nature can enhance amyloid protein A amyloidosis in mice: Cross-seeding as a disease mechanism”. In: *Proceedings of the National Academy of Sciences* 102.17 (2005), pp. 6098–6102.
- [58] Kenjiro Ono et al. “Cross-seeding effects of amyloid β -protein and α -synuclein”. In: *Journal of neurochemistry* 122.5 (2012), pp. 883–890.
- [59] Rundong Hu et al. “Cross-seeding interaction between β -amyloid and human islet amyloid polypeptide”. In: *ACS chemical neuroscience* 6.10 (2015), pp. 1759–1768.
- [60] Ping Cao et al. “The ability of rodent islet amyloid polypeptide to inhibit amyloid formation by human islet amyloid polypeptide has important implications for the mechanism of amyloid formation and the design of inhibitors”. In: *Biochemistry* 49.5 (2010), pp. 872–881.

- [61] Hui Wang et al. “Analysis of the ability of pramlintide to inhibit amyloid formation by human islet amyloid polypeptide reveals a balance between optimal recognition and reduced amyloidogenicity”. In: *Biochemistry* 54.44 (2015), pp. 6704–6711.
- [62] Eri Chatani and Naoki Yamamoto. “Recent progress on understanding the mechanisms of amyloid nucleation”. In: *Biophysical reviews* 10.2 (2018), pp. 527–534.
- [63] Baiping Ren et al. “Fundamentals of cross-seeding of amyloid proteins: an introduction”. In: *Journal of materials chemistry B* 7.46 (2019), pp. 7267–7282.
- [64] Magdalena I Ivanova et al. “Biophysical processes underlying cross-seeding in amyloid aggregation and implications in amyloid pathology”. In: *Biophysical chemistry* 269 (2021), p. 106507.
- [65] Rodrigo Morales et al. “Molecular cross talk between misfolded proteins in animal models of Alzheimer’s and prion diseases”. In: *Journal of Neuroscience* 30.13 (2010), pp. 4528–4535.
- [66] Mingzhen Zhang et al. “Interfacial interaction and lateral association of cross-seeding assemblies between hIAPP and rIAPP oligomers”. In: *Physical Chemistry Chemical Physics* 17.16 (2015), pp. 10373–10382.
- [67] Tsuranobu Shirahama and Alan S Cohen. “High-resolution electron microscopic analysis of the amyloid fibril”. In: *The Journal of cell biology* 33.3 (1967), pp. 679–708.
- [68] Zhigang Wang et al. “AFM and STM study of β -amyloid aggregation on graphite”. In: *Ultramicroscopy* 97.1-4 (2003), pp. 73–79.
- [69] Alexander J Howie and Douglas B Brewer. “Optical properties of amyloid stained by Congo red: history and mechanisms”. In: *Micron* 40.3 (2009), pp. 285–301.
- [70] Liza Nielsen et al. “Effect of environmental factors on the kinetics of insulin fibril formation: elucidation of the molecular mechanism”. In: *Biochemistry* 40.20 (2001), pp. 6036–6046.
- [71] Marlena E Gąsior-Głogowska, Natalia Szulc, and Monika Szeftczyk. “Challenges in Experimental Methods”. In: *Computer Simulations of Aggregation of Proteins and Peptides*. Springer, 2022, pp. 281–307.
- [72] Jennifer A Siepen and David R Westhead. “The fibril_one on-line database: mutations, experimental conditions, and trends associated with amyloid fibril formation”. In: *Protein science* 11.7 (2002), pp. 1862–1866.
- [73] Sandrine Pawlicki, Antony Le Béhec, and Christian Delamarche. “AMYPdb: a database dedicated to amyloid precursor proteins”. In: *BMC bioinformatics* 9 (2008), pp. 1–11.
- [74] Jacinte Beerten et al. “WALTZ-DB: a benchmark database of amyloidogenic hexapeptides”. In: *Bioinformatics* 31.10 (2015), pp. 1698–1700.
- [75] Nikolaos Louros et al. “WALTZ-DB 2.0: an updated database containing structural information of experimentally determined amyloid-forming peptides”. In: *Nucleic acids research* 48.D1 (2020), pp. D389–D393.

- [76] Pawel P Wozniak and Malgorzata Kotulska. “AmyLoad: website dedicated to amyloidogenic protein fragments”. In: *Bioinformatics* 31.20 (2015), pp. 3395–3397.
- [77] A Mary Thangakani et al. “CPAD, curated protein aggregation database: a repository of manually curated experimental data on protein and peptide aggregation”. In: *PLoS One* 11.4 (2016), e0152949.
- [78] Puneet Rawat et al. “CPAD 2.0: a repository of curated experimental data on aggregating proteins and peptides”. In: *Amyloid* 27.2 (2020), pp. 128–133.
- [79] Mihaly Varadi et al. “AmyPro: a database of proteins with validated amyloidogenic regions”. In: *Nucleic acids research* 46.D1 (2018), pp. D387–D392.
- [80] Kristóf Takács, Bálint Varga, and Vince Grohmsuz. “PDB _Amyloid: an extended live amyloid structure list from the PDB”. In: *FEBS Open Bio* 9.1 (2019), pp. 185–190.
- [81] Michał Burdukiewicz et al. “AmyloGraph: a comprehensive database of amyloid–amyloid interactions”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D352–D357.
- [82] Ana-Maria Fernandez-Escamilla et al. “Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins”. In: *Nature biotechnology* 22.10 (2004), pp. 1302–1306.
- [83] Frederic Rousseau, Joost Schymkowitz, and Luis Serrano. “Protein aggregation and amyloidosis: confusion of the kinds?” In: *Current opinion in structural biology* 16.1 (2006), pp. 118–126.
- [84] Michael J Thompson et al. “The 3D profile method for identifying fibril-forming segments of proteins”. In: *Proceedings of the National Academy of Sciences* 103.11 (2006), pp. 4074–4078.
- [85] Lukasz Goldschmidt et al. “Identifying the amyloids, proteins capable of forming amyloid-like fibrils”. In: *Proceedings of the National Academy of Sciences* 107.8 (2010), pp. 3487–3492.
- [86] Antonio Trovato et al. “Insight into the structure of amyloid fibrils from the analysis of globular proteins”. In: *PLoS computational biology* 2.12 (2006), e170.
- [87] Ian Walsh et al. “PASTA 2.0: an improved server for protein aggregation prediction”. In: *Nucleic acids research* 42.W1 (2014), W301–W307.
- [88] Oscar Conchillo-Solé et al. “AGGRESKAN: a server for the prediction and evaluation of " hot spots " of aggregation in polypeptides”. In: *BMC bioinformatics* 8 (2007), pp. 1–17.
- [89] Natalia Sánchez De Groot et al. “Mutagenesis of the central hydrophobic cluster in A β 42 Alzheimer’s peptide: Side-chain properties correlate with aggregation propensities”. In: *The FEBS journal* 273.3 (2006), pp. 658–668.
- [90] Changsik Kim et al. “NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation”. In: *Nucleic acids research* 37.suppl_2 (2009), W469–W473.
- [91] Jian Tian et al. “Prediction of amyloid fibril-forming segments based on a support vector machine”. In: *BMC bioinformatics* 10 (2009), pp. 1–8.

- [92] Sergiy O Garbuzynskiy, Michail Yu Lobanov, and Oxana V Galzitskaya. “FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence”. In: *Bioinformatics* 26.3 (2010), pp. 326–332.
- [93] Sebastian Maurer-Stroh et al. “Exploring the sequence determinants of amyloid structure using position-specific scoring matrices”. In: *Nature methods* 7.3 (2010), pp. 237–242.
- [94] Kimon K Frousios et al. “Amyloidogenic determinants are usually not buried”. In: *BMC structural biology* 9.1 (2009), pp. 1–9.
- [95] Antonios C Tsolis et al. “A consensus method for the prediction of ‘aggregation-prone’ peptides in globular proteins”. In: *PloS one* 8.1 (2013), e54175.
- [96] Mathieu Emily, Anthony Talvas, and Christian Delamarche. “MetAmyl: a METa-predictor for AMYLoid proteins”. In: *PloS one* 8.11 (2013), e79722.
- [97] Pawel Gasior and Malgorzata Kotulska. “FISH Amyloid—a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of aminoacids”. In: *BMC bioinformatics* 15.1 (2014), pp. 1–8.
- [98] Abdullah B Ahmed et al. “A structure-based approach to predict predisposition to amyloidosis”. In: *Alzheimer’s & Dementia* 11.6 (2015), pp. 681–690.
- [99] Stanislav A Bondarev et al. “BetaSerpentine: a bioinformatics tool for reconstruction of amyloid structures”. In: *Bioinformatics* 34.4 (2018), pp. 599–608.
- [100] Carlos Familia et al. “Prediction of peptide and protein propensity for amyloid formation”. In: *PloS one* 10.8 (2015), e0134679.
- [101] Michał Burdukiewicz et al. “Amyloidogenic motifs revealed by n-gram analysis”. In: *Scientific reports* 7.1 (2017), p. 12961.
- [102] Gabriele Orlando et al. “Accurate prediction of protein beta-aggregation with generalized statistical potentials”. In: *Bioinformatics* 36.7 (2020), pp. 2076–2081.
- [103] Jakub W Wojciechowski and Małgorzata Kotulska. “Path-prediction of amyloidogenicity by threading and machine learning”. In: *Scientific Reports* 10.1 (2020), pp. 1–9.
- [104] Nikolaos Louros et al. “Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities”. In: *Nature communications* 11.1 (2020), p. 3314.
- [105] László Keresztes et al. “The budapest amyloid predictor and its applications”. In: *Biomolecules* 11.4 (2021), p. 500.
- [106] Alessia Auriemma Citarella et al. “ENTAIL: yEt aNoTher amyloid fibrils cLassifier”. In: *BMC bioinformatics* 23.1 (2022), pp. 1–15.
- [107] Rafael Zambrano et al. “AGGRESKAN3D (A3D): server for prediction of aggregation properties of protein structures”. In: *Nucleic acids research* 43.W1 (2015), W306–W313.
- [108] Aleksander Kuriata et al. “Aggrescan3D standalone package for structure-based prediction of protein aggregation properties”. In: *Bioinformatics* 35.19 (2019), pp. 3834–3835.

- [109] Aleksander Kuriata et al. “Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility”. In: *Nucleic acids research* 47.W1 (2019), W300–W307.
- [110] Michal Jamroz, Andrzej Kolinski, and Sebastian Kmiecik. “CABS-flex: server for fast simulation of protein structure fluctuations”. In: *Nucleic acids research* 41.W1 (2013), W427–W431.
- [111] Kannan Sankar et al. “AggScore: prediction of aggregation-prone regions in proteins based on the distribution of surface patches”. In: *Proteins: Structure, Function, and Bioinformatics* 86.11 (2018), pp. 1147–1156.
- [112] Mengting Niu et al. “RFAMyloid: a web server for predicting amyloid proteins”. In: *International journal of molecular sciences* 19.7 (2018), p. 2071.
- [113] Phasit Charoenkwan et al. “iAMY-SCM: Improved prediction and analysis of amyloid proteins using a scoring card method with propensity scores of dipeptides”. In: *Genomics* 113.1 (2021), pp. 689–698.
- [114] Phasit Charoenkwan et al. “AMYPred-FRL is a novel approach for accurate prediction of amyloid proteins by using feature representation learning”. In: *Scientific Reports* 12.1 (2022), p. 7697.
- [115] Runtao Yang, Jiaming Liu, and Lina Zhang. “ECAMyloid: An amyloid predictor based on ensemble learning and comprehensive sequence-derived features”. In: *Computational Biology and Chemistry* 104 (2023), p. 107853.
- [116] Annelies Vandersteen et al. “A comparative analysis of the aggregation behavior of amyloid- β peptide variants”. In: *FEBS letters* 586.23 (2012), pp. 4088–4093.
- [117] Fabrizio Chiti et al. “Rationalization of the effects of mutations on peptide and protein aggregation rates”. In: *Nature* 424.6950 (2003), pp. 805–808.
- [118] Kateri F DuBay et al. “Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains”. In: *Journal of molecular biology* 341.5 (2004), pp. 1317–1326.
- [119] Charles W O’Donnell et al. “A method for probing the mutational landscape of amyloid structure”. In: *Bioinformatics* 27.13 (2011), pp. i34–i42.
- [120] Puneet Rawat, Sandeep Kumar, and M Michael Gromiha. “An in-silico method for identifying aggregation rate enhancer and mitigator mutations in proteins”. In: *International journal of biological macromolecules* 118 (2018), pp. 1157–1167.
- [121] Mohamed Raef Smaoui et al. “Computational assembly of polymorphic amyloid fibrils reveals stable aggregates”. In: *Biophysical journal* 104.3 (2013), pp. 683–693.
- [122] Mohamed Raef Smaoui, Henri Orland, and Jérôme Waldispühl. “Probing the binding affinity of amyloids to reduce toxicity of oligomers in diabetes”. In: *Bioinformatics* 31.14 (2015), pp. 2294–2302.
- [123] Witold Dyrka et al. “Searching for universal model of amyloid signaling motifs using probabilistic context-free grammars”. In: *BMC bioinformatics* 22.1 (2021), p. 222.

- [124] Witold Dyrka and Jean-Christophe Nebel. “A stochastic context free grammar based framework for analysis of protein sequences”. In: *BMC bioinformatics* 10 (2009), pp. 1–24.
- [125] Andrej Šali and Tom L Blundell. “Comparative protein modelling by satisfaction of spatial restraints”. In: *Journal of molecular biology* 234.3 (1993), pp. 779–815.
- [126] Min-yi Shen and Andrej Sali. “Statistical potential for assessment and prediction of protein structures”. In: *Protein science* 15.11 (2006), pp. 2507–2524.
- [127] Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J Gray. “PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta”. In: *Bioinformatics* 26.5 (2010), pp. 689–691.
- [128] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.
- [129] Miron B Kurasa and Witold R Rudnicki. “Feature selection with the Boruta package”. In: *Journal of statistical software* 36 (2010), pp. 1–13.
- [130] Xuan Wang and Matthew R Chapman. “Sequence determinants of bacterial amyloid formation”. In: *Journal of molecular biology* 380.3 (2008), pp. 570–580.
- [131] Lee Sewell et al. “NMR insights into the pre-amyloid ensemble and secretion targeting of the curli subunit CsgA”. In: *Scientific reports* 10.1 (2020), p. 7896.
- [132] Catherine H Rankin, Christine DO Beck, and Catherine M Chiba. “Caenorhabditis elegans: a new model system for the study of learning and memory”. In: *Behavioural brain research* 37.1 (1990), pp. 89–92.
- [133] Ghulam Jeelani Pir, Bikash Choudhary, and Eckhard Mandelkow. “Caenorhabditis elegans models of tauopathy”. In: *The FASEB Journal* 31.12 (2017), pp. 5137–5148.
- [134] Javier Alvarez et al. “Modeling Alzheimer’s disease in Caenorhabditis elegans”. In: *Biomedicines* 10.2 (2022), p. 288.
- [135] Elizabeth K Marsh and Robin C May. “Caenorhabditis elegans, a model organism for investigating immunity”. In: *Applied and environmental microbiology* 78.7 (2012), pp. 2075–2081.
- [136] Nguyen Phuoc Long, Jong Seong Kang, and Hyung Min Kim. “Caenorhabditis elegans: a model organism in the toxicity assessment of environmental pollutants”. In: *Environmental Science and Pollution Research* (2023), pp. 1–15.
- [137] Fan Zhang et al. “Caenorhabditis elegans as a model for microbiome research”. In: *Frontiers in microbiology* 8 (2017), p. 485.
- [138] Christine Xue et al. “Thioflavin T as an amyloid dye: fibril quantification, optimal concentration and effect on aggregation”. In: *Royal Society open science* 4.1 (2017), p. 160696.
- [139] Alba Espargaró et al. “On the binding of Congo Red to amyloid fibrils”. In: *Angewandte Chemie* 132.21 (2020), pp. 8181–8184.

- [140] Nobuhisa Iwata et al. “Metabolic regulation of brain A β by neprilysin”. In: *Science* 292.5521 (2001), pp. 1550–1552.
- [141] Samuel IA Cohen et al. “A molecular chaperone breaks the catalytic cycle that generates toxic A β oligomers”. In: *Nature structural & molecular biology* 22.3 (2015), pp. 207–213.
- [142] Théo Falgarone et al. “Census of exposed aggregation-prone regions in proteomes”. In: *bioRxiv* (2022), pp. 2022–12.
- [143] Koki Ikeda et al. “Presence of intrinsically disordered proteins can inhibit the nucleation phase of amyloid fibril formation of A β (1–42) in amino acid sequence independent manner”. In: *Scientific reports* 10.1 (2020), p. 12334.
- [144] Michael P Hughes et al. “Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks”. In: *Science* 359.6376 (2018), pp. 698–701.
- [145] Jessie Uehling, Aurélie Deveau, and Mathieu Paoletti. “Do fungi have an innate immune response? An NLR-based comparison to plant and animal immune systems”. In: *PLoS pathogens* 13.10 (2017), e1006578.
- [146] Zane Duxbury, Chih-hang Wu, and Pingtao Ding. “A comparative overview of the intracellular guardians of plants and animals: NLRs in innate immunity and beyond”. In: *Annual review of plant biology* 72 (2021), pp. 155–184.
- [147] Sven J Saupe. “Amyloid signaling in filamentous fungi and bacteria”. In: *Annual Review of Microbiology* 74 (2020), pp. 673–691.
- [148] Asen Daskalov et al. “Signal transduction by a fungal NOD-like receptor based on propagation of a prion amyloid fold”. In: *PLoS biology* 13.2 (2015), e1002059.
- [149] Martin Steinegger and Johannes Söding. “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets”. In: *Nature biotechnology* 35.11 (2017), pp. 1026–1028.
- [150] Michael Remmert et al. “HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment”. In: *Nature methods* 9.2 (2012), pp. 173–175.
- [151] “Database resources of the national center for biotechnology information”. In: *Nucleic acids research* 46.D1 (2018), pp. D8–D13.
- [152] Weizhong Li and Adam Godzik. “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”. In: *Bioinformatics* 22.13 (2006), pp. 1658–1659.
- [153] Timothy L Bailey et al. “The MEME suite”. In: *Nucleic acids research* 43.W1 (2015), W39–W49.
- [154] Sujeet S Bhoite et al. “Mechanistic insights into accelerated α -synuclein aggregation mediated by human microbiome-associated functional amyloids”. In: *Journal of Biological Chemistry* 298.7 (2022).

Appendix



OPEN

PATH - Prediction of Amyloidogenicity by Threading and Machine Learning

Jakub W. Wojciechowski & Małgorzata Kotulska

Amyloids are protein aggregates observed in several diseases, for example in Alzheimer's and Parkinson's diseases. An aggregate has a very regular beta structure with a tightly packed core, which spontaneously assumes a steric zipper form. Experimental methods enable studying such peptides, however they are tedious and costly, therefore inappropriate for genomewide studies. Several bioinformatic methods have been proposed to evaluate protein propensity to form an amyloid. However, the knowledge of aggregate structures is usually not taken into account. We propose PATH (Prediction of Amyloidogenicity by THreading) - a novel structure-based method for predicting amyloidogenicity and show that involving available structures of amyloidogenic fragments enhances classification performance. Experimental aggregate structures were used in templatebased modeling to recognize the most stable representative structural class of a query peptide. Several machine learning methods were then applied on the structural models, using their energy terms. Finally, we identified the most important terms in classification of amyloidogenic peptides. The proposed method outperforms most of the currently available methods for predicting amyloidogenicity, with its area under ROC curve equal to 0.876. Furthermore, the method gave insight into significance of selected structural features and the potentially most stable structural class of a peptide fragment if subjected to crystallization.

Amyloids are unbranched, fibrillar protein aggregates, which produce characteristic diffraction pattern in X-ray diffraction experiments¹. For a long time their occurrence was associated exclusively with severe neurodegenerative diseases, such as Alzheimer's and Parkinson's diseases. However, more recent studies showed that these proteins play other roles in a wide variety of organisms, from bacteria and fungi to human². Plenty of studies showed that formation of amyloid fibers depends on the presence of short fragments with an appropriate sequence patterns, called hot-spots³. These fragments are responsible for formation of a steric zipper - tightly packed structure which involves two beta sheets that form a core of the amyloid aggregate. High resolution studies, using X-ray diffraction, shed light on molecular details of the steric zipper and revealed different forms of packing peptides into such structures. Theoretical extrapolation, based on symmetry operations, led to proposing ten putative structural classes of the zipper structure, formed by parallel or antiparallel beta sheets⁴. Currently, using X-ray diffraction, seven structural classes of the crystal zipper structure have been identified. Classes 3, 9, and 10 have not been proven, yet.

Amyloidogenic hot-spots can be identified experimentally and computationally. Experiments typically use Congo Red⁵ and Thioflavin T⁶ staining, high resolution techniques such as electron microscopy⁷ and atomic force microscopy⁸. Recently, infrared spectroscopy has become one of the leading methods, due to its simplicity and efficiency⁹. However, experimental techniques are expensive and time consuming, which hampers their use in genome wide studies.

To overcome these limitations several bioinformatic methods for amyloid prediction have been proposed. Some of them, like PASTA 2.0¹⁰ or ArchCandy¹¹, use structural information. Others like Waltz¹² or AGGREGSCAN¹³ employ statistical analysis of a sequence. FoldAmyloid¹⁴ utilizes density of a protein's contact sites. Along with a growing number of known amyloidogenic sequences, machine learning methods, such as FISH Amyloid¹⁵, APPNN¹⁶, or AmyloGram¹⁷ were proposed. Finally, consensus predictors, such as MetAmyl¹⁸ or Amylpred2¹⁹, are also available. Machine learning methods benefit from statistically significant patterns, which can be found in datasets, capable of providing predictions with a good accuracy. Not all of them reveal relations between the features representing solutions to the problem. For example, most of the bioinformatic methods do

Department of Biomedical Engineering, Wrocław University of Science and Technology, 50-370, Wrocław, Poland.
✉ e-mail: malgorzata.kotulska@pwr.edu.pl

| Structural class | PDB code | Sequence | Origin |
|------------------|----------|----------|----------------------|
| 1 | 1YJO | NNQQNY | yeast prion Sup35 |
| 2 | 2Y3J | AIIGLM | amyloid-beta |
| 4 | 2ONV | GGVVIA | amyloid-beta |
| 5 | 3LOZ | LSFSKD | beta 2 microglobulin |
| 6 | 3PZZ | GAIIGL | amyloid-beta |
| 7 | 3OW9 | KLVFFA | amyloid-beta |
| 8 | 3NHC | GYMLGS | human prion PrP |

Table 1. Hexapeptide structures representing different classes, used as templates for modeling (class numbering in accordance with⁴).

not provide much insight into structure of amyloidogenic fragments, especially when aggregates are formed by short sequences. Studying short peptides is relevant, since they represent the phenomenon of amyloidogenicity triggering amyloid pathways of longer peptides or proteins. Moreover, most of the currently known amyloidogenic sequences are hexapeptides^{20,21}. Despite polymorphism in amyloid structures, amyloid crystals represent the ground state of the protein folding energy landscape in short peptides, hence including them in modeling amyloidogenicity may bring essential knowledge into these methods²².

Accordingly, we combined a structural approach to modeling amyloidogenicity with machine learning methods, and developed PATH (Prediction of Amyloidogenicity by THreading). While classifying amyloidogenic propensity of peptides, PATH should provide structural insight into steric zipper structures formed by their crystals. Furthermore, we aimed to identify the most important energy terms characterizing these structures, which split them between potential amyloids and non-amyloids.

Methods

Data set. The data that we used in our study included four data sets of hexapeptides. Peptide fragments of this length are regarded as very good representatives of amyloid hot-spots, which are believed to include between 4 and 10 amino acids. Moreover, they constitute the majority of instances in databases of amyloidogenic sequences.

The first data set, *Templates*, consisted of structural templates that were applied to modeling potential structures of amyloid aggregates formed by other hexapeptides. Based on the structural classification of amyloid hexapeptides, proposed in⁴, seven crystallographic structures of steric zippers were selected from the Protein Data Bank. Their crystallographic structures of steric zippers were selected from the Protein Data Bank. Each of them represented one of the experimentally confirmed structural class of amyloid hexapeptides (see Table 1). For the purpose of this study, the available structures were processed. All non-protein fragments, such as small organic molecules assisting crystallization, ions, or water molecules were removed from the structures. Since the original structures differed in numbers of chains forming the zipper, our final templates were built with six peptide chains in each beta sheet. In this procedure, copying existing chains and translating them by an appropriate vector was performed, based on the crystallographic data of the original structures.

In the first stage of our study, we modeled structural classes of amyloidogenic peptides. For this purpose we collected a set of validating structures from the Protein Data Bank. The obtained set, *Amyloid Structures*, consisted of 24 amyloid fibers with experimentally determined structures, which were available in the Protein Data Bank and already assigned to a structural class⁴.

The data used in the final classification between two classes (amyloids and non-amyloids) constituted the *Classification set*, which was extracted from the Waltz database²³. It consisted of 1080 unique hexapeptides, experimentally assigned either to amyloidogenic (244 peptides) or non-amyloidogenic (836 peptides) class. These data were used for design, training, and testing the final effectiveness of our method. Before training the method, the classification set was divided into two separate sets - the training and testing data. During development of the classification methods, k-folds cross validation was performed with $k=5$, on the training data set only. The final testing set consisted of 326 randomly selected peptides (30% of the total classification set), in which 85 were amyloidogenic and 241 non-amyloidogenic. It was used to evaluate the complete method.

Our approach was additionally tested with the use of hexapeptides included in another benchmark set, *pep424*, applied by the authors of Pasta 2.0 for evaluation of their method¹⁰. The benchmark set consisted of 164 hexapeptides from *pep424*, in which 67 were amyloidogenic and 97 non-amyloidogenic. The advantage of this data is such that it is much better balanced with regard to representation of both classes. Similarly to our previously described classification test, it was used to train and then test the potential of our algorithm. In both cases the sets were randomly divided into training and test sets containing 70% and 30% of samples, respectively.

Modeling. Each query sequence, from the set of amyloid structures and from training part of the classification set, was threaded into seven previously described templates. The structural modeling was performed with Modeller 9.21, which is designed for homology or comparative modeling, and its automodel class with default parameters and the model-multichain.py procedure²⁴. Ten models were obtained for each structural class (70 models in total), for each of the query sequences. All of the models were scored using DOPE statistical potential implemented in Modeller. The model with the lowest value of this score, representing each of the sequences, was then chosen for further analysis.

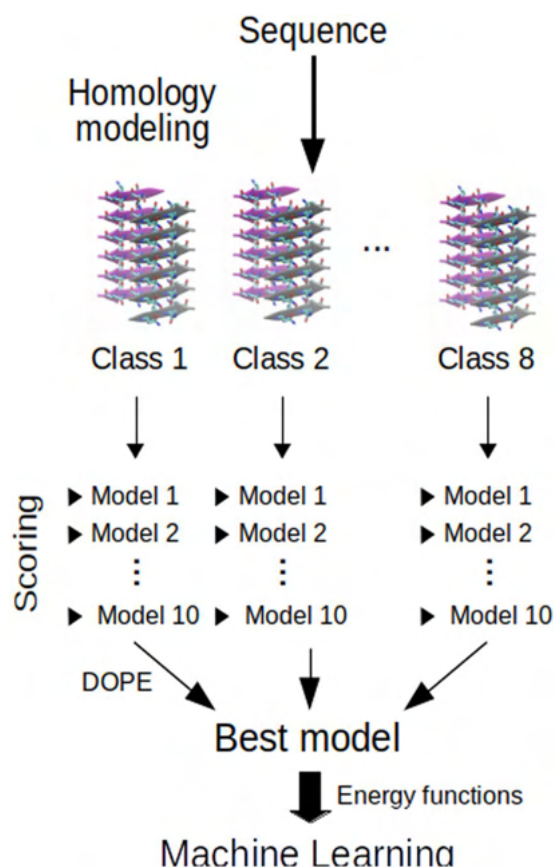


Figure 1. Prediction procedure. Using comparative modeling, query sequence was threaded into seven templates representing different structural classes (class numbering in accordance with⁴). For each of them ten models were proposed and the model with the lowest DOPE score was selected. For this model REF15 and selected PyRosetta energy terms were calculated and used as an input for machine learning classifiers.

First, based on the best model structures obtained from Modeller, we attempted to predict structural classes of peptide fragments. In this study, only 24 peptides included in the set of amyloid structures were modeled.

The next study was applied to the sequences from the training part of the classification set. For these sequences, Rosetta Energy Function (REF15)²⁵ and some of its components were calculated for each of their optimal model structures selected in the first stage. The statistical potentials corresponded to the following energy terms: van der Waals interactions (*fa_atr*, *fa_rep*, *fa_intra_rep*), electrostatic interactions (*fa_elec*), interactions with a solvent (*fa_sol*, *lk_ball_wtd*, *fa_intra_sol_xover4*), and statistical parameters describing amino acid conformation (*omega*, *fa_dun*, *p_aa_pp*, *ref*, *rama_prepro*). All of them were calculated using PyRosetta²⁶. These terms, as well as previously computed DOPE and other scores provided by Modeller, were normalized and further used as an input for machine learning classifiers (see Fig. 1). Logistic regression, support vector machines (SVM) with three different kernel functions, and random forest methods were tested.

All classifiers were built using scikit-learn Python library²⁷. The logistic regression model with L1 regularization was built using *LogisticRegression* function with default parameters. SVM with linear, polynomial and RBF kernels were built using *sklearn.svm.SVC* method with default parameters. The random forest consisted of 100 decision trees with the maximum depth of 4, and it was built with *RandomForestClassifier* using cross-entropy as the loss function. Other parameters of models were default. To make sure that the methods do not overfit to the data, k-folds cross validation was performed on the training data set, with $k=5$.

Methods were trained and tested on both classification and benchmark sets. In both cases the sets were randomly divided into training and test sets containing 70% and 30% of samples, respectively. To assess the performance of the method, accuracy, which is defined as a fraction of correctly classified samples, sensitivity, specificity, Area Under ROC Curve (AUC) and Matthew Correlation Coefficient (MCC) were calculated.

Feature selection. In order to identify the most important features that distinguish models of amyloids from non-amyloids, feature selection was performed, using three approaches. In the first step, we analyzed coefficients of logistic regression. In general, input parameters with the largest absolute values of coefficients contribute more to the classification than parameters with coefficients close to zero. Furthermore, the proposed model used L1 regularization, which penalized the contribution of less significant inputs. Feature selection was performed using

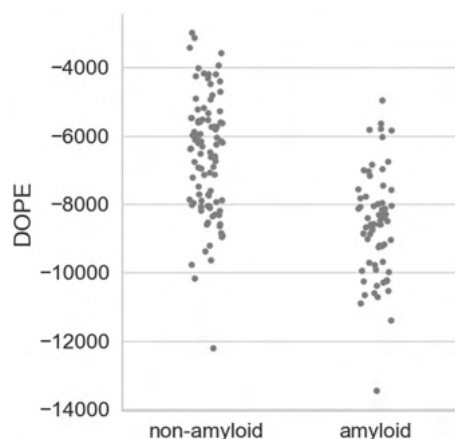


Figure 2. DOPE score of amyloidogenic and non-amyloidogenic sequences threaded onto steric zipper structures for the benchmark set.

| class | 1 | 2 | 4 | 5 | 6 | 7 | 8 |
|--------------------|------|------|------|------|------|------|------|
| number of peptides | 5 | 4 | 3 | 2 | 1 | 4 | 5 |
| accuracy | 0.40 | 0.75 | 0.67 | 0.50 | 0.00 | 0.25 | 0.40 |

Table 2. Accuracy of class prediction for different classes from the study on the set of Structures.

previously described random forest classifier and `feature_importances_` method. Finally feature selection was performed using the algorithm Boruta²⁸.

Results and Discussion

Hexapeptides were modeled using seven templates representing all different structural classes of the steric zipper form. For each sequence the model with the lowest DOPE score was chosen. Figure 2 shows the obtained DOPE values for the best models of amyloidogenic and non-amyloidogenic fragments from the benchmark set. As expected, amyloids obtained lower scores when threaded onto steric zippers, which indicate that they formed more energetically favourable, and thus more stable, structures. However, this method alone does not allow unambiguously distinguishing between amyloids and non-amyloids since both classes partially overlap (Fig. 2). Nevertheless, this initial study showed that DOPE may give some clue to the nature of a query hexapeptide.

Structural class prediction. Based on the comparative modeling and available class templates, we tested whether it is possible to predict the potentially most probably structural class which a query peptide sequence could assume. This hypothesis assumed that the model structure with the lowest energy, represented by DOPE value, should be the most stable and closest to the native structure. Our results showed that the model with the lowest DOPE score did not always correspond to the experimentally observed class. A predicted structural class matched the one experimentally derived in only 11 out of 24 peptides from the set of structures. Table 2 shows the accuracy of the classification in the form one vs all.

Although the results turned out below the expectations, it should be noted that the testing set was very small. Unfortunately, there are very few X-ray structures of amyloid hexapeptides that have been annotated to representative classes⁴, therefore many structural classes were strongly underrepresented. There are also other reasons why such methods might struggle to identify the correct class. For example, most statistical potentials including DOPE are parametrized using a set of globular proteins. Thus, they might describe protein aggregates, such as amyloids, with a certain level of inaccuracy. Another reason may be structural polymorphism of amyloid structures. To certain extent, the final structure of a fiber depends on the experimental conditions, and even in the same environment a population containing different structures can be observed²⁹. However, there are studies showing that crystal structures of short amyloid fragments assume a very stable and well defined structure, in contrast to polymorphic fibrils²². In the case of our studies, energy differences between amyloids and non-amyloids were much greater than between different structural classes. Therefore, a certain inaccuracy in the class prediction should not affect prediction of amyloidogenicity. Energy differences between amyloids and non-amyloids were much greater than between different structural classes.

Predicting amyloidogenicity. Using calculated structural scores and several energy terms of the assumably most stable structural models, which were obtained in the first stage, we trained several machine learning methods to classify amyloidogenicity of hexapeptides. Figure 3 shows the receiver operating curves (ROC) of all methods. Table 3 shows metrics of the best classifiers, such as area under ROC curve (AUC), sensitivity and specificity, obtained on the testing subsets of the data used to build our method. These results are very close to the results obtained from the k-fold cross-validation. The performance of selected methods was within the same

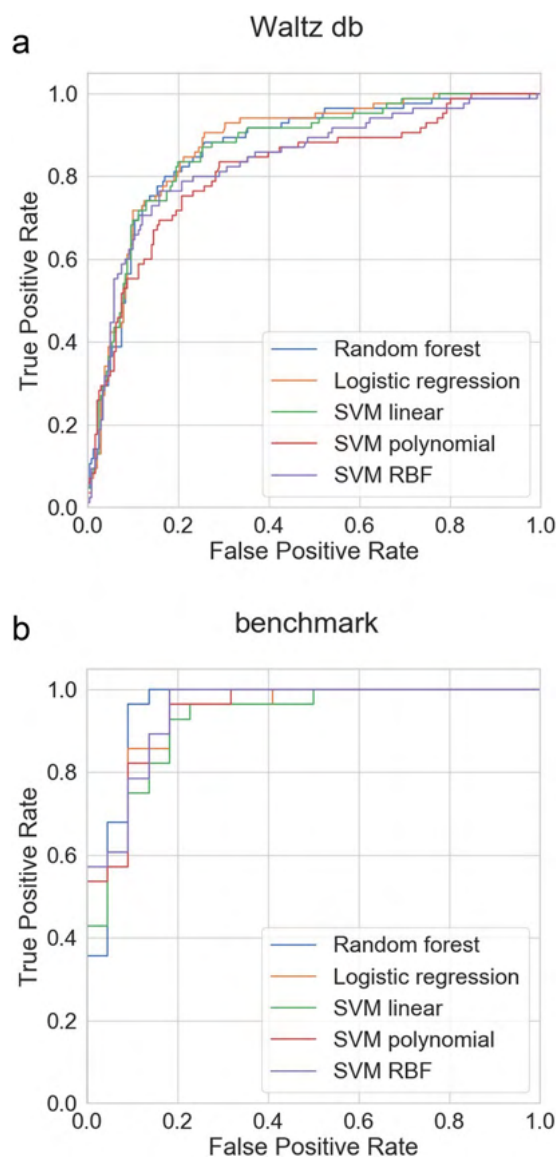


Figure 3. ROC curves for compared machine learning methods, for two test sets.

| Method | AUC [95% CI] | | Sensitivity [95% CI] | | Specificity [95% CI] | |
|-----------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| | Waltz db | benchmark | Waltz db | benchmark | Waltz db | benchmark |
| Logistic regression | 0.8762 [0.8351–0.9161] | 0.9379 [0.9091–0.9609] | 0.4235 [0.3082–0.5404] | 0.8569 [0.7945–0.9077] | 0.9414 [0.9095–0.9712] | 0.9069 [0.8528–0.9528] |
| SVM linear kernel | 0.8681 [0.8208–0.9115] | 0.9232 [0.8899–0.9500] | 0.3540 [0.2597–0.4578] | 0.7828 [0.7109–0.8403] | 0.9551 [0.9271–0.9811] | 0.8655 [0.8024–0.9170] |
| SVM polynomial kernel | 0.8165 [0.7577–0.8732] | 0.9418 [0.9153–0.9645] | 0.2132 [0.1219–0.3084] | 0.6792 [0.6095–0.7449] | 0.9792 [0.9577–0.9955] | 0.9089 [0.8565–0.9540] |
| SVM RBF kernel | 0.8375 [0.7828–0.8867] | 0.9479 [0.9237–0.9670] | 0.4843 [0.3681–0.5951] | 0.7865 [0.7222–0.8418] | 0.9409 [0.9087–0.9710] | 0.8627 [0.8000–0.9134] |
| Random forest | 0.8668 [0.8193–0.9090] | 0.9544 [0.9289–0.9780] | 0.4273 [0.3188–0.5409] | 0.8915 [0.8403–0.9387] | 0.9259 [0.8929–0.9585] | 0.9086 [0.8542–0.9526] |

Table 3. Performance of machine learning classifiers trained and tested on subsets of classification (Waltz database) and benchmark sets. 95% confidence intervals (CI) were calculated using bootstrap.

range. Finally, for PATH we chose logistic regression, which was the simplest and one of the best performing classifiers. An additional benefit of using this method is that it is highly interpretable because it fits several coefficients of the linear function during the training. All methods were trained and tested on the training subsets of two data sets: classification database and hexapeptides from pep424 data set (see Methods).

The performance of PATH (on its test set) was compared to selected three other top predictors of amyloidogenicity (Table 4). The performance metrics of other methods are based on the data reported by their authors,

| Method | AUC | Sensitivity | Specificity | MCC |
|-------------|--------|-------------|-------------|--------|
| PATH | 0.8762 | 0.4235 | 0.9414 | 0.4444 |
| Pasta 2.0 | 0.8550 | 0.3826 | 0.9519 | 0.4291 |
| AmyloGram | 0.8856 | 0.6779 | 0.9037 | 0.6057 |
| FoldAmyloid | 0.7531 | 0.7517 | 0.7185 | 0.4526 |

Table 4. Comparison of PATH with several state of the art predictors of amyloidogenicity. PATH showed one of the best performances among tested methods.

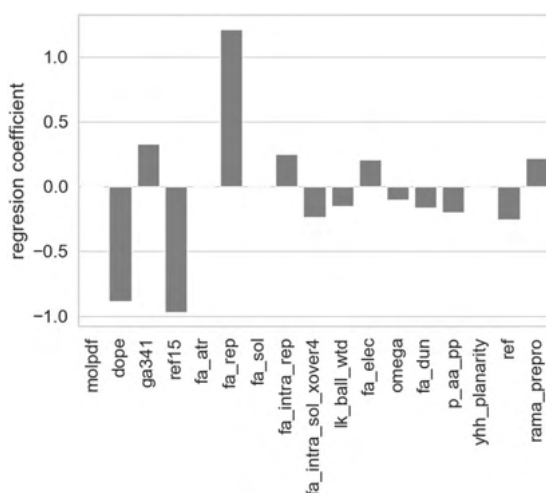


Figure 4. Coefficients of logistic regression model for energy terms.

calculated using their test sets. The observed differences between the three best performing predictors were negligible. PATH performed very well compared to the methods used here for a comparison.

Additionally, as suggested by one of the reviewers, we tested the approach from PATH on FVFLM pentapeptide, which was shown to inhibit aggregation of Abeta amyloid involved in the Alzheimer's disease, although it is strongly amyloidogenic itself³⁰. This peptide was too short for many predictors of amyloidogenicity. Moreover only four out of more than ten predictors that could analyze it, identified it correctly as amyloidogenic. PATH was able to classify it correctly as an amyloid, even though our method was not trained on any pentapeptides or hexapeptides that included this sequence.

Feature selection. In the next step, we identified the most important features for our model. This was done by studying the logistic regression coefficients (see Fig. 4). Three most important features turned out to be DOPE score, REF15 score and the repulsive component of Lennard-Jones term (fa_rep). The first two terms represent complex energy functions that are not trivial to interpret. In general, both of them should approximate the free energy of a protein, thus low values of these functions indicate highly stable conformations. As expected, amyloidogenic fragments obtained lower scores when threaded into a steric zipper structure. The last identified term, fa_rep , has a physical interpretation - it describes repulsion between atoms, arising from Pauli repulsion. This indicates that structures with high values of this function are too tightly packed.

A similar analysis was performed using a random forest classifier and the algorithm Boruta. Figure 5 shows the most important features. DOPE and fa_rep terms showed a large impact on the classification. Since this classifier is more sophisticated and capable of capturing nonlinear relationships in data, more features were identified in this case. The term fa_atr , describing the attractive part of the Lennard-Jones potential between two atoms on different residues separated by a certain distance, approximates van der Waals interactions, which are known to be important for fiber stabilization. A similar interpretation has fa_intra_rep , but it is calculated for atoms within the same residue. Finally, ref is a reference energy for a given amino acid type in an unfolded state and it was introduced in Rosetta as a tool for protein design. It reflects the importance of the amino acid composition of amyloidogenic fragments. Since logistic regression and random forest used slightly different features, we tested if both methods produced consistent results. Comparing the results from both classifiers, it turned out that less than 7% of sequences were classified differently.

Conclusions

Recognizing amyloidogenic propensity of short peptides provides more knowledge on their potentially adverse behavior, especially if they appear inside longer functional proteins. Bioinformatical methods offering their fast and faultless identification are indispensable tools to advance the prediction of amyloids.

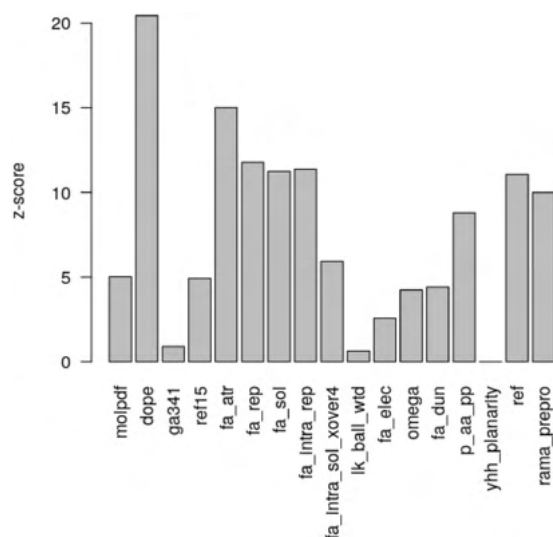


Figure 5. Mean z-scores for energy terms obtained from Boruta feature selection method.

We proposed a new method predicting amyloidogenic propensity of hexapeptides. PATH is based on threading potentially amyloidogenic sequences on zipper-like amyloid structures, corresponding to all representative and experimentally confirmed structural classes of short amyloids. An affinity of a sequence to each structural class was first evaluated with regard to the total energy of its structure. The structures were obtained using comparative modeling with regard to structures of class representatives. A model with the minimal DOPE statistical potential, representing total energy of the structure, was assumed as the most stable and the most accurate for each tested sequence. Although this energy could hint on potential amyloidogenicity of a sequence, since the median energy of amyloids was lower than that of non-amyloids, it did not allow for unambiguous split into two classes. Some of the non-amyloidogenic sequences, forced to assume an amyloid structure, received energies lower than amyloidogenic sequences.

In the other study, only amyloidogenic sequences were considered. The objective was prediction of the most suitable structural class of a query sequence, based on the total energy of its model structure. The correct structural class was accurately predicted for 46% of sequences. The class corresponding to the model with the lowest energy was selected out of 7 possible classes. This analysis, however, could not be regarded as conclusive - the set of available instances is very scarce, containing only 24 sequences. There might be also other reasons which additionally hamper this kind of modeling, such as the choice of the statistical potential, representing energy of the structures, or tendency of amyloid structures to polymorphism.

Since the general structural features are not sufficient for differentiating between amyloids and non-amyloids, in the next stage of the classification we applied more specific and descriptive energy terms from PyRosetta. The extended set of statistical potentials, corresponding to the best model structure from the first stage of modeling, was used to build computational machine learning models. Out of several available algorithms, we finally selected logistic regression, as the one which gave the best accuracy. Additionally, this method is not so much of a black-box type, allowing for interpretation of the results. Our method was verified on two data sets, using also cross-validation. PATH showed a very good potential for classification of amyloidogenicity, with AUC ROC at 0.88, sensitivity $S_n = 0.42$, specificity $S_p = 0.94$, and MCC equal to 0.44. The relatively low value of sensitivity is a problem for many other amyloid predictors and is related to the low ratio of amyloids to non-amyloids in the available experimental data. This was the case in our method. Applying the other benchmark data set, consisting of hexapeptides from the pep424 set, the sensitivity was much higher ($S_n = 0.8569$), without deterioration of the specificity ($S_p = 0.9069$). This was due to much better balanced number of instances representing both classes in this data set.

Confronting our method with other available predictors, we note that its effectiveness is very high and it could effectively support modeling amyloidogenicity. One of its assets, compared to other methods, is the combination of the structural approach with machine learning on numerous instances. It appears that a somehow similar approach of modeling amyloidogenicity was very recently applied in the version 2.0 of Waltz database class prediction²¹, in which experimental data of the instances are accompanied with their structural models and their energy values. The authors, however, do not reveal all details regarding their method and its performance.

Due to the high interpretability of our method, it was possible to identify the most important features that distinguished amyloids from non-amyloids in the classification. Apart from differences in total energies, such as DOPE from Modeler and REF15 from PyRosetta, some other energy terms appeared to play a role. All applied methods for feature selection showed the importance of the *fa_rep* energy term, representing repulsive van der Waals interactions, which approximate Pauli repulsion, whose high values may indicate clashes in structures and imply that non-amyloidogenic fragments may not be fitted well into a steric zipper structure. It could explain relatively low accuracy of the first stage of modeling, in which non-amyloids threaded on amyloid structures, are indeed faulty. Also, a relatively high importance of the energy term describing repulsion of atoms within the same

residue (`fa_intra_rep`) was observed. A relatively low importance of statistical terms describing conformation of the backbone and side chains were observed. It should be noted, however, that statistical potentials were mostly fitted to describe structures of globular proteins and their use with other proteins may not be optimal¹¹, therefore using better suited descriptors might improve the results and give more insight into the most influential features of the structures.

Code availability

All the scripts are available on <https://github.com/KubaWojciechowski/PATH>.

Received: 11 December 2019; Accepted: 24 March 2020;

Published online: 07 May 2020

References

- Eisenberg, D. & Jucker, M. The amyloid state of proteins in human diseases. *Cell* **148**, 1188–1203 (2012).
- McGlinchey, R. P. & Lee, J. C. Why study functional amyloids? Lessons from the repeat domain of pmel17. *J. molecular biology* **430**, 3696–3706 (2018).
- de la Paz, M. L. & Serrano, L. Sequence determinants of amyloid fibril formation. *Proc. Natl. Acad. Sci.* **101**, 87–92 (2004).
- Eisenberg, D. S. & Sawaya, M. R. Structural studies of amyloid proteins at the molecular level. *Annu. review biochemistry* **86**, 69–95 (2017).
- Howie, A. J. & Brewer, D. B. Optical properties of amyloid stained by congo red: history and mechanisms. *Micron* **40**, 285–301 (2009).
- Nielsen, L. *et al.* Effect of environmental factors on the kinetics of insulin fibril formation: elucidation of the molecular mechanism. *Biochemistry* **40**, 6036–6046 (2001).
- Shirahama, T. & Cohen, A. S. High-resolution electron microscopic analysis of the amyloid fibril. *The J. cell biology* **33**, 679–708 (1967).
- Wang, Z. *et al.* Afm and stm study of b-amyloid aggregation on graphite. *Ultramicroscopy* **97**, 73–79 (2003).
- Sarroukh, R., Goormaghtigh, E., Ruyschaert, J.-M. & Raussens, V. Atr-ftir: a “rejuvenated” tool to investigate amyloid proteins. *Biochimica et Biophys. Acta (BBA)-Biomembranes* **1828**, 2328–2338 (2013).
- Walsh, I., Seno, F., Tosatto, S. C. & Trovato, A. Pasta 2.0: an improved server for protein aggregation prediction. *Nucleic acids research* **42**, W301–W307 (2014).
- Ahmed, A. B., Znassi, N., Château, M.-T. & Kajava, A. V. A structure-based approach to predict predisposition to amyloidosis. *Alzheimer's & Dementia* **11**, 681–690 (2015).
- Maurer-Stroh, S. *et al.* Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. methods* **7**, 237 (2010).
- Conchillo-Solé, O. *et al.* Aggrescan: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC bioinformatics* **8**, 65 (2007).
- Garbuzynskiy, S. O., Lobanov, M. Y. & Galzitskaya, O. V. Foldamyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* **26**, 326–332 (2009).
- Gasior, P. & Kotulska, M. Fish amyloid—a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of aminoacids. *BMC bioinformatics* **15**, 54 (2014).
- Família, C., Dennison, S. R., Quintas, A. & Phoenix, D. A. Prediction of peptide and protein propensity for amyloid formation. *PLoS one* **10**, e0134679 (2015).
- Burdukiewicz, M. *et al.* Amyloidogenic motifs revealed by n-gram analysis. *Sci. reports* **7**, 12961 (2017).
- Emily, M., Talvas, A. & Delamarque, C. Metamyli: a meta-predictor for amyloid proteins. *PLoS one* **8** (2013).
- Tsolis, A. C., Papandreou, N. C., Iconomidou, V. A. & Hamodrakas, S. J. A consensus method for the prediction of ‘aggregation-prone’ peptides in globular proteins. *PLoS one* **8** (2013).
- Wozniak, P. P. & Kotulska, M. Amyloid: website dedicated to amyloidogenic protein fragments. *Bioinformatics* **31**, 3395–3397 (2015).
- Louros, N. *et al.* Waltz-db 2.0: an updated database containing structural information of experimentally determined amyloid-forming peptides. *Nucleic Acids Res.* **48**, D389–D393 (2020).
- Reynolds, N. P. *et al.* Competition between crystal and fibril formation in molecular mutations of amyloidogenic peptides. *Nat. communications* **8**, 1338 (2017).
- Beerten, J. *et al.* Waltz-db: a benchmark database of amyloidogenic hexapeptides. *Bioinformatics* **31**, 1698–1700 (2015).
- Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. molecular biology* **234**, 779–815 (1993).
- Alford, R. F. *et al.* The rosetta all-atom energy function for macromolecular modeling and design. *J. chemical theory computation* **13**, 3031–3048 (2017).
- Chaudhury, S., Lyskov, S. & Gray, J. J. Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. *Bioinformatics* **26**, 689–691 (2010).
- Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. machine learning research* **12**, 2825–2830 (2011).
- Kursa, M. B. & Rudnicki, W. R. *et al.* Feature selection with the boruta package. *J. Stat Softw* **36**, 1–13 (2010).
- Tycko, R. Amyloid polymorphism: structural basis and neurobiological relevance. *Neuron* **86**, 632–645 (2015).
- Kouza, M., Banerji, A., Kolinski, A., Buhimschi, I. A. & Kloczkowski, A. Oligomerization of fvflm peptides and their ability to inhibit beta amyloid peptides aggregation: consideration as a possible model. *Phys. Chem. Chem. Phys.* **19**, 2990–2999 (2017).

Acknowledgements

Wrocław Centre for Networking and Supercomputing at Wrocław University of Science and Technology is acknowledged.

Author contributions

J.W.W. and M.K. developed the concept. J.W.W. implemented the algorithms. J.W.W. and M.K. analyzed data and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020



OPEN

Bioinformatics methods for identification of amyloidogenic peptides show robustness to misannotated training data

Natalia Szulc^{1,2}, Michał Burdukiewicz^{3,4,7}✉, Marlena Gąsior-Głogowska¹, Jakub W. Wojciechowski¹, Jarosław Chilimoniuk⁵, Paweł Mackiewicz⁵, Tomas Šneideris⁶, Vytautas Smirnovas⁶ & Małgorzata Kotulska^{1,7}✉

Several disorders are related to amyloid aggregation of proteins, for example Alzheimer's or Parkinson's diseases. Amyloid proteins form fibrils of aggregated beta structures. This is preceded by formation of oligomers—the most cytotoxic species. Determining amyloidogenicity is tedious and costly. The most reliable identification of amyloids is obtained with high resolution microscopies, such as electron microscopy or atomic force microscopy (AFM). More frequently, less expensive and faster methods are used, especially infrared (IR) spectroscopy or Thioflavin T staining. Different experimental methods are not always concurrent, especially when amyloid peptides do not readily form fibrils but oligomers. This may lead to peptide misclassification and mislabeling. Several bioinformatics methods have been proposed for *in-silico* identification of amyloids, many of them based on machine learning. The effectiveness of these methods heavily depends on accurate annotation of the reference training data obtained from *in-vitro* experiments. We study how robust are bioinformatics methods to weak supervision, encountering imperfect training data. AmyloGram and three other amyloid predictors were applied. The results proved that a certain degree of misannotation in the reference data can be eliminated by the bioinformatics tools, even if they belonged to their training set. The computational results are supported by new experiments with IR and AFM methods.

Amyloids are a group of proteins folding into assemblies of insoluble fibrils of very regular and tightly packed β -structures, which resemble a steric zipper. Despite the importance of amyloids, which is related to their roles in various diseases, their formation and unique behavior are not fully explained¹. One of the challenges associated with amyloid studies is to establish computationally, whether a protein can form amyloids. Currently available tools addressing this question use statistical and physical models^{2,3}. The statistical methods are only based on the amino acid composition of previously annotated amyloid and non-amyloid proteins and use computational models recognizing regularities in the sequences^{4–6}. The physical models, on the other hand, determine folding of proteins into fibrils and use structural constraints^{7–9}. All these methods first require reference data, i.e. a collection of sequences and/or structures of proteins labeled with their ability or inability to form amyloid fibrils. This information is crucial and its imperfection may introduce a bias into prediction methods¹⁰. However, the process of labeling potential amyloid sequences and confirming the ability to form amyloid fibrils is costly and laborious, usually involving a set of diverse experiments.

Amyloids can be recognized by a characteristic cross- β sheet diffraction pattern observable in X-ray studies. However, to identify the occurrence of an amyloid, less precise methods are usually applied, some of which are direct and others indirect. Direct methods involve microscopy and spectroscopy^{11,12}. High resolution microscopic techniques, such as atomic force microscopy (AFM) or transmission electron microscopy (TEM), allow for direct examination of amyloid fibril structures. These methods are focused on their topology and mechanical

¹Department of Biomedical Engineering, Wrocław University of Science and Technology, 50-370 Wrocław, Poland. ²University of Lorraine, CNRS, 54000 Nancy, France. ³Medical University of Białystok, 15-089 Białystok, Poland. ⁴Institute of Biochemistry and Biophysics, Polish Academy Sciences, 02-106 Warsaw, Poland. ⁵Faculty of Biotechnology, University of Wrocław, 50-137 Wrocław, Poland. ⁶Life Sciences Center, Institute of Biotechnology, Vilnius University, 01513 Vilnius, Lithuania. ⁷These authors contributed equally: Michał Burdukiewicz and Małgorzata Kotulska. ✉email: michalburdukiewicz@gmail.com; malgorzata.kotulska@pwr.edu.pl

properties, such as Young modulus^{13,14}. Spectroscopic methods involve vibrational spectroscopy¹⁵, especially IR spectroscopy¹⁶. In addition to precise information about the kinetics of self-assembly and details about their secondary structures, spectroscopic methods reveal the fraction of amyloid aggregates in the structure.

Indirect techniques rely on the detection (usually through fluorescence) of probes selectively binding to amyloid fibrils. Thioflavin T (ThT) is considered to be the most reliable probe¹⁷, but Congo Red can also be applied¹⁸. Although indirect methods are less expensive, there are some concerns regarding their specificity¹⁹. Therefore, it is helpful if such methods are complemented with direct experimental verification.

As direct and indirect methods focus on different aspects of amyloid fibrils, their results may differ. The problem of experimental validation is further heightened by the elusiveness of amyloid properties²⁰. Experimental conditions, such as incubation time, pH and ionic strength, may greatly affect the kinetics of self-assembly, which effectively prevent the development of amyloid fibrils²¹. Therefore, even experimental results bring only partial confidence into the amyloid properties of a peptide or protein.

Such a situation leads to a classical problem of weak labeling (weak supervision)²², where some labels (amyloid or non-amyloid) are wrongly assigned to reference instances (proteins or peptides). The weak supervision is common in all applications of machine learning and significantly lowers the performance of a model. Among several approaches proposed to solve this issue, it is suggested to detect mislabeled training data by applying a computational model as a filter, capable of identifying outliers²³. Here, the outliers are defined as instances predicted computationally with a high probability to have a label opposite to that obtained from a reference dataset. This approach can enhance the classification accuracy achieved by learning algorithms by improving the quality of training data. However, a potential obstacle should be considered, related to overfitting of prediction methods, which may not so easily find mislabeled data in their own training data sets.

To investigate the impact of weak supervision in computational prediction of amyloid proteins, we decided to test AmyloGram, as a filter on training data, which may be mislabeled in databases. The objective was verifying the filtering approach and detecting possible outliers in the learning set. To do this, we selected a subset of peptides for which bioinformatics predictions by AmyloGram were opposite to their labels assigned in experimental AmyLoad and Waltz databases^{24,25}. The most extreme outliers, with the highest probability of a predicted label being opposite to that in databases, were then evaluated experimentally. It allowed to verify if the filtering properties of AmyloGram were sufficient to clean the training data from doubtful instances. To strengthen the analysis, we also tested three different bioinformatics predictors of amyloids in this regard. The results revealed how robust are bioinformatics predictors of amyloids to errors in learning datasets.

Materials and methods

Data selection. Peptides were uploaded from AmyLoad²⁴ database. The original dataset used for training AmyloGram included 421 amyloid peptides and 1044 non-amyloid peptides (1465 sequences in total). In terms of their amyloid propensities, all these peptides were also identically annotated in Waltz 2.0 database²⁵. The flow chart of the data selection procedure is presented in Fig. 1. First, all sequences with six residues (hexapeptides) and without atypical amino acids were selected. The obtained set included 1088 sequences. It was then divided into two subsets, based on their origin. The first subset contained 158 (67 amyloid and 91 non-amyloid) sequences which were based on the original AmylHex database²⁶, and the other set of 930 (180 amyloid and 750 non-amyloid) sequences was based on instances from other sources. AmylHex was the first available data set of amyloid peptides and, although still valuable, it has a strongly biased pattern related to the method by which it was obtained. Therefore, the division in our data processing was introduced to avoid overrepresentation of the AmylHex sequences in the final set and diminish the influence of these biases. Then, all non-redundant amino acid sequences of hexapeptides were converted into the simplified amino acid alphabet obtained in AmyloGram and redundant sequences were removed, leading to 184 encoded amyloid sequences and 683 encoded non-amyloid sequences⁴. Importantly, each of these sequences previously belonged to the reference training dataset and were used to develop AmyloGram.

Since the original experimental annotations do not necessarily have to agree with the classifications obtained with a computational method, the peptides were again classified, now computationally, with AmyloGram (AmyloGram available at: <http://www.smorfland.uni.wroc.pl/shiny/AmyloGram/>). Peptides that obtained a high probability of classification in agreement with their original database annotations were defined as references. Peptides with a high probability of labels opposite to their original database annotations were defined as outliers. Finally, 10 sequences out of the references were selected and represented with the full amino acid alphabet—we denote this dataset as the *reference dataset*. Similarly, 24 sequences from outliers (represented here with the full amino acid alphabet) were selected and labeled as the *test dataset*. Both sets were used in further experimental validations. The first set served to set up and validate our experimental and chemometric methods, while the other to verify whether the original database annotations of the peptides were correct.

Materials. All hexapeptide sequences selected for experimental validation were provided by CASLO (CASLO ApS, Denmark). The experiments were carried out on 34 sequences, out of which 10 were *reference* sequences (FNPQGG, FTFLQF, ISFLIF, KPAESD, LVFYQQ, NPQGGY, SFLIFL, TKPAES, YLLYYT, YTVIIE), and 24 were *test* sequences (ALEEYT, ASSSNY, DETVIV, ELNIYQ, FGELFE, FQKQKQ, FTPTEK, HGFNQQ, HLFNLT, HSSNNE, MIENIQ, MIHFGN, MMHFGN, NIFNIT, NNSGPN, NTIFVQ, QANKHI, QEMRHF, SHVIIIE, STTIE, STVVIE, SWVIIIE, WSYFLL, YYTEFT). The purity of synthesized peptides was in the range between 95% and 99.6%.

Sample preparation. First, lyophilized hexamers were dissolved and vortexed in 0.1 M NaOH. Next, phosphate-buffered saline (50 mM, pH 7.2) was added to obtain pH = 7. Samples were diluted to the final con-

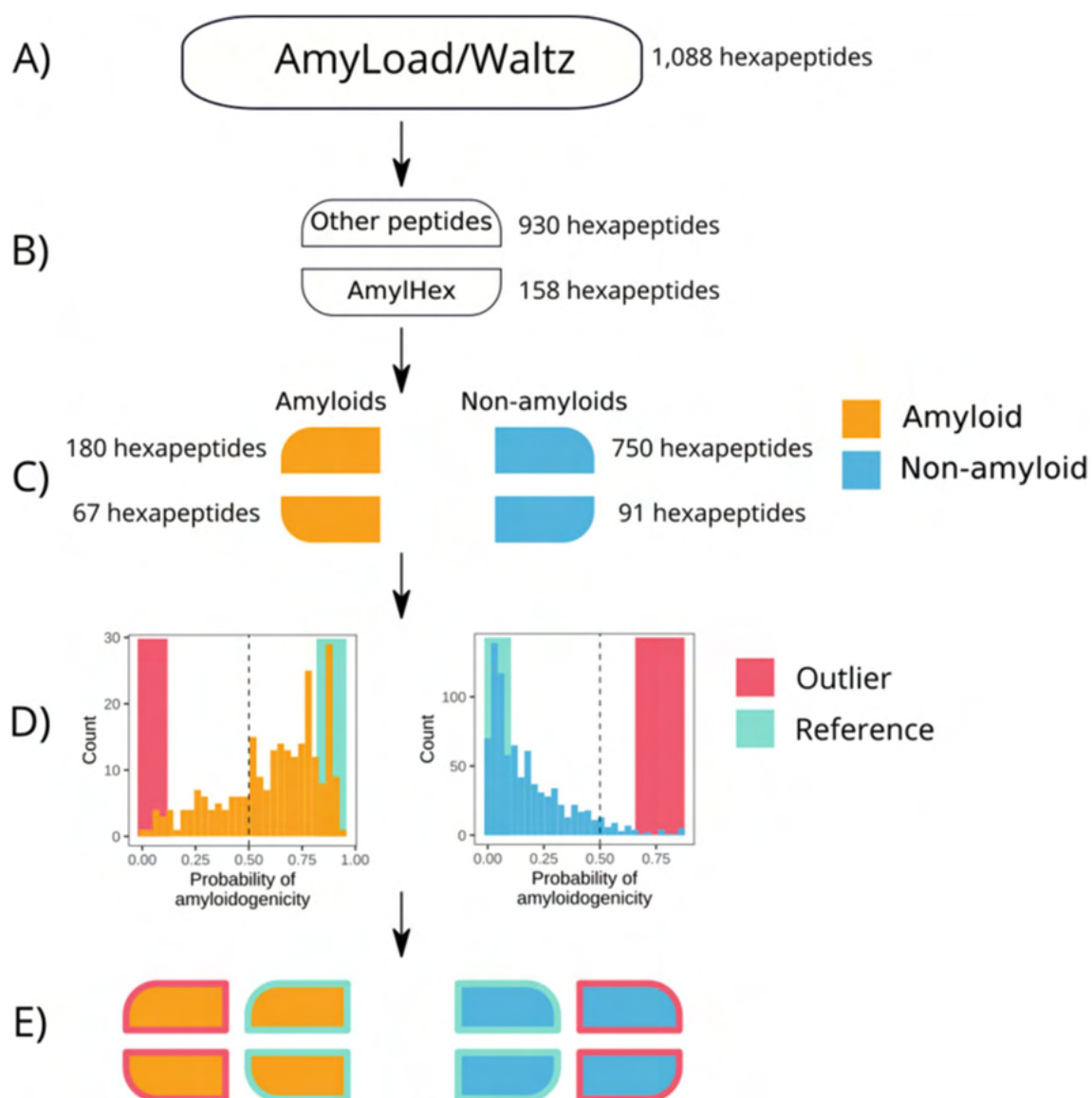


Figure 1. Scheme of peptide selection. (A) 1088 hexapeptides in the simplified amino acid alphabet were used to train AmyloGram. (B) Two subsets of the sequences were defined. (C) Sequences were divided into amyloids and non-amyloids according to their annotations in the database. (D) Each peptide was classified with AmyloGram. Peptides with a high probability of classification in agreement with their original annotations were defined as references. Peptides with a high probability of classification opposite to their original annotations were defined as outliers. (E) Ten references and 24 outliers were selected for experiments.

centration of 4 mg/ml with Milli-Q water. Then, they were incubated at 37 °C for one month. To assure the reproducibility of new experimental results, reported in this work, the table based on the MIRRAGGE protocol²⁷ is available in the Supplement 1, 2, Table 1.

Experimental evaluation. To keep the experimental validation robust, we employed three direct techniques: two methods of IR spectra measurements and AFM. They complement each other in terms of the presence of aggregates and the exact morphology of fibrils.

Atomic force microscopy. AFM images were recorded using Dimension Icon (Bruker) atomic force microscope operating in tapping mode and equipped with a silicon cantilever RTESPA-300 (40 N/m, Bruker), with a typical tip radius of curvature 8 nm. Images (4×4 , 5×5 and $10 \times 10 \mu\text{m}^2$) of sample topography were recorded at the resolution of 1024×1024 pixels. The scan rate was 0.5–1.0 Hz. In each experiment, 20 μl of peptide solution was deposited on freshly etched mica surface and incubated for 10 min. Subsequently, samples were rinsed with 1 ml of MilliQ water and dried under gentle airflow.

| No | Sequence | Database | IR microscopy | | ATR-FTIR | | AFM | Consensus with database annotation |
|----|----------|----------|------------------------------|-------|-------------------------------|-------|------|------------------------------------|
| | | | Amide I [cm^{-1}] | Class | Amide I [cm^{-1}] | Class | | |
| 1 | FNPQGG | No | 1679(m)/ 1641(s) | No | 1655(s,br) | No | No | Yes |
| 2 | FTFIQF | Yes | 1689(m,sh)/ 1628(s) | Yes | 1690(w)/ 1622(s) | Yes | Yes* | Yes |
| 3 | ISFLIF | Yes | 1689(m,sh)/ 1631(s) | Yes | 1685(w)/ 1631(s) | Yes | Yes | Yes |
| 4 | KPAESD | No | 1665(s,br) | No | 1678(s,br) /1640(m,sh) | No | No | Yes |
| 5 | LVFYQQ | Yes | 1631(s) | Yes | 1683(w,sh)/ 1629(s) | Yes* | Yes | Yes |
| 6 | NPQGGY | No | 1658(s,br) | No | 1658(s,br) | No | No | Yes |
| 7 | SFLIFL | Yes | 1689(m)/ 1633(s) | Yes* | 1632(s) | Yes | Yes* | Yes |
| 8 | TKPAES | No | 1652(s,br) | No | 1678(s) /1640(sh) | No | No | Yes |
| 9 | YLLYYT | Yes | 1686(m,sh)/ 1629(s) | Yes | 1685(m)/ 1630(s) | Yes | Yes* | Yes |
| 10 | YTVIIE | Yes | 1685(m)/ 1627(s) | Yes | 1684(m)/ 1626(s) | Yes | Yes | Yes |

Table 1. Reference data set of sequences and their amyloid propensity by different experimental methods ('Yes'—identified as amyloid, 'No'—non-amyloid, 'Yes*'—oligomer, 's'—strong band, 'm'—medium band, 'w'—weak band, 'br'—broad band, 'sh'—shoulder band, band maxima in bold). The results agree with the original database annotations, which were also in agreement with AmyloGram predictions.

Infrared spectroscopy. Two vibrational spectroscopic techniques²⁸, commonly used in the field of peptide aggregation, were used in the study: Attenuated Total Reflection—Fourier Transform Infrared (ATR-FTIR)²⁹, and Fourier Transform Infrared Microscopy using transmission mode (IR microscopy)³⁰. The main drawback of examining proteins in aqueous solutions by means of IR spectroscopy is strong absorbance of water in the region of approximately 1634 cm^{-1} ³¹. Therefore, in our procedures of spectroscopic measurements we used a dry-film technique³².

The ATR-FTIR spectra were collected using a Nicolet 6700 spectrometer (Thermo Scientific, USA) equipped with ATR Accessory with Heated Diamond Top-plate (PIKE Technologies, USA). The spectrometer was continuously purged with dry air. Peptides aliquots of 20 μl volumes were pipetted onto the ATR crystal and allowed to dry out. Spectra were recorded with a resolution of 4 cm^{-1} with 128 co-added scans over the range of 3600–150 cm^{-1} , at the constant temperature of 25 °C. The background spectrum was recorded before measurement of the sample spectra using 512 scans under resolution 4 cm^{-1} .

The spectra from IR microscopy were recorded using Nicolet iN10 FTIR microscope (Thermo Scientific, USA). Samples were measured with a liquid nitrogen cooled mercury cadmium telluride (MCT-A) detector at the spatial resolution of 10 μm . The microscope was continuously purged with dry air. An area of 450 $\mu\text{m} \times 450 \mu\text{m}$ was first selected with the upper aperture (100/5 = 50 μm), then the data were collected. All spectra were recorded in the wave number range from 4000 to 500 cm^{-1} ; 64 interferograms per sample at the resolution of 4 cm^{-1} were collected. The volume of 10 μl of the solution was applied to barium chloride window cell and allowed to dry out until the coffee-ring was formed³³. The measurements were carried out at room temperature. For each spectral map the average spectrum was calculated.

Using two IR methods with different acquisition modes allowed us to verify the observations and avoid ambiguity that may arise due to high water absorption³⁴. ATR-FTIR spectrophotometer provides one average single spectra obtained from a small area (typically of 3 mm^2). The FTIR microscopy allows for mapping the probe with a step of 10 μm or less. The liquid nitrogen cooled MCT-A detector is more sensitive and allows to measure smaller aliquots. The built-in camera allows to choose a region of interest, significant for non-homogeneous deposition patterns, created in film techniques. Although IR microscopy is a more precise method and was finally selected as our reference experimental method, we also examined whether ATR-FTIR, which is a cheaper and a more widespread method, would provide different annotations of the peptides.

Spectroscopic data processing. All spectra were analyzed using the OriginPro 2019 program (OriginLab Corporation, USA). The spectra preprocessing included: baseline correction³⁵ and normalization for the Amide I band maximum. The second derivative (DII)³⁶ was performed in the range of 1720–1580 cm^{-1} to identify the local maximum of the component bands. The second derivative spectra were smoothed with the Savitzky-Golay filter (parameters: polynomial order 2, window 30)³⁷.

Chemometric analysis. For both types of the IR spectra, Principal Component Analysis (PCA)^{38,39} was performed on DII of the described region, using PCA function from scikit-learn Python library⁴⁰ with default parameters.

Bioinformatics methods. The hexapeptide sequences were classified by bioinformatics methods, such as AmyloGram⁴ (<http://www.smorfland.uni.wroc.pl/shiny/AmyloGram/>), PATH⁴¹ (in-house software), FoldAmyloid⁶ (<http://bioinfo.protres.ru/fold-amyloid/>), and PASTA 2.0⁹ (<http://old.protein.bio.unipd.it/pasta2/>). AmyloGram is a tool based on machine learning methods, FoldAmyloid and PASTA 2.0 are based on physical models, whereas PATH is our latest method combining physical modeling with machine learning. AmyloGram and PATH were previously trained on the reference peptide sequences, which included all

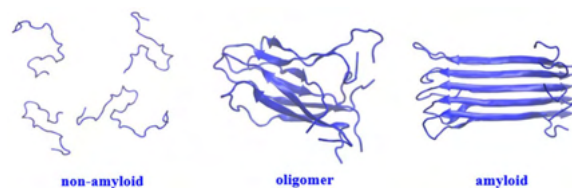


Figure 2. Schemes of peptide classes, representing a general idea.

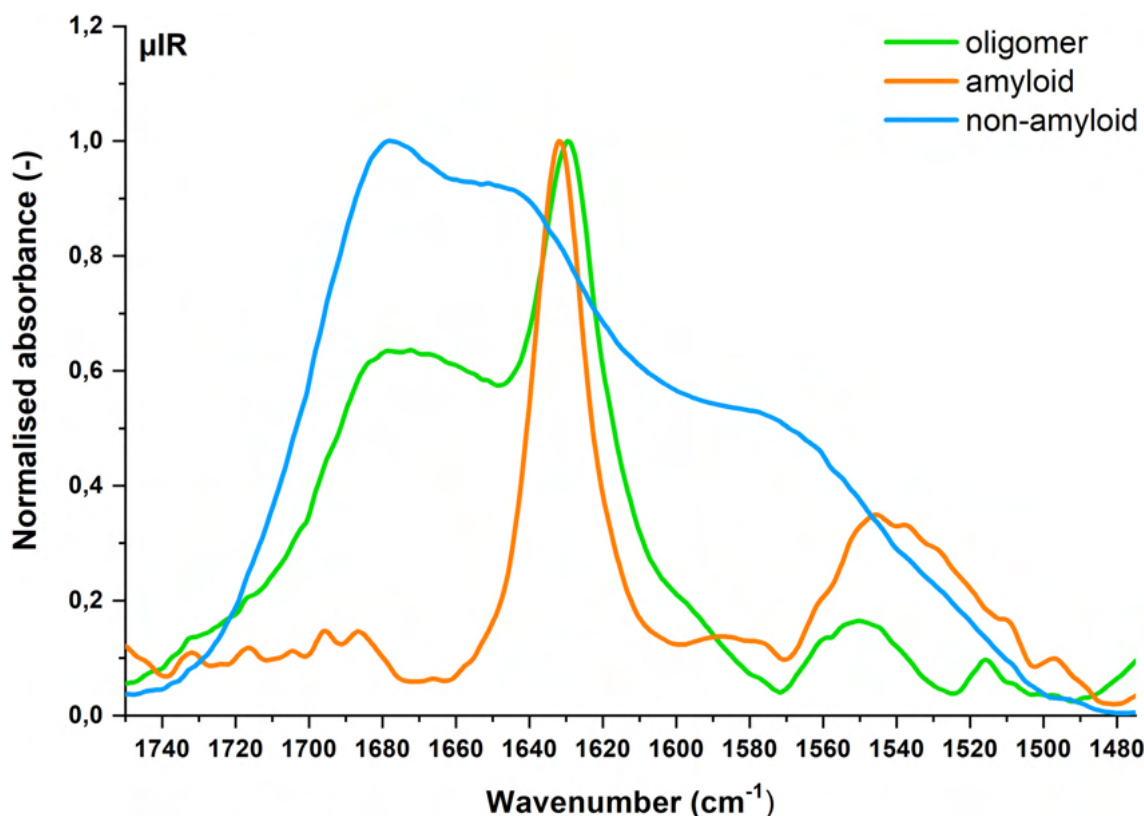


Figure 3. Representative IR microscopy spectra: amyloid (LVFYQQ) in red, oligomer (SFLIFL) in green, non-amyloid (KPAESD) in blue.

sequences verified here anew (*reference* and *test sets*), using their original annotations in the database. All predictors, excluding PASTA 2.0, were used with their default parameters. In PASTA 2.0, the *peptide* option was chosen to set the thresholds. The presented statistics of classification results included: Accuracy (*Acc*) calculated as the ratio of correctly assigned data labels, Sensitivity (*Sn*) denoting the ratio of correctly identified true positives versus actual positives, and Specificity (*Sp*) meaning the ratio of true negatives versus actual negatives.

Results

Experimental verification of the reference dataset of sequences. First, we examined the *reference set*, whose instances had identical annotations in reference databases (AmyLoad and Waltz) and classifications by AmyloGram. The direct microscopy method AFM and two IR methods (ATR-FTIR and IR microscopy) were used to experimentally verify these instances, as well as calibrate our empirical and chemometric methods.

Based on the AFM micrographs (Supplement 1, 1.1) and spectral characteristics (Supplement 1, 2.1 and 2.2), peptides were annotated into three classes: positive (amyloids), negative (non-amyloids), and oligomers (Fig. 2). The last class is not considered by any bioinformatics method but is evident in experimental analyses and may pose a problem for computational tools in its correct classification.

The IR spectra can be fairly easily analyzed in terms of potential amyloidogenicity of the peptides, showing different characteristics for non-amyloids, small assemblies of amyloid aggregates known as oligomers, and mature fibrils. Exemplary spectra of our *reference set*, representing each of these classes, are presented in Fig. 3.

Amide bands characteristic of peptide bonds dominate in the protein infrared spectra. The most intensive, Amide I, occurs in the range of 1700–1600 cm^{-1} , which corresponds to C=O stretching vibrations³⁴. Amyloid fibrils show absorbance between 1611 and 1630 cm^{-1} , usually close to 1630 cm^{-1} , while for native β -sheet proteins it extends from 1630 to 1643 cm^{-1} . This method also enables recognition of typical amyloid oligomers, indicated

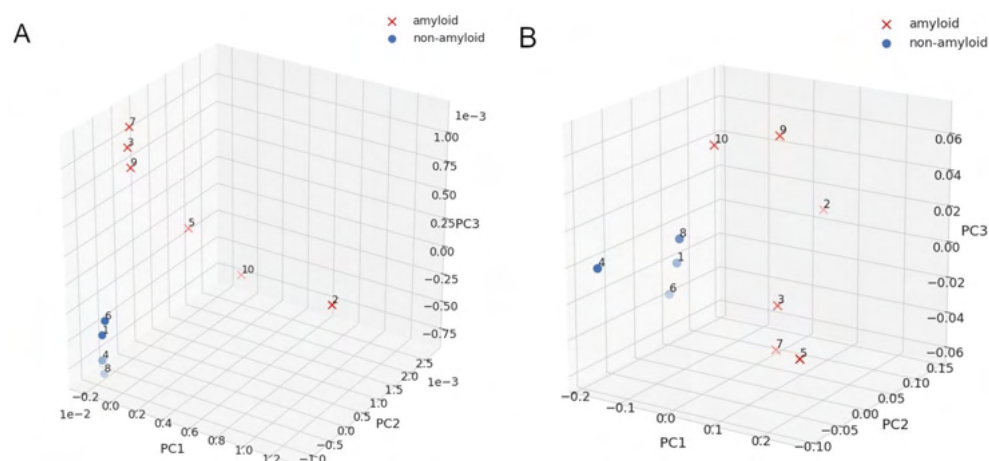


Figure 4. PCA plots for IR spectra of the *reference set*: (A) ATR-FTIR. (B) IR microscopy. Crosses denote amyloids and dots represent non-amyloids, as identified on the spectra by a human expert.

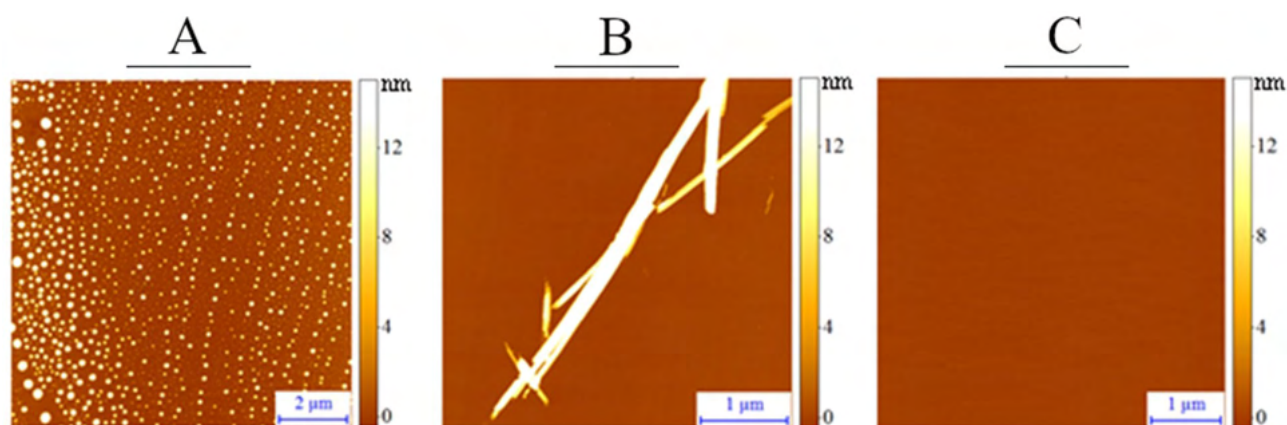


Figure 5. Representative AFM micrographs: (A) oligomer (FTFIQF), (B) amyloid (LVFYQQ), C. non-amyloid (NPQGGY).

by the presence of two local maxima in Amide I region. The major one is located at 1630 cm^{-1} , and the minor peak, resulting from a strong dipolar coupling, ranges between 1695 and 1685 cm^{-1} . The latter peak is often approximately five-fold weaker than the absorption at 1630 cm^{-1} (Fig. 3)^{29,35,36}.

Both IR methods, used in our studies, provided compatible results. As expected, they were in general agreement with their original annotations in the databases (Table 1). However, there were differences, which may have resulted from the experimental specifics (see Materials and Methods), or the oligomer class. The sequence SFLIFL provided slightly different spectra in both IR methods: transmission (microscopy) and attenuated reflection (ATR-FTIR) (Table 1 and Supplement 1, 2.4, Table 7), indicating formation of oligomers which did not transform into fibrils.

The differences may be caused by the artifacts incited by the thickness of the sample—thicker samples can raise the spectrum in the transmission mode in IR microscopy. On the other hand, the signal registered with ATR-FTIR could be influenced by water molecules in contact with the crystal⁴². The contact of peptide molecules with the diamond surface in ATR-FTIR can accelerate the aggregation process. Therefore, IR microscopy could be regarded as a more accurate experimental method. The study confirmed that infrared spectroscopy could be used as a time-efficient tool to investigate the formation of different types of aggregates.

Furthermore, for fast and more robust identification of amyloids and non-amyloids, we applied principal component analysis (PCA) on the IR spectra^{38,39}. PCA separated out 4 sequences in the ATR-FTIR spectra of the *reference set*: NPQGGY, FNPQGG, KPAESD, TKPAES. All these sequences were identified as non-amyloids by a human expert based on different experimental methods. Each of the remaining sequences, more dispersed in the plot, was previously identified either as an amyloid or oligomer—based on the same experimental methods. Similarly, PCA for IR microscopy spectra also distinguished the group of non-amyloid peptides (Figs. 4A,B).

| No | Sequence | IR microscopy | AmyloGram | FoldAmyloid | PASTA 2.0 | PATH (LR) | PATH (RF) | Consensus with IR (%) |
|----|----------|---------------|-----------|-------------|-----------|-----------|-----------|-----------------------|
| 1 | FNPQGG | No | No | No | No | No | No | 100 |
| 2 | FTFIQF | Yes | Yes | Yes | No | Yes | Yes | 80 |
| 3 | ISFLIF | Yes | Yes | Yes | Yes | Yes | Yes | 100 |
| 4 | KPAESD | No | No | No | No | No | No | 100 |
| 5 | LVFYQQ | Yes | Yes | Yes | No | Yes | Yes | 80 |
| 6 | NPQGGY | No | No | No | No | No | No | 100 |
| 7 | SFLIFL | Yes* | Yes | Yes | Yes | Yes | Yes | 100 |
| 8 | TKPAES | No | No | No | No | No | No | 100 |
| 9 | YLLYYT | Yes | Yes | Yes | No | Yes | Yes | 80 |
| 10 | YTVIIE | Yes | Yes | Yes | Yes | No | Yes | 80 |

Table 2. Reference sequences and their amyloid propensity obtained by different bioinformatic methods, compared to IR microscopy ('Yes'—amyloid, 'No'—non-amyloid, 'Yes*'—oligomer).

The results obtained by means of IR spectroscopy were verified with high resolution microscopy using AFM (Fig. 5, Supplement 1, 2.1, Table 2). In these studies, the process of hexapeptide self-assembly was observed a few minutes after preparation of the peptide solution.

Bioinformatics analysis of the reference dataset. The annotations based on IR microscopy results were compared with all bioinformatics methods, including not only AmyloGram, but also FoldAmyloid, PASTA 2.0 and PATH (Table 2). Generally, all methods recognized the sequences correctly and in agreement with IR spectroscopy. Concurrence of the IR microscopy and computational results was at a high level, reaching 75 or 100%. We want to emphasize that due to the very small size of the set and the method of its selection (based on the strong prediction probabilities by AmyloGram), the prediction results from different bioinformatics methods by no means should be treated as benchmarks of their individual general performances.

Annotations of sequences in the test dataset. The experiments on the *reference dataset* showed that IR spectroscopy is in good agreement with much more laborious and expensive AFM method. Therefore, IR spectroscopy was selected for experimental validation of the annotations in the *test set*, which was the main objective of our studies. The results obtained for 24 sequences that constituted this set are presented in Table 3. These data did not take into account the component bands from aromatic amino acids, such as: phenylalanine (1600), tyrosine (1616) and tryptophan (1620)⁴³.

Out of 24 hexapeptides, only one peptide, STTIIIE, gave an ambiguous result in terms of IR spectroscopic methods (Table 3 and Supplement 1, 3.2, Table 12). For STTIIIE, we observed in IR microscopy two local maxima, 1657 cm⁻¹ corresponding to the strong band from α -helix and 1607 cm⁻¹ assigned to tyrosine vibrations. Therefore, this peptide was labeled as non-amyloid. Although Amide I band is very broad, there are many component bands, which are confirmed by the second derivative (Supplement 1, 3.1.2.2., Table 11). This fact cannot exclude that the oligomerization process could have occurred. However, based on the ATR-FTIR, this structure can be identified as oligomer, therefore in terms of classification by bioinformatics tools—positively. Two local maxima characteristic of oligomers can be observed in the spectrum. The first maximum at 1684 cm⁻¹ and the second, more intense, at 1633 cm⁻¹ (Supplement 1, 3.2). The spectral features can be assigned to anti-parallel oligomeric β -sheets. For the remaining 23 sequences both IR techniques provided consistent results.

Based on the results presented in Table 4, we observed that in the *test set*, for which AmyloGram's classification disagreed with the original database annotations, 17 (71%) peptides were indeed misannotated, 12 (70%) of them were false positives and 5 (30%) were false negatives. In the set of misannotated sequences, five were actually amyloids and all of them (100%) were misannotated, while 19 were non-amyloids and 12 (63%) of them were misannotated. A variety of reasons could have contributed to it, which is shown in Supplement 2, Table 1.

Importantly, all these sequences were previously used for training of AmyloGram, using the misannotated labels. However, AmyloGram was capable of recognizing misannotated instances in its training dataset, which showed its robustness with regard to incorrect labeling. Only 7 sequences out of this set were correctly annotated in the database and misclassified by AmyloGram. The majority of them were sequences rich in aromatic and charged amino acids.

IR spectra of the *test set* were analyzed with PCA. Similar to the *reference set*, a good separation between amyloids and non-amyloids (as previously identified by the human expert) was obtained for majority of the sequences (Fig. 6), especially good agreement was obtained for the data from IR microscopy (Fig. 6B). The automated PCA analysis on the spectra from ATR-FTIR located the sequence no 20 (STTIIIE), which was ambiguous with regard to IR experiments, outside the amyloid and non-amyloid clusters. As expected, PCA based on the spectra from the IR microscopy assigned it to the cluster of non-amyloids. A few other sequences were also located outside the aggregated clusters, either in the PCA analysis on ATR-FTIR or IR microscopy, but there was no overlap between them, except the sequence no 4 (ELNIYQ). Interestingly, although this sequence was experimentally verified as non-amyloid, it was predicted by AmyloGram and FoldAmyloid as a potential amyloid.

| No | Sequence | Database | IR microscopy | | ATR-FTIR | | Consensus with database annotation |
|----|----------|----------|-----------------------------|-------|---------------------------------------|-------|------------------------------------|
| | | | Amide I [cm ⁻¹] | Class | Amide I [cm ⁻¹] | Class | |
| 1 | ALEEYT | Yes | 1655(s,br) | No | 1654(s) | No | No |
| 2 | ASSSNY | Yes | 1649(m,sh) | No | 1655(m,br) | No | No |
| 3 | DETVIV | No | 1685(w)/ 1635(s) | Yes* | 1685(m)/ 1633(s) | Yes* | No |
| 4 | ELNIYQ | No | 1661(w,sh)/ 1635(s) | No | 1681(m,br)/1668(m,br)/ 1635(s) | No | Yes |
| 5 | FGELFE | No | 1660(s) /1650(w) | No | 1659(s) | No | Yes |
| 6 | FQKQK | No | 1660(s,br) | No | 1682(s,br) | No | Yes |
| 7 | FTPTEK | No | 1660(s,br) | No | 1680(s,br) | No | Yes |
| 8 | HGFNQK | Yes | 1662(s,br) | No | 1682(s,br) | No | No |
| 9 | HLFNLT | Yes | 1674(s,br) | No | 1680(s,br) /1633(m,br) | No | No |
| 10 | HSSNNF | Yes | 1649(m,br) | No | 1680(s) /1646(m,sh) | No | No |
| 11 | MIENIQ | Yes | 1656(s,br) | No | 1655(s,br) | No | No |
| 12 | MIHFGN | Yes | 1677(s,br) | No | 1680(s,br) /1646(m,br) | NO | NO |
| 13 | MMHFGN | Yes | 1675(s) | No | 1676(s,br) | No | No |
| 14 | NIFNIT | Yes | 1657(s) | No | 1663(s,br) | No | No |
| 15 | NNSGPN | Yes | 1676(sh)/ 1648(s,br) | No | 1676(s,br) /1654(m,br) | No | No |
| 16 | NTIFVQ | No | 1629(s) | Yes | 1682(w)/ 1631(s) | Yes* | No |
| 17 | QANKHI | Yes | 1680(s,br) | No | 1681(s) /1653(sh) | No | No |
| 18 | QEMRHF | Yes | 1679(s,br) | No | 1676(s,br) /1655(sh) | No | No |
| 19 | SHVIIIE | No | 1688(m)/ 1630(s) | Yes | 1684(m)/ 1633(s) | Yes | No |
| 20 | STIIIE | No | 1657(s,br) | No | 1681(m)/ 1630(s) | Yes* | Yes ambiguous |
| 21 | STVVIE | No | 1685(w,br)/ 1633(s) | Yes | 1682(w,br)/ 1630(s) | Yes* | NO |
| 22 | SWVIIIE | No | 1682(w,sh)/ 1631(s) | Yes | 1684(w)/ 1631(s) | Yes | No |
| 23 | WSFYLL | No | 1658(s,br) | No | 1675(w,sh)/ 1637(s) | No | Yes |
| 24 | YYTEFT | No | 1665(s,br) | No | 1659(s,br) | No | Yes |

Table 3. Test sequences and their amyloid propensities ('Yes'—identified as amyloid, 'No'—non-amyloid, 'Yes*'—oligomer, 's'—strong band, 'm'—medium band, 'w'—weak band, 'br'—broad band, 'sh'—shoulder band, band maxima in bold), compared with the original database annotation (all in disagreement with AmyloGram predictions).

The annotations from IR microscopy for the *test set* were compared with results from other bioinformatics predictors, out of which PATH is another method also trained on the set including the misannotated sequences, which can use either logistic regression (LR) or random forest (RF) classification methods. Except for AmyloGram and PATH, other bioinformatics methods might have not been trained on the misannotated data (methods not developed in our group). The majority of methods agreed with our IR results (Table 4, detailed scores in Supplement 2: Table 2 and Table 3), including the cases in which the original annotation in the database was contradicted by the experiments presented in Table 3. There were a few less obvious instances. For example, the consensus between bioinformatics methods dropped for two sequences: DETVIV and ELNIYQ. In case of DETVIV, the IR microscopy result was also ambiguous—it showed oligomeric rather than fibril aggregates. In case of ELNIYQ, PCA-based classification of the spectra did not locate it in the cluster of non-amyloids. The bioinformatics analysis identified the sequence no 20 (STIIIE), which was ambiguous regarding IR experiments, as non-amyloid (3 out of 4 methods), which agrees with IR microscopy and associated PCA analysis. AmyloGram was the only method which misclassified it as amyloid. Table 5 presents aggregated results of the bioinformatics analysis.

All computational methods correctly identified the majority of misannotated sequences. Again, we want to emphasize that due to the size of the set and the method of its selection (based on the strong adverse predictions by AmyloGram), the prediction results from different bioinformatics methods should not be treated as benchmarks of their general performances.

Discussion

Amyloid aggregates may lead to serious health problems, when peptides enter the amyloid pathway, therefore it is crucial to recognize them correctly and identify specific sequence features, which can be associated with amyloidogenicity. Although several direct and indirect experimental methods are available to determine the amyloid propensity of a sequence, all of them are laborious and expensive. What is even more important, the results of the experiments are not always conclusive and identical, if obtained with different experimental methods. This may lead to misannotation of the sequences regarding their amyloidogenicity. Moreover, errors occurring in databases, related to data retrieval or curation, may additionally contribute to mislabeling of the data.

Many bioinformatics methods have been developed to classify amyloidogenicity of amino acid sequences. These methods readily and efficiently support experiments, saving time and money. However, all

| No | Sequence | Database | IR microscopy | AmyloGram | PATH (LR) | PATH (RF) | FoldAmyloid | PASTA 2.0 | Bioinformatics consensus with IR [%] |
|----|----------|----------|---------------|-----------|-----------|-----------|-------------|-----------|--------------------------------------|
| 1 | ALEEYT | Yes | No | No | No | No | No | No | 100 |
| 2 | ASSSNY | Yes | No | No | No | No | No | No | 100 |
| 3 | DETVIV | No | Yes* | Yes | No | Yes | No | Yes | 60 |
| 4 | ELNIYQ | No | No | Yes | No | No | Yes | No | 60 |
| 5 | FGELFE | No | No | Yes | No | No | No | No | 80 |
| 6 | FQKQKQ | No | No | Yes | No | No | No | No | 80 |
| 7 | FTPTEK | No | No | Yes | No | No | No | No | 80 |
| 8 | HGFNQQ | Yes | No | No | No | No | No | No | 100 |
| 9 | HLFNLT | Yes | No | No | No | Yes | Yes | No | 60 |
| 10 | HSSNNF | Yes | No | No | No | No | No | No | 100 |
| 11 | MIENIQ | Yes | No | No | No | No | No | No | 100 |
| 12 | MIHFGN | Yes | No | No | No | No | No | No | 100 |
| 13 | MMHFGN | Yes | No | No | No | No | No | No | 100 |
| 14 | NIFNIT | Yes | No | No | No | Yes | Yes | No | 60 |
| 15 | NNSGPN | Yes | No | No | No | No | No | No | 100 |
| 16 | NTIFVQ | No | Yes | YES | Yes | Yes | Yes | No | 80 |
| 17 | QANKHI | Yes | No | No | No | No | No | No | 100 |
| 18 | QEMRHF | Yes | No | No | No | No | No | No | 100 |
| 19 | SHVIIIE | No | Yes | Yes | No | No | Yes | Yes | 60 |
| 20 | STIIIE | No | No | Yes | No | No | No | No | 80 |
| 21 | STVVIE | No | Yes | Yes | No | Yes | Yes | Yes | 80 |
| 22 | SWVIIIE | No | Yes | Yes | No | Yes | Yes | Yes | 80 |
| 23 | WSFYLL | No | No | Yes | Yes | Yes | Yes | No | 80 |
| 24 | YYTEFT | No | no | Yes | No | No | No | No | 80 |

Table 4. Test sequences and their amyloid propensities predicted by different bioinformatics methods and compared with IR microscopy ('Yes'—amyloid, 'No'—non-amyloid, 'Yes*'—oligomer). For comparison, the 'Database' column presents original annotations from the databases.

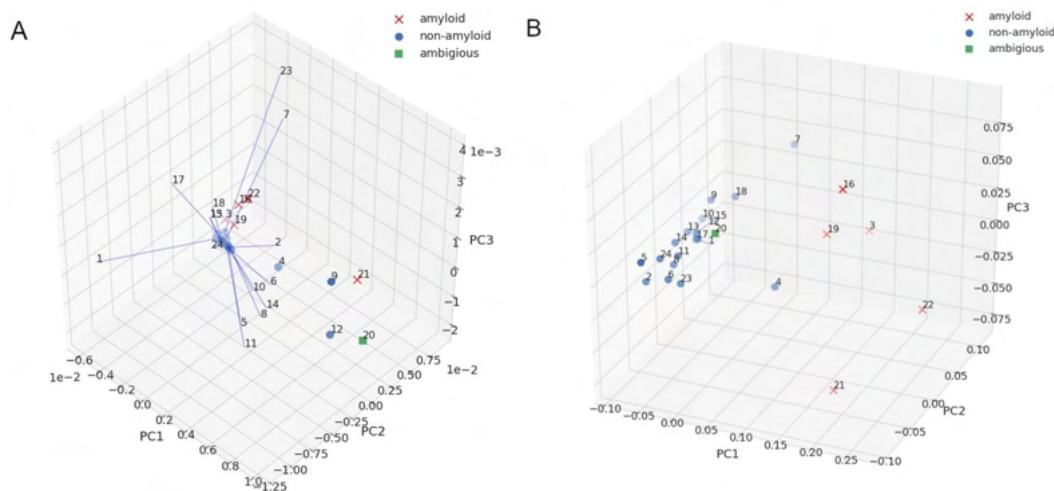


Figure 6. PCA plots for IR spectra of the *test set*: (A) ATR-FTIR. (B) IR microscopy. Crosses denote amyloids and dots represent non-amyloids, as identified on the spectra by a human expert.

computational methods, like modeling in general, heavily depend on the data used in the model construction. Data including misannotated instances may lead to an incorrect model, not even revealed by standard evaluation methods, which would also rely on the mislabeled reference data.

Therefore, we posed a question: How robust could be bioinformatics methods to the problem of certain misannotations in the reference data? The problem occurred when we observed that some of the computational classifications did not always agree with labeling of the reference training data. To address the question, we

| | AmyloGram | | | PATH (LR) | | | PATH (RF) | | | FoldAmyloid | | | PASTA 2.0 | | |
|---|-----------|----|------|-----------|-----|------|-----------|-----|------|-------------|-----|------|-----------|-----|----|
| | Acc | Sn | Sp | Acc | Sn | Sp | Acc | Sn | Sp | Acc | Sn | Sp | Acc | Sn | Sp |
| A | 0.71 | 1 | 0.63 | 0.79 | 0.2 | 0.95 | 0.83 | 0.8 | 0.84 | 0.79 | 0.8 | 0.79 | 0.92 | 0.8 | 1 |
| B | 1 | 1 | 1 | 0.76 | 0.2 | 1 | 0.82 | 0.8 | 0.83 | 0.82 | 0.8 | 0.83 | 0.94 | 0.8 | 1 |

Table 5. Consensus between annotations obtained from bioinformatics methods and IR microscopy (Accuracy *Acc*, Sensitivity *Sn*, Specificity *Sp*). Presented results are for: (A) all 24 sequences from the *test set*, (B) only 17 sequences from the *test set*, which turned out misannotated in databases.

selected a set of sequences and tested their amyloidogenicity by experimental and computational methods. The first part of the set, when classified by our predictor AmyloGram, strongly agreed with the initial labeling in the database, as it was expected. We used it to set up our experimental and chemometric methods, including two IR spectroscopy methods, ATR-FTIR and IR microscopy, and AFM microscopy. The second part of the set included sequences whose classification by AmyloGram strongly disagreed with the initial labeling in the reference databases. Besides amyloids and non-amyloids, we also noted that a third class of structures, i.e. oligomers, should be included in the analyses.

As a result, we observed that 17 out of 24 non-compatible sequences were actually misannotated in the original databases. Therefore, the bioinformatics predictor proved resistant to overfitting, and able to find errors in its own training data. Tests on other bioinformatics predictors showed that all of them were able to classify the misannotated data correctly, with accuracies reaching at least 80% or more—also for methods which were trained on all these mislabeled data. This proves that bioinformatics methods can be successfully applied to evaluate quality of experimental data and used for their filtering. However, we underline that the fraction of mislabeled instances cannot be excessively high in the training set.

Received: 22 September 2020; Accepted: 8 March 2021

Published online: 26 April 2021

References

- Iadanza, M. G. *et al.* A new Era for understanding amyloid structures and disease. *Nat. Rev. Mol. Cell Biol.* **19**(12), 755–773 (2018).
- Navarro, S. & Ventura, S. Computational re-design of protein structures to improve solubility. *Expert Opin. Drug Discov.* **14**(10), 1077–1088 (2019).
- Bondarev, S. A. *et al.* Structure-based view on [PSI⁺] prion properties. *Prion* **9**(3), 190–199 (2015).
- Burdukiewicz, M. *et al.* Amyloidogenic motifs revealed by n-gram analysis. *Sci. Rep.* **7**(1), 12961 (2017).
- Gasior, P. & Kotulska, M. FISH Amyloid—a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of aminoacids. *BMC Bioinformatics* **15**, 54 (2014).
- Garbuzynskiy, S. O., Lobanov, M. Y. & Galzitskaya, O. V. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* **26**(3), 326–332 (2010).
- Bondarev, S. A., Bondareva, O. V., Zhouravleva, G. A. & Kajava, A. V. BetaSerpentine: a bioinformatics tool for reconstruction of amyloid structures. *Bioinformatics* **34**(4), 599–608 (2018).
- Conchillo-Solé, O. *et al.* AGGRESKAN: A server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC Bioinform.* **8**, 65 (2007).
- Walsh, I., Seno, F., Tosatto, S. C. & Trovato, A. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res.* **42**, 301–307 (2014).
- Kotulska, M. & Unold, O. On the amyloid datasets used for training PAFIG—how (not) to extend the experimental dataset of hexapeptides. *BMC Bioinform.* **14**, 351 (2013).
- Adamcik, J. *et al.* Measurement of intrinsic properties of amyloid fibrils by the peak force QNM method. *Nanoscale* **4**(15), 4426–4429 (2012).
- Cristóvão, J. S., Henriques, B. J. & Gomes, C. M. Biophysical and spectroscopic methods for monitoring protein misfolding and amyloid aggregation. *Methods Mol. Biol.* **1873**, 3–18 (2019).
- Ruggeri, F. S., Sneideris, T., Vendruscolo, M. & Knowles, T. P. J. Atomic force microscopy for single molecule characterisation of protein aggregation. *Arch. Biochem. Biophys.* **664**, 134–148 (2019).
- Knowles, T. P. *et al.* Role of intermolecular forces in defining material properties of protein nanofibrils. *Science* **318**(5858), 1900–1903 (2007).
- Martial, B., Lefèvre, T. & Auger, M. Understanding amyloid fibril formation using protein fragments: structural investigations via vibrational spectroscopy and solid-state NMR. *Biophys. Rev.* **10**(4), 1133–1149 (2018).
- Moran, S. D. & Zanni, M. T. How to get insight into amyloid structure and formation from infrared spectroscopy. *J. Phys. Chem. Lett.* **5**(11), 1984–1993 (2014).
- Gade Malmos, K. *et al.* ThT 101: a primer on the use of thioflavin T to investigate amyloid formation [Internet]. *Amyloid* **24**(1), 1–16 (2017).
- Yakupova, E. I. *et al.* Congo Red and amyloids: History and relationship. *Biosci. Rep.* **39**(1), 62 (2019).
- Biancardi, A. *et al.* Mechanistic aspects of thioflavin-T self-aggregation and DNA binding: evidence for dimer attack on DNA grooves. *Phys. Chem. Chem. Phys.* **16**, 2006–2072 (2014).
- Tycko, R. Amyloid polymorphism: structural basis and neurobiological relevance. *Neuron* **86**(3), 632–645 (2015).
- Hoyer, W. *et al.* Dependence of α -synuclein aggregate morphology on solution conditions. *J. Mol. Biol.* **322**(2), 383–393 (2002).
- Zhou, Z.-H. Special topic: machine learning a brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **5**(1), 44–53 (2018).
- Brodley, C. E. & Friedl, M. A. Identifying mislabeled training data. *J. Artificial Intell. Res.* **11**, 131–167 (1999).
- Wozniak, P. P. & Kotulska, M. AmyLoad: website dedicated to amyloidogenic protein fragments. *Bioinformatics* **31**, 3395–3397 (2015).
- Louros, N. *et al.* WALTZ-DB 2.0: an updated database containing structural information of experimentally determined amyloid-forming peptides. *Nucleic Acids Res.* **48**(1), D389–D393 (2020).

26. Thompson, M. J. *et al.* The 3D profile method for identifying fibril-forming segments of proteins. *Proc. Natl. Acad. Sci. USA* **103**(11), 4074–4078 (2006).
27. Martins, P. M. *et al.* MIRRAGGE—minimum information required for reproducible AGGregation experiments. *Front. Mol. Neurosci.* **222**(13), 139 (2020).
28. Li, H., Lantz, R. & Du, D. Vibrational approach to the dynamics and structure of protein amyloids. *Molecules* **24**(1), E186 (2019).
29. Ruyschaert, J. M. & Raussens, V. ATR-FTIR analysis of amyloid proteins. *Methods Mol. Biol.* **1777**, 69–81 (2018).
30. Baker, M. J. *et al.* Using Fourier transform IR spectroscopy to analyze biological materials. *Nat. Protoc.* **9**, 1771–1791 (2014).
31. Barth, A. Infrared spectroscopy of proteins. *Biochim. Biophys. Acta Bioenerg.* **1767**(9), 1073–1101 (2007).
32. Allara, D. & Stapleton, J. Methods of IR spectroscopy for surfaces and thin films. *Springer Ser. Surf. Sci.* **51**(1), 59–98 (2013).
33. Choi, S. & Birarda, G. Protein mixture segregation at coffee-ring: real-time imaging of protein ring precipitation by FTIR spectromicroscopy. *J. Phys. Chem.* **121**(30), 7359–7365 (2017).
34. Sharaha, U. *et al.* Fast and reliable determination of *Escherichia coli* susceptibility to antibiotics: Infrared microscopy in tandem with machine learning algorithms. *J. Biophotonics* **12**(7), e201800478 (2019).
35. Saroukh, R. *et al.* ATR-FTIR: a “rejuvenated” tool to investigate amyloid proteins. *Biochim. Biophys. Acta Biomembr.* **1828**(10), 2328–2338 (2013).
36. Seo, J. *et al.* An infrared spectroscopy approach to follow β -sheet formation in peptide amyloid assemblies. *Nat. Chem.* **9**(1), 39–44 (2017).
37. Savitzky, A. & Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**, 1627–1639 (1964).
38. Baranska, M., Roman, M. & Majzner, K. General overview on vibrational spectroscopy applied in biology and medicine. In *Optical Spectroscopy and Computational Methods in Biology and Medicine* (ed. Baranska, M.) 3–14 (Springer, 2014).
39. Szymanska-Chargot, M. & Zdunek, A. Use of FT-IR spectra and PCA to the bulk characterization of cell wall residues of fruits and vegetables along a fraction process. *Food Biophys.* **8**, 29–42 (2013).
40. Pedregosa, F. *et al.* Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
41. Wojciechowski, J. W. & Kotulska, M. PATH-prediction of amyloidogenicity by threading and machine learning. *Sci. Rep.* **10**(1), 7721 (2020).
42. Goldberg, M. E. & Chaffotte, A. F. Undistorted structural analysis of soluble proteins by attenuated total reflectance infrared spectroscopy. *Protein Sci.* **14**(11), 2781–2792 (2005).
43. Hernández, B. *et al.* Vibrational analysis of amino acids and short peptides in hydrated media. VIII. Amino acids with aromatic side chains: L-phenylalanine, l-tyrosine, and l-tryptophan. *J. Phys. Chem. B.* **114**(46), 15319–15330 (2010).

Acknowledgements

This work was partially supported by the National Science Centre, Poland, Grant 2019/35/B/NZ2/03997 (MK, MB, MGG, JW), National Centre for Research and Development, Poland under POWR.03.02.00-00-I003/16 (NS) and under PWR.03.02.00-00-I037/16-01/16 (JC) and Wroclaw Center of Biotechnology program “The Leading National Research Center (KNOW) for years 2014–2018” (MB, PM, JC). Access to Wroclaw Centre for Networking and Supercomputing is greatly acknowledged. Funding was provided by Wroclawskie Centrum Sieciowo-Superkomputerowe, Politechnika Wroclawska (Grant Number 98).

Author contributions

N.S.: Experimental, Investigation, Writing; M.B.: Conceptualization, Writing, Revision; M.G.-G.: Experimental, Investigation, Writing; J.W.W.: Bioinformatic analysis, Writing; J.C.: AFM studies; P.M.: Conceptualization, Writing, Revision; T.Š.: AFM studies; V.S.: Conceptualization, Writing, Revision; M.K.: Conceptualization, Writing, Revision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-86530-6>.

Correspondence and requests for materials should be addressed to M.B. or M.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021



Article

Variability of Amyloid Propensity in Imperfect Repeats of CsgA Protein of *Salmonella enterica* and *Escherichia coli*

Natalia Szulc ^{1,2} , Marlena Gašior-Głogowska ¹ , Jakub W. Wojciechowski ¹, Monika Szefczyk ³, Andrzej M. Żak ⁴ , Michał Burdukiewicz ^{5,6,7,*} and Malgorzata Kotulska ^{1,*}

¹ Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland; natalia.szulc@pwr.edu.pl (N.S.); marlena.gasior-glogowska@pwr.edu.pl (M.G.-G.); jakub.wojciechowski@pwr.edu.pl (J.W.W.)

² LPCT, CNRS, Université de Lorraine, F-54000 Nancy, France

³ Department of Bioorganic Chemistry, Faculty of Chemistry, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland; monika.szefczyk@pwr.edu.pl

⁴ Electron Microscopy Laboratory, Faculty of Mechanical Engineering, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland; andrzej.zak@pwr.edu.pl

⁵ Clinical Research Centre, Medical University of Białystok, Jana Kilińskiego 1, 15-089 Białystok, Poland

⁶ Institute of Biochemistry and Biophysics, Polish Academy Sciences, 02-106 Warsaw, Poland

⁷ Faculty of Natural Sciences, Brandenburg University of Technology Cottbus-Senftenberg, 01968 Senftenberg, Germany

* Correspondence: michalburdukiewicz@gmail.com (M.B.); malgorzata.kotulska@pwr.edu.pl (M.K.)



Citation: Szulc, N.;

Gašior-Głogowska, M.;

Wojciechowski, J.W.; Szefczyk, M.;

Żak, A.M.; Burdukiewicz, M.;

Kotulska, M. Variability of Amyloid

Propensity in Imperfect Repeats of

CsgA Protein of *Salmonella enterica*

and *Escherichia coli*. *Int. J. Mol. Sci.*

2021, *22*, 5127. [https://doi.org/](https://doi.org/10.3390/ijms22105127)

10.3390/ijms22105127

Academic Editors:

Vytautas Smirnovas and Konstantin

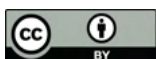
K. Turoverov

Received: 9 April 2021

Accepted: 7 May 2021

Published: 12 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: CsgA is an aggregating protein from bacterial biofilms, representing a class of functional amyloids. Its amyloid propensity is defined by five fragments (R1–R5) of the sequence, representing non-perfect repeats. Gate-keeper amino acid residues, specific to each fragment, define the fragment's propensity for self-aggregation and aggregating characteristics of the whole protein. We study the self-aggregation and secondary structures of the repeat fragments of *Salmonella enterica* and *Escherichia coli* and comparatively analyze their potential effects on these proteins in a bacterial biofilm. Using bioinformatics predictors, ATR-FTIR and FT-Raman spectroscopy techniques, circular dichroism, and transmission electron microscopy, we confirmed self-aggregation of R1, R3, R5 fragments, as previously reported for *Escherichia coli*, however, with different temporal characteristics for each species. We also observed aggregation propensities of R4 fragment of *Salmonella enterica* that is different than that of *Escherichia coli*. Our studies showed that amyloid structures of CsgA repeats are more easily formed and more durable in *Salmonella enterica* than those in *Escherichia coli*.

Keywords: functional amyloids; curli; aggregation; biofilm; ATR-FTIR; FT-Raman

1. Introduction

Functional amyloids are spread across nearly the whole tree of life, including archaea, bacteria, fungi, protozoa, and viruses [1]. Although functional amyloids, similarly to pathological amyloids, self-assemble into fibers, their aggregates are involved in a wide range of crucial molecular tasks, including hormone storage, signaling, enhancing cell adhesion, and biofilm formation [2]. Aside from these functionalities, some bacterial functional amyloids constitute a proteinaceous skeleton of the extracellular matrix, called biofilm. Bacteria produce biofilms to create an environment protecting them from adverse conditions. This ability is widespread in nature and can be seen as one of the most common survival strategies adopted by bacteria [3]. It is estimated that between 40% and 80% of all bacterial cells are part of biofilms [4].

As the ability to form stable biofilms depends on the bacterial natural niche and their genotypic characteristics [5], it is also affected by the phylogenetic variability of functional amyloids involved in this process. Here, we focus on one of the best-studied amyloids

involved in biofilm formation, curli fibers [6]. Curli form non-branched fibrils on the cell surface, which are very resistant to degradation by proteases and detergents. A primary structural component of these fibrils is CsgA protein. The most widely studied functional amyloid is CsgA of *Escherichia coli* (*E. coli*). It is a 151 amino acid long protein, including N terminal signaling peptide, which is proteolytically cleaved, and a core amyloid domain transported outside the cell by CsgG protein [7]. CsgA forms amyloid fibrils along with CsgB protein, which enhances the fibril formation [8]. The curli homologs are prevalent among other Enterobacteriaceae, although in many cases they exhibit a large structural diversity. The CsgA sequence consists of five imperfect repeats labeled as R1–R5 fragments, which, in *E. coli*, follows a common pattern S-[X]5-Q-[X]-G-[X]-G-N-[X]-A-[X]3-Q. The motif depends on bacterial species and it can be altered in curli of other species. Fragments R1 and R5, as the most amyloidogenic, are critical for seeding and the curli formation in *E. coli* [9,10]. The other three fragments are less prone to aggregation or non-aggregating at all.

Despite their variability, it has been shown that even very distant CsgA homologs can together contribute to the formation of the heterogeneous curli fibrils [11]. This phenomenon probably occurs due to the widespread presence of imperfect repeats, which appear more or less commonly in all CsgA and CsgA-like proteins [12].

This ability to interact with other amyloids is also a source of putative negative effect of curli. It has been shown that CsgA produced by the gut microflora can promote aggregation of human proteins by the phenomenon called cross-seeding, for example, facilitating aggregation of α -synuclein or amyloid A β , which are involved in human amyloid diseases [13,14]. This poorly understood phenomenon, termed “mapranosis” (microbiota-associated proteopathy and neuroinflammation), may lead to contribution of the microbiome to neurodegenerative diseases [15].

Currently, one of the challenges in understanding the self-assembly propensity of CsgA is the evolutionary variability of imperfect repeats. To shed more light on the structural determinants of the amyloid propensity of imperfect repeats, we studied the behavior of CsgA homolog from *Salmonella enterica* (*S. enterica*), a common foodborne pathogen that creates many challenges in medicine and food industry [16]. Its CsgA has a slightly altered motif; one of the glycines is not always present (R1 lacking the first glycine, R3 and R4 lacking the second one). Although it is very similar to CsgA from *E. coli*, the aggregation kinetics of these two proteins may be different. Therefore, we compare in vitro the aggregation propensities of CsgA fragments *S. enterica* with *E. coli* strain K12. The knowledge of their properties may lead to better understanding of the principles of functional amyloid aggregation and, thus, help in developing anti-biofilm agents [17].

2. Results

2.1. Sequence Alignment

CsgA proteins from *E. coli* and *S. enterica* are closely related homologues with 75% of identity, as calculated by BLAST [18]. To further investigate differences between corresponding fragments, pairwise alignments were performed for each pair of the fragments (Figure 1). The alignment confirmed that the sequences were highly similar to their counterparts. Almost all of the peptides share similar features: they are rather hydrophilic, with their ends often containing charged amino acids and a highly flexible glycine rich linker in the middle. Such an architecture suggests the propensity to form beta arch structure, which agrees with the computational results of ArchCandy for almost all of these sequences [19] (Table S1 in SI).

in the amyloid propensity of CsgA proteins, depending on their location [21–23]. Not only the contributing residues affect the peptide aggregation susceptibility but also the sequence order is of great importance. This fact was discovered by statistical analyses, which led to the release of several bioinformatics predictors. The sequence alignment of corresponding pairs from both bacterial species shows that different amyloid propensities of the CsgA fragments could not be excluded.

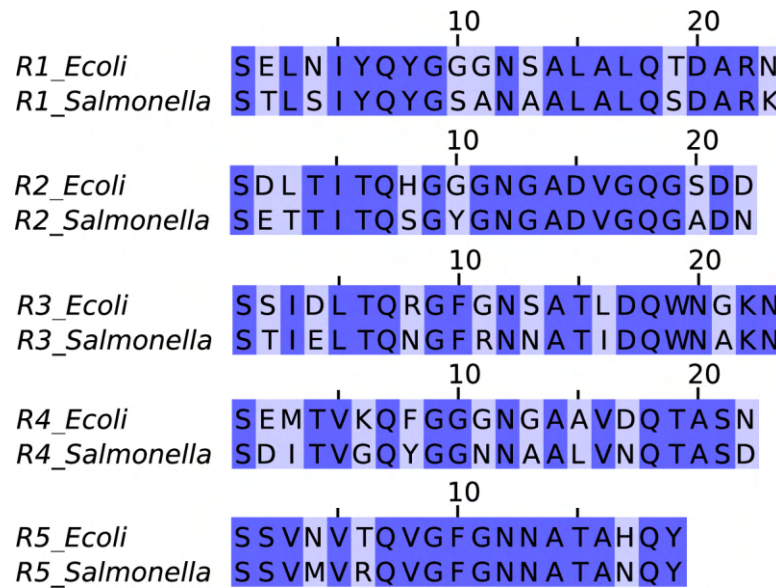


Figure 1. Pairwise sequence alignment between CsgA fragments of *E. coli* and *S. enterica* bacteria. The differences in the amino acid compositions are highlighted with light purple.

2.2. Bioinformatics Analysis

However, their aggregation propensities may differ in terms of physicochemical properties. We applied bioinformatics tools of CsgA genes from *E. coli* and *S. enterica* to identify amino acid positions stabilizing or destabilizing in the peptide structure, thus forming a potential impact of genetic importance to the aggregates [20]. Similarly, the charged residues may play an important role in the amyloid propensity of CsgA proteins, depending on their location [21–23]. Not only the contributing residues affect the peptide aggregation susceptibility but also the sequence order is of great importance. This fact was discovered by statistical analyses, which led to the release of several bioinformatics predictors. The sequence alignment of corresponding pairs from both bacterial species shows that different amyloid propensities of the CsgA fragments could not be excluded.

Therefore, this fragment was selected as a candidate that may have different aggregation propensities. For a comparison, we also applied several other bioinformatic predictors, such as Pasta 2.0 [28], Waltz [29], AmylPred2 [30], FoldAmyloid [31], MetAmyl [32], and Tango [33]. The results of their predictions were not unanimous (Table S1 in SI).

The analysis was performed with our amyloid predictors, which showed a very high accuracy. AmyloGram [24] and PATH [25], both methods were trained on hexapeptides collected in AmyLoad [26] and Waltz 2.0 databases [27]. The only difference between corresponding fragments was obtained for R4. In this case, despite significant similarity of the sequences, AmyloGram provided different results. R4 fragment from *S. enterica* was reported as amyloidogenic, while R4 from *E. coli* was reported as non-amyloidogenic.

Since experimental data on the fragments were only published for *E. coli*, we verified our predictions based on their reported aggregation propensities. None of the predictors provided the classification results in good agreement with experimental data, which report R1, R3, and R5 of *E. coli* as capable of forming aggregates [9]. The best agreement was obtained for the R1 fragment (AmyloGram, PATH, and Waltz). The lack of agreement between computational and experimental results may partly come from the functional character of these sequences. All of the presented methods were trained mostly on fragments of pathological amyloids and their mutants, which somehow showed different characteristics from functional amyloids [34, 35]. However, we believe that these methods are still sensitive enough to detect differences between highly similar fragments, whose

Since experimental data on the fragments were only published for *E. coli*, we verified our predictions based on their reported aggregation propensities. None of the predictors provided the classification results in good agreement with experimental data, which report R1, R3, and R5 of *E. coli* as capable of forming aggregates [9]. The best agreement was obtained for the R1 fragment (AmyloGram, PATH, and Waltz). The lack of agreement between computational and experimental results may partly come from the functional

character of these sequences. All of the presented methods were trained mostly on fragments of pathological amyloids and their mutants, which somehow showed different characteristics from functional amyloids [34,35]. However, we believe that these methods are still sensitive enough to detect differences between highly similar fragments whose scarce point mutations are indicated as potentially leading to changing of the aggregation propensity. Therefore, we expected differences in amyloid aggregation of R4 fragments from *E. coli* and *S. enterica* bacteria, which we tested experimentally.

2.3. Experimental Analysis

Spectroscopic techniques (CD, ATR-FTIR, and FT-Raman) were used to study aggregation propensity of all *E. coli* and *S. enterica* fragments. These methods provide general information about the secondary structure and allow for monitoring of the fibrillization process [36–39]. Finally, we performed transmission electron microscopy (TEM) to analyze morphology of the selected fragments [40].

2.3.1. Circular Dichroism

Circular dichroism (CD) spectroscopy was used to elucidate general characteristics of the secondary structure of the CsgA fragments. CD spectra of *E. coli* fragments are presented in Figure 2A. On the day of sample dissolving, for all *E. coli* fragments, a minimum of ca. 200 nm could be observed in all recorded spectra, which is characteristic of a random coil conformation. The spectra resemble the results presented for the whole CsgA studied by Shu et al. 2012 [41]. The data show that CsgA initially exhibited a random structure; however, after 13 days of incubation, it showed the presence of a β -sheet structure.

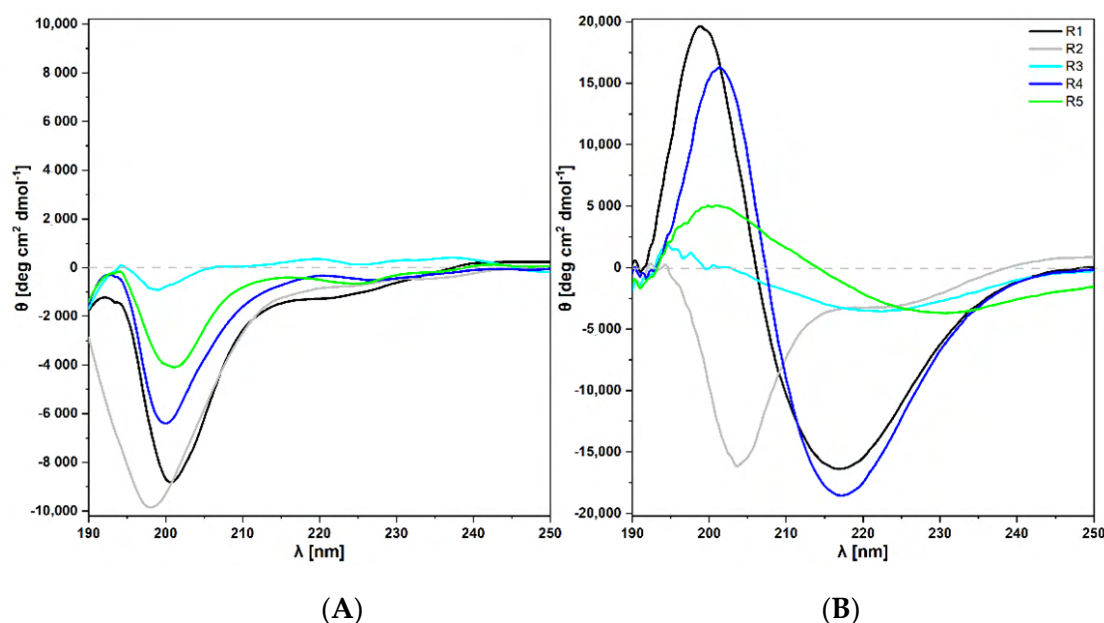


Figure 2. CD spectra of CsgA fragments on the day of the dissolving. (A) Spectra for *E. coli* fragments (R1–R5) and (B) Spectra for *S. enterica* fragments (R1–R5) on the day of the dissolving (final peptide concentration 500 μ M).

CD spectra of *S. enterica* fragments are depicted in Figure 2B. The spectrum of R1 fragment showed a single maximum at 198 nm and a single minimum at 217 nm, which were characteristic of the β -sheet conformation. Analysis of R2 fragment revealed a single maximum at 198 nm and a single minimum at 217 nm, which were characteristic of the β -sheet conformation. Analysis of R3 fragment revealed a single maximum at 194 nm and a broad minimum at 220 nm, which indicated the presence of the β -sheet. R4 fragment showed a maximum at 201 nm and a single minimum at 216 nm, which was also characteristic of the β -sheet conformation. In case of R5 fragment, the broad maximum at 200 nm and broad single minimum at 230 nm could also be assigned to the β -sheet conformation.

In summary, CsgA fragments of *S. enterica* show two types of structures: β -sheet con-

broad maximum at 200 nm and broad single minimum at 230 nm could also be assigned to the β -sheet conformation.

In summary, CsgA fragments of *S. enterica* show two types of structures: β -sheet conformations can be assigned to fragments R1 and R3–R5, the random coil conformation is present in case of fragment R2. Shifting in peaks positions and a weak negative Cotton effect, observed especially for fragments R3 and R5, can be a consequence of the aggregation process occurring during the measurement. This was also observed with other techniques, as we discuss further.

The results of the secondary structure analysis, based on CD spectra, indicate that the rate of assuming the β -sheet conformation is higher in the fragments from *S. enterica* than in their counterparts from *E. coli*. While all *E. coli* fragments were still in the phase of the random coil structure, all potentially aggregating fragments of *S. enterica* showed β -sheet conformations on the day of dissolving. This process seemed most advanced for R1 and R4 fragments; however, the results for R5 and R3 fragments also indicated the onset of amyloid aggregation.

2.3.2. ATR-FTIR

ATR-FTIR spectra in the wavenumber range of 1725–1590 cm^{-1} were used for a more advanced secondary structure analysis of the CsgA fragments.

We compared the spectra of *E. coli* fragments obtained on the day of dissolving (Figure 3A) with those after one month of incubation at 37 °C (Figure 3B). Fragments R1, R3, and R5, on the day of dissolving, showed the main band located below 1630 cm^{-1} , which corresponds to cross- β amyloid architecture (Figure 3A). It indicated the presence of aggregates [37]. These repeating units were considered as highly amyloidogenic. Wang et al. showed that R1 and R5 fragments are critical for CsgA protein to form fibrils [9]. The analysis of the second derivative spectrum of R3 unveiled that R3 formed more rigid and ordered fibrils than R1 or R5. It was revealed by the location of its main negative peak at lower wavenumbers at c.a. 1621 cm^{-1} , as well as the band width. For fibril forming fragments R1, R3, and R5, the Amide I band had an additional local maximum located at approximately 1665 cm^{-1} , which is typically attributed to the parallel β -sheet structure that shows a high-frequency component between 1670–1660 cm^{-1} [42]. It can also be assigned to turn structures [43,44], as well as loops [45]. High absorbances in that loop–turn region are characteristic of parallel β -helix structure and observed, for example, in infrared spectra of HET-s [46] or PrP^{Sc} [47], which are known to adopt beta solenoid conformations.

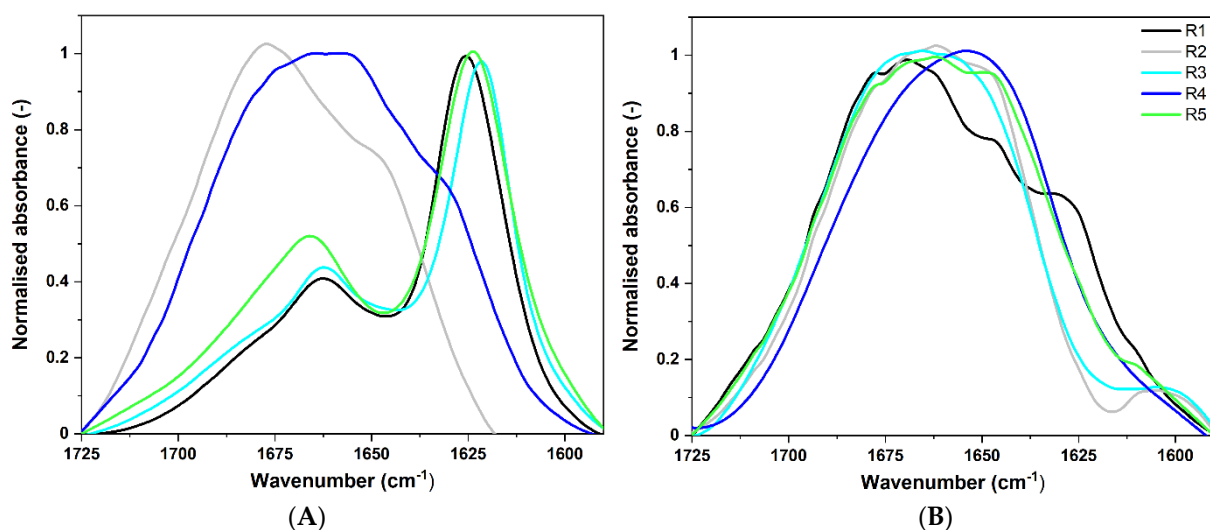


Figure 3. Normalized ATR-FTIR spectra of *E. coli* fragments in the wavenumber range of 1725–1590 cm^{-1} (Amide I), smoothed SG 35 (see Methods). (A) on the day of the dissolving (B) after one month of incubation at 37 °C. Peptide concentration was 500 μM .

After 30 days of incubation at 37 °C we observed that Amide I bands in ATR-FTIR spectra, registered for R1, R3, and R5 fragments of *E. coli*, broadened (Figure 3B). They lost spectral signatures typical of aggregates. Nevertheless, in the second derivative, the spectrum of R1 (Figure S6 in SI) was still clearly visible, leading to the conclusion that the R1

In turn, the fragment R2 had a more complex spectral characteristic, with a broad local maximum at 1678 cm^{-1} assigned to β -turns. The shoulder at 1648 cm^{-1} was typical of a random coil [48]. ATR-FTIR spectrum of R2 lacked the β -sheet component below 1640 cm^{-1} and, due to its low absorbance below 1630 cm^{-1} , we concluded that this fragment did not manifest aggregation properties. These observations are consistent with CD results (see Section 2.3.1) and with literature [49,50]. Fragments R2 and R4 from *E. coli*'s CsgA are considered incapable of self-assembly into ordered amyloid fibers in vitro; albeit, it is worth adding that Wang et al. [51] observed fibers with TEM experiments in which R2 or R4 were incubated at 2 mg/mL at room temperature for 5 days. As expected, peptide R4 under our experimental conditions did not manifest significant amyloidogenic properties. However, the secondary derivative in the range of $1725\text{--}1590\text{ cm}^{-1}$ (Figure S5 in SI) revealed the presence of β -sheet low frequency component at ca. 1628 cm^{-1} , and high frequency component in the range of $1710\text{--}1690\text{ cm}^{-1}$, typical of anti-parallel β -sheet [42]. The absorbance of these sub-bands was relatively low.

After 30 days of incubation at $37\text{ }^{\circ}\text{C}$ we observed that Amide I bands in ATR-FTIR spectra, registered for R1, R3, and R5 fragments of *E. coli*, broadened (Figure 3B). They lost spectral signatures typical of aggregates. Nevertheless, in the second derivative, the spectrum of R1 (Figure S6 in SI) was still clearly visible, leading to the conclusion that the R1 fragment remained partially aggregated. Moreover, all fragments, excluding R4, exhibited local minima at about 1693 and 1678 cm^{-1} , assigned to antiparallel β -sheet and β -turns, respectively. This spectral characteristic is typical of oligomers [36,42,52]. In turn, Amide I bands in the ATR-FTIR spectra of R2 and R4 were dominated by the sub-band at around 1644 cm^{-1} , assigned to random structures.

We concluded that the aggregates formed by R1, R3, and R5 fragments of CsgA from *E. coli* were not stable in time under studied conditions. This phenomenon can be caused by deamination of asparagine and glutamine residues present in the fragments. This non-enzymatic reaction leads to carboxylic acid derivatives, confirmed by higher absorption in the range of $1725\text{--}1710\text{ cm}^{-1}$. The process is known to occur during the incubation in vitro [53,54] and also during peptide synthesis [55]. Deamidation process is generally slow, but it can be strengthened by experimental conditions, such as increased temperature. In our case, two factors may have influenced the observed effect: incubation time (30 days) and temperature ($37\text{ }^{\circ}\text{C}$). After three months of the incubation process, disintegration of *E. coli* fragments appeared. This observation is very interesting with regard to the fact that CsgA protein is a functional amyloid from an organism frequently co-existing with humans, and as such, its fibrils should not be as stable as those from pathological amyloids.

Similarly, we compared the spectra of *S. enterica* fragments obtained on the day of dissolving (Figure 4A) with those measured after one month of incubation at $37\text{ }^{\circ}\text{C}$ (Figure 4B). Spectra of fragments R1, R3, R4, and R5, directly after dissolving, showed a high intensive absorbance at about 1622 cm^{-1} (Figure 4A). This indicates the presence of long and rigid amyloid fibrils [36,56]. For the R2 fragment, a broad band located at about 1645 cm^{-1} in the Amide I, which is characteristic of disordered proteins [57], could be observed. After 30 days of incubation at $37\text{ }^{\circ}\text{C}$, we did not notice any significant changes in all studied ATR-FTIR spectra in the range of $1725\text{--}1590\text{ cm}^{-1}$ (Figure 4B). Contrary to *E. coli*, all *S. enterica* fragments maintained the same structures as they assumed directly after dissolving. This result indicates that CsgA fragments of *S. enterica* are more stable than those from *E. coli*. However, after 3 months of incubation at $37\text{ }^{\circ}\text{C}$, we observed that all fragments of *S. enterica* also disintegrated (data not shown).

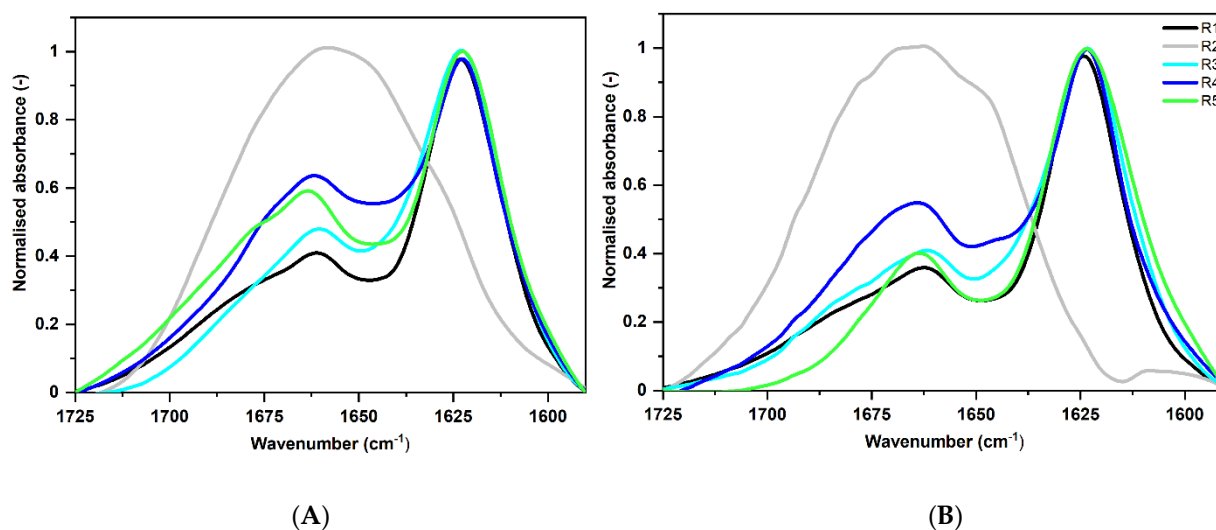


Figure 4. Normalized ATR-FTIR spectra of *S. enterica* fragments in the wavenumber range of 1725–1590 cm^{-1} (Amide I), smoothed with SG 35 (see Methods). (A) on the day of the dissolving (B) after month incubation at 37 °C. Peptide concentration was 500 μM .

The results from ATR-FTIR do not exactly match those obtained from CD, as we also observed here. ATR-FTIR experiments may speed up formation of amyloid fibrils due to interaction of peptides with the hydrophobic surface of the ATR accessory diamond. For example early occurrence of amyloids in fragments of *E. coli* *Las* as shown by our FTIR experiments could be related to the increased rate of aggregation (not observed with CD at this stage).

A summary of the Amide I spectral analysis is presented in Table 1. The corresponding fragments are compared, including effects of locations of the characteristic spectral components, shows propensity of peptide to aggregate, aggregation details on details on tertiary structures.

Table 1. Secondary structure assignments of the studied peptides with sequences from *E. coli* and *S. enterica* on the basis of Amide I ($\nu(\text{C=O})$ 80%, $\nu(\text{NH})$ 20%) band in ATR-FTIR spectra. The results are from experiments on the day of dissolving and after incubation for 30 days at 37 °C. Band positions (cm^{-1}) are presented, along with tentative assignments based on the most intense local minima of the second derivatives.

| | | After Dissolving | | 30 Days | |
|--------------------|----|------------------|------------------|------------------|------------------|
| | | After Dissolving | | 30 Days | |
| | | Assignment | Assignment | Assignment | Assignment |
| | | cm^{-1} | cm^{-1} | cm^{-1} | cm^{-1} |
| <i>E. coli</i> | | | | | |
| R1 | R1 | 1626 | aggregates | 1679 | turns |
| R2 | R2 | 1667 | turns | 1646 | 1679 |
| R3 | R3 | 1621 | aggregates | 1679 | 1646 |
| R4 | R4 | 1654 | random | 1642 | 1679 |
| R5 | R5 | 1624 | aggregates | 1646 | 1642 |
| <i>S. enterica</i> | | | | | |
| R1 | R1 | 1626 | aggregates | 1679 | turns |
| R2 | R2 | 1645 | random | 1646 | 1679 |
| R3 | R3 | 1622 | aggregates | 1623 | 1646 |
| R4 | R4 | 1622 | aggregates | 1623 | 1623 |
| R5 | R5 | 1622 | aggregates | 1624 | 1623 |

We carried out principal component analysis (PCA) based on normalized ATR-FTIR spectra after application of SG 35, in the range of 1725–1590 cm^{-1} (see Methods). PCA

We carried out principal component analysis (PCA) based on normalized ATR-FTIR spectra after application of SG 35, in the range of 1725–1590 cm^{-1} (see Methods). PCA analysis distinguished a class of aggregates in the set of studied peptides, based on the first three components (Figure 5 and Figure S1 in SI). Based on ATR-FTIR spectra of CsgA fragments, the loading plot of PC1 was obtained (Figure S2 in SI). It shows that the Amide I component at 1620 cm^{-1} strongly contributes to the separation of aggregates and non-aggregates. PC2 distinguished oligomers, due to high contributions of 1690, 1680, and 1630 cm^{-1} . These features are characteristic of anti-parallel structures [52]. The results of PCA analysis matched those by a human expert, as presented in Table 1.

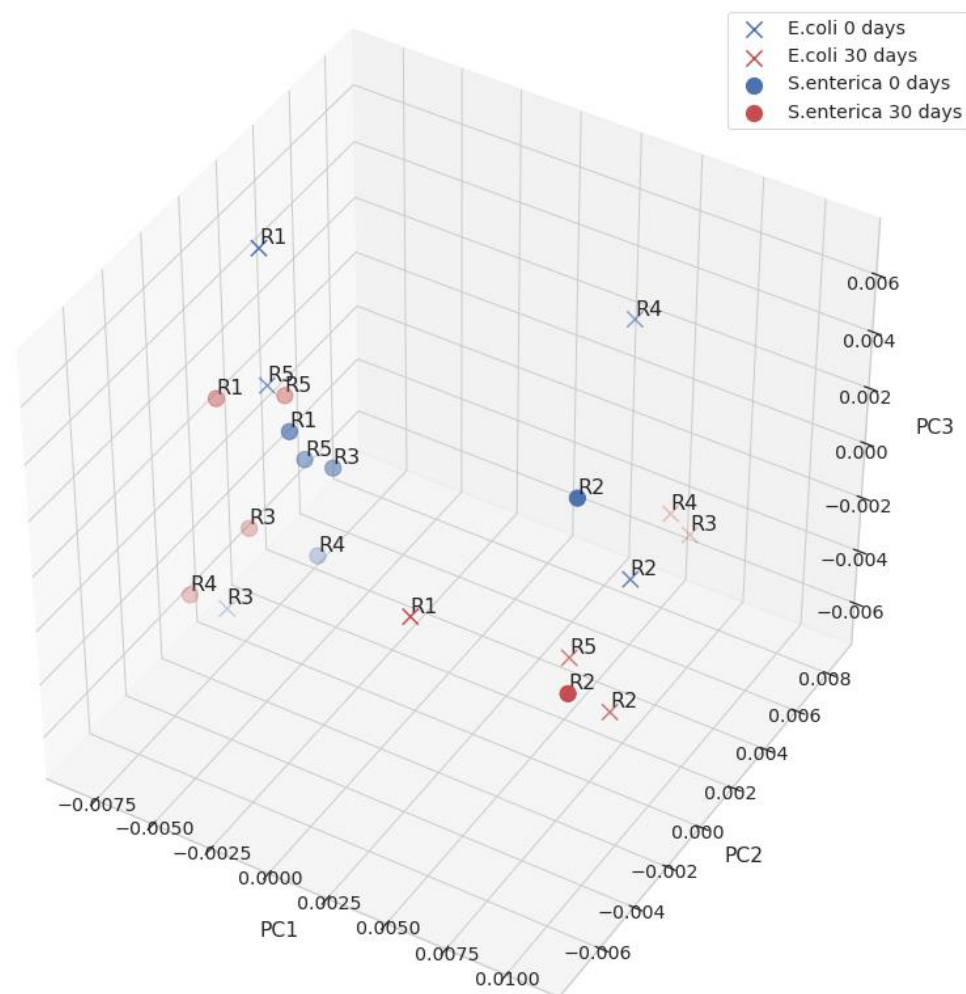


Figure 5. PCA plot for *E. coli* and *S. enterica*. Samples on the day of the dissolving and after 30 days of incubation at 37 °C. Points on the left side correspond to aggregates, on the right side to random structures.

2.3.3. FT-Raman

2.3.3. FT-Raman

For further structure analysis of the long-incubated CsgA fragments, FT-Raman spectroscopy was applied for R1–R5 fragments of *E. coli* and *S. enterica* after 30 days of incubation at 37 °C. This study can bring information complementary to ATR-FTIR regarding amyloid structures. FT-Raman technique is not frequently used for studying amyloids, although it can shed new light on structural analysis of aggregates. Results of FT-Raman spectra of CsgA fragments from *E. coli* and *S. enterica* are presented in Figure 6, including Amide I (1725–1575 cm^{-1}) (second derivative spectra are available in Figure S7 in SI), and Amide III (1375–1185 cm^{-1}) bands (Figure S3 in SI). The main drawback of studying peptides and proteins in aqueous solutions using FTIR spectroscopy is a strong water absorbance band at approximately 1635 cm^{-1} in Amide I band. Contrary to ATR-FTIR spectra, those from FT-Raman are usually analyzed in all these bands because there is

gates. The differences in its intensities can indicate various exposures of aromatic amino acids to the external environment, most probably caused by changes in tertiary structure of peptides during the aggregation process [65]. The analysis of Amide III (Figure 3A, Table S3 in SI) confirmed that all fragments possessed dominant β conformation but, in addition, revealed some differences in secondary structures between studied fragments. While fragments R1, R2, R4, and R5 had higher intensities near 1267 cm^{-1} , which corresponded to the β -turns, the location of an amide III band at 1250 cm^{-1} obtained for R3 was typical of random and loose β structures. In all second derivative spectra in the range of $1375\text{--}1195\text{ cm}^{-1}$, the minimum at $\sim 1230\text{ cm}^{-1}$ was present. It was most intensive in the spectrum of R5, and it could be assigned to the β -sheet structure [66]. Importantly, simultaneous analysis of two regions enables higher certainty of structure assignments.

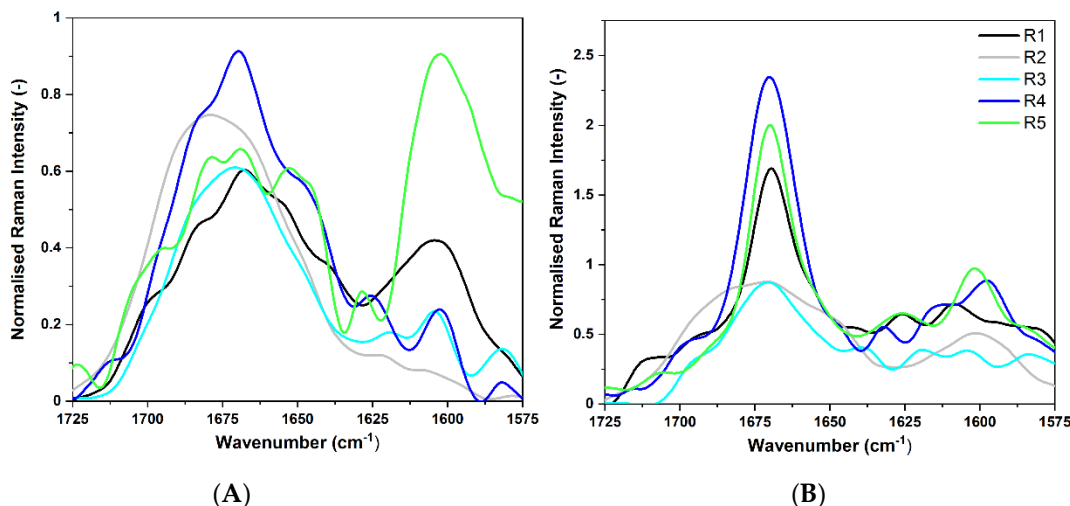


Figure 6. Normalized FT-Raman spectra of CsgA protein fragments, smoothed with SG 35 (see Methods), in the wavenumber range of $1725\text{--}1575\text{ cm}^{-1}$ (Amide I). (A) Spectra for *E. coli* fragments after 30 days of incubation at $37\text{ }^{\circ}\text{C}$, (B) Spectra for *S. enterica* fragments after 30 days of incubation at $37\text{ }^{\circ}\text{C}$. Peptide concentration was $500\text{ }\mu\text{M}$.

Raman results prove a partial deamidation of peptides corresponding to five imperfect repeating units of CsgA from *E. coli*. The FT-Raman spectra of *E. coli* subunits in the Amide I range are much more complex than the spectra of *S. enterica*, but in general, the Amide I band positions are characteristic of β structures [59]. Four fragments (R1, R3, R4, R5) of *E. coli* assumed a high percentage of β -sheet structures, which is indicated by the presence of a second derivative minimum at about 1670 cm^{-1} [60,61]. Bands in the range of $1695\text{--}1680\text{ cm}^{-1}$ are usually assigned to the β -turn structure [44]. However, disordered proteins also exhibit high contribution in that region [62]. Therefore, the R2 fragment of *E. coli* showed a broad Amide I band with the maximum at about 1691 cm^{-1} , with an associated (shoulder) band at 1647 cm^{-1} . The assignment of the sub-band at around 1645 cm^{-1} is debatable in literature, but in our opinion, it should be assigned to disordered structures. The appearance of additional Amide I mode at 1645 cm^{-1} is correlated with the strongly enhanced band of $1245\text{--}1255\text{ cm}^{-1}$ in the Amide III region (Figure S3 in SI), typically attributed to unordered structures.

Additionally, all spectra, except for R2, revealed higher intensities at about 1600 cm^{-1} , due to surface-enhanced Raman spectroscopy (SERS) effect, which occurred upon the adsorption of the peptide on metal surfaces [63]. This broad band can be mainly attributed to ring modes of phenylalanine and tyrosine [64], but it also overlaps with other spectral features in that region, i.e., Amide II. The most intensive band at $\sim 1600\text{ cm}^{-1}$ was observed for R1 and R5 fragments (Figure 6), which indicates the presence of the most rigid aggregates. The differences in its intensities can indicate various exposures of aromatic amino acids to the external environment, most probably caused by changes in tertiary structure of peptides during the aggregation process [65]. The analysis of Amide III (Figure 3A, Table S3 in SI) confirmed that all fragments possessed dominant β conformation but, in addition, revealed some differences in secondary structures between studied fragments. While fragments R1, R2, R4, and R5 had higher intensities near 1267 cm^{-1} , which corresponded to the β -turns, the location of an amide III band at 1250 cm^{-1} obtained for R3 was typical of random and loose β structures. In all second derivative spectra in the range of $1375\text{--}1195\text{ cm}^{-1}$, the minimum at $\sim 1230\text{ cm}^{-1}$ was present. It was most intensive in the spectrum of R5, and it could be assigned to the β -sheet structure [66].

In FT-Raman spectra of *S. enterica*, we observed that all peptides exhibited Amide I band maxima near 1670 cm^{-1} , which is typical of β -structures [67,68]. As mentioned above, β -turns give an additional contribution to the spectra in the range of $1715\text{--}1675\text{ cm}^{-1}$. For the R2 fragment of *S. enterica*, the Amide I band was very broad with the full width at half maximum (FWMH) = 60 cm^{-1} , indicating a complex structure (Table S4 in SI). The narrow (FWMH is 19 cm^{-1}) and intensive Amide I band, as of R1, indicated the presence of well-ordered β -strands [38] (Figure 6B). The R3 fragment also exhibited a complex spectrum; however, the dominant maximum at 1671 cm^{-1} marked the signature of β -sheet conformation. All Amide bands of R3 consisted of more sub-bands than bands from other peptides (Figure 6B, Figure S3B in SI). Additionally, the spectrum had an increased intensity at about 1600 cm^{-1} , which can be interpreted as a contribution from aggregates. The analysis of Amide III band confirms all above observations (see Figure S3, Table S3 in SI). The wavenumber range of $1375\text{--}1185\text{ cm}^{-1}$ was dominated by signatures typical of β structures. Fragments R4 and R5 exhibited intensive features at 1225 cm^{-1} , which arose from β -sheet conformations [61,69].

A summary of the spectral analysis based on FT-Raman experiments is presented in Table 2. Locations of the most characteristic spectral components show propensity of each peptide to certain secondary structures.

Table 2. Main band positions of Amide I in FT-Raman spectra of studied peptides in aqueous solution after 30 days of incubation at $37\text{ }^{\circ}\text{C}$. Band positions (cm^{-1}) along with tentative assignments based on the most intensive local minima of the second derivatives.

| | | 30 Days | |
|--------------------|----|------------------|----------------------|
| | | cm^{-1} | Assignment |
| <i>E. coli</i> | | | |
| | R1 | 1668 | β -sheet |
| | R2 | 1691 | turns |
| | R3 | 1667 | β -sheet |
| | R4 | 1669 | β -sheet |
| | R5 | 1668, 1680 | β -sheet/turns |
| <i>S. enterica</i> | | | |
| | R1 | 1670 | β -sheet |
| | R2 | 1670, 1697 | β -sheet/turns |
| | R3 | 1671 | β -sheet |
| | R4 | 1670 | β -sheet |
| | R5 | 1670 | β -sheet |

Summarizing the results, FT-Raman spectroscopy showed that R1, R3, and R4 fragments of *E. coli* after 30 days of incubation contained a high number of β -sheet conformations, while R2 fragments had a dominant β -turn conformation. The presence of β -turns was also detected in R1, R4, and R5. Only R1 and R5 fragments showed some symptoms of amyloid aggregates. However, the Amide I band in FT-Raman spectrum registered for the R5 fragment of *E. coli* was typical of a random coil. It may indicate that fragment R5 formed less structured aggregates in comparison to the structure of R1. In the case of *S. enterica*, R1, R3–R5 fragments formed amyloid aggregates. However, R3 formed amyloid aggregates with additional contribution of other complex secondary structures. R2 had a complex non-aggregated secondary structure. Spectral signatures for this fragment are typical of a disordered conformation. R1, R4, and R5 represented a structure with dominant β -sheets.

Based on FT-Raman results, it is evident that all *S. enterica* CsgA sequence repeats show higher stability than corresponding fragments of *E. coli*. The fibril forming units exhibited intensive and narrow Amide I bands located at $\sim 1670\text{ cm}^{-1}$, while in the case of *E. coli*, broad and complex FT-Raman signatures were observed in the range of $1725\text{--}1590\text{ cm}^{-1}$. FT-Raman findings are consistent with previously presented results from ATR-FTIR.

Based on FT-Raman results, it is evident that all *S. enterica* CsgA sequence repeats show higher stability than corresponding fragments of *E. coli*. The fibril forming units exhibited intensive and narrow Amide I bands located at $\sim 1670\text{ cm}^{-1}$, while in the case of *E. coli*, broad and complex FT-Raman signatures were observed in the range of $1725\text{--}1590\text{ cm}^{-1}$. FT-Raman findings are consistent with previously presented results from ATR-FTIR.

2.3.4. Transmission Electron Microscopy

To observe the morphology and size of aggregates from *S. enterica* fragments, which have not been studied and published so far, we used transmission electron microscopy (TEM) [70,71]. Presence of fibrils was observed in case of R1, R4, and R5 fragments. The observed fibrils were organized into rigid and high ordered structures that ranged from 100 nm to 1000 nm in diameter and from 500 nm to more than 1 μm in length. Interestingly, the morphology of fragment R3 differed from other fragments of *S. enterica*. The aggregates were composed of many connected oligomers/monomers. The observation matches the results from FT-Raman technique, which indicated a complex structure. This result shows that fragment R3 of *S. enterica* does not have as strong amyloid propensity as R1 and R5 similar to R3 of *E. coli* [49]. In turn, the micrographs of the R2 fragment did not reveal fibrils (Figure 7). These results are in agreement with ATR-FTIR and FT-Raman techniques, which also classified R2 as a non-aggregating fragment.

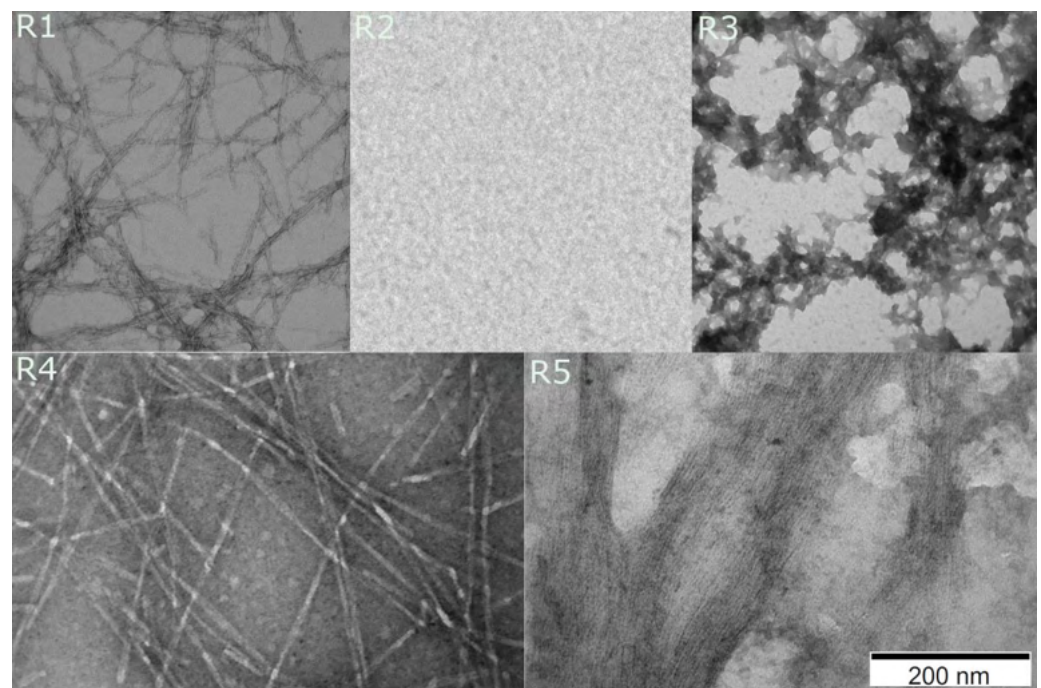


Figure 7. Electron micrographs of *S. enterica* fragments after one to seven days of incubation in 37°C . Images registered at the magnification of 200 μm . Peptide concentration was $0.5\ \mu\text{M}$. Images registered at the magnification of 200 nm. Peptide concentration was $0.5\ \mu\text{M}$.

2.4. Comparative Analysis of R4 Fragments from *S. enterica* and *E. coli*

Different prediction results from bioinformatics tools indicated that the R4 fragment may differ in the aggregation properties for sequences from *E. coli* and *S. enterica*. Our spectroscopy and microscopy results confirmed the computational prediction and demonstrated that R4 is the only fragment with different amyloid characteristics. Therefore, we conducted additional analyses regarding both sequences of R4.

2.4.1. ThT Assay

To track the fibrillation kinetics, we performed a thioflavin-T (ThT) fluorescence assay [72]. The results are presented in Figure 8. A significant increase in the fluorescence emission was observed for *S. enterica*, which confirmed fibril assembly [73]. The fluorescence of R4 of *S. enterica* was about nine times higher than that of *E. coli* fragment. The lag phase was not observed there, which indicated rapid aggregation. The fibrillation steps of R4 of *S. enterica* indicated immediate elongation phase and saturation phase.

emission was observed for *S. enterica*, which confirmed fibril assembly [73]. The fluorescence of R4 of *S. enterica* was about nine times higher than that of *E. coli* fragment. The aggregation phase was not observed there, which indicated rapid aggregation. The fibrillation steps of R4 of *S. enterica* indicated immediate elongation phase and saturation phase. The fibrillation steps of R4 of *S. enterica* indicated immediate elongation phase and saturation phase.

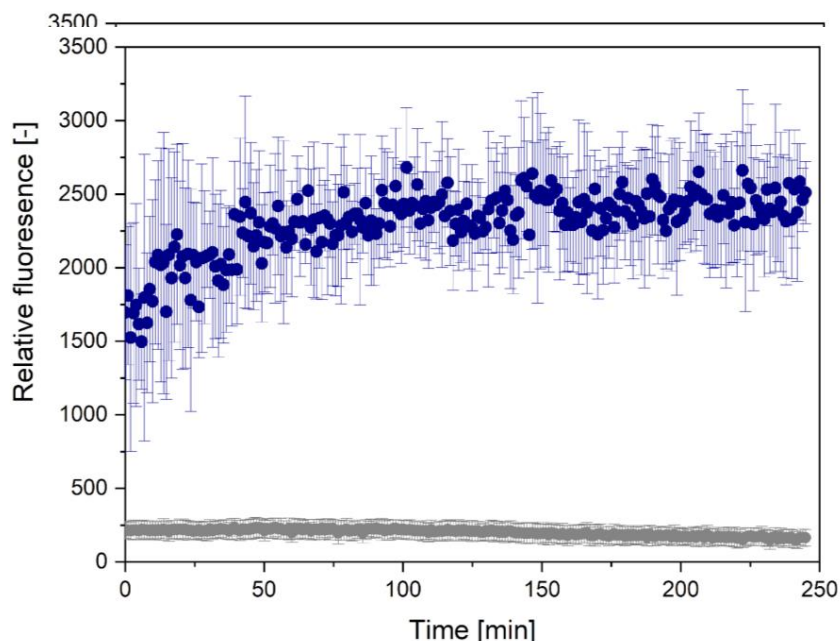


Figure 8. Time-dependent ThT fluorescence curves for R4 fragments. Here, grey dots represent *E. coli*, blue ones *S. enterica*. Peptide concentration was 500 μ M.

2.4.2. Comparative Transmission Electron Microscopy Micrographs

The micrographs of R4 fragment from *E. coli*, measured on the day of the dissolving, did not reveal fibrils (Figure 9A). This result contrasted with the micrographs of *S. enterica*, which showed fibrils (Figure 9B). However, we made a very interesting observation regarding R4 fragment of *E. coli*. The fibrils of R4 were also observed for *E. coli*, but only after seven days of incubation at 37 °C (Figure 9C). Our results show that the R4 fragment has an amyloid propensity in both bacterial species; however, the aggregation process of isolated *S. enterica* fragments is much faster than that in *E. coli*.

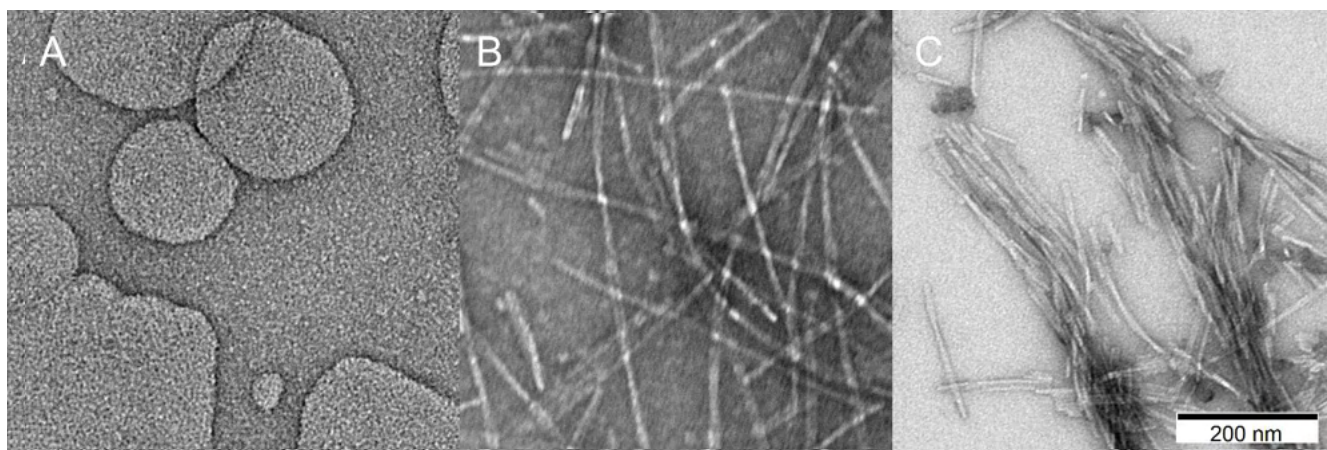


Figure 9. Comparison R4 fragments: (A) *E. coli* on the day of dissolving, (B) *S. enterica* on day of dissolving, (C) *E. coli* after 7 days of incubation in 37 °C. Images registered at the magnification of 200 nm. Peptide concentration was 0.5 μ M.

3. Discussion

Bacteria produce biofilms to create an environment protecting them from adverse conditions using amyloid forming fibrils, such as CsgA curli protein. Understanding of the self-assembly propensity of CsgA can be based on evolutionary variability of imperfect repeats. The structural and functional nature of the CsgA protein is defined by R1–R5

fragments, representing non-perfect repeats of amino acid sequences. In this work we compared aggregation propensities of CsgA fragments in vitro from *E. coli* with those from *S. enterica*, not reported so far. The results may help in better understanding of the principles of curli aggregation and potential effects on human health, especially in the case of salmonellosis.

Based on previous publications on *E. coli* [9,49], it was expected that R1 and R5 fragments are the aggregation seeds, driving the amyloid propensity of CsgA. Fragment R3 is also capable of aggregation, while R2 and R4 are non-aggregating. Tendency to form amyloid fibrils is defined by certain gate-keeper residues, specific to each fragment. Generally, the repeat fragments are defined by the common motif: S-[X]5-Q-[X]-[G{Ec}/X{Se}]-[X]-[G{Ec}/X{Se}]-N-[X]-A-[X]3-Q, where {Ec} and {Se} indicate differences in motifs characteristic of *E. coli* and *S. enterica*, respectively. The discovered gate-keepers are located at the mutating positions, denoted here with [X].

One of those gate-keeper residues, observed in R1 fragment of *E. coli*, is a negatively charged glutamic acid (Q) at its position 7 and aspartic acid (N) at position 12. These positions are not mutated in *S. enterica*. It should also be noted that negatively charged aspartic acid (N) at position 23 is replaced by positively charged lysine (K) in *S. enterica*.

The non-amyloid nature of R2 is secured by the presence of glycine (G) at positions 13 and 17, as well as negatively charged aspartic acid (D) at position 15. This pattern is identical in both bacterial species, which suggests that R2 keeps its non-amyloid character also in *S. enterica*.

Fragment R3, which is a weak amyloid in *E. coli*, is slowed down in its amyloid susceptibility by two gate-keeper residues: aspartic acid (D) at positions 4 and 17. One of these residues is mutated in *S. enterica*—aspartic acid at position 4 is replaced by glutamic acid (D4E). This substitution should lead to stronger amyloid propensity in R3 from *S. enterica* when compared to its counterpart from *E. coli*.

An even greater change concerning gate-keeper residues appears in the R4 fragment. Here, aggregation in *E. coli* is slowed down by glycine (G) at position 13 and aspartic acid (D) at position 17. Both positions are substituted in *S. enterica*, glycine by alanine (G13A) and aspartic acid by glutamic acid (D17N). This substitution may lead to increased amyloid propensity in R4 from *S. enterica*.

Strong amyloid propensity of R5 from *E. coli* was assigned to negatively charged glutamic acid (N) residues at positions 4 and 12 and histidine (H) at position 17. Fragment R5 in *S. enterica* has two substitutions at these positions; negatively charged aspartic acid is replaced by uncharged methionine (N4M) and positively charged histidine by negatively charged aspartic acid (H17N). The substitutions of gate-keeper residues could affect the amyloid nature of R5.

We studied self-aggregation and secondary structures of the repeat fragments of *S. enterica* and *E. coli* and comparatively analyzed their potential effects on these proteins in a bacterial biofilm. Different methods were applied, including bioinformatics prediction, ATR-FTIR and FT-Raman spectroscopy techniques, circular dichroism, and transmission electron spectroscopy.

Bioinformatics predictors were not unanimous in their results and, unfortunately, not very accurate when confronted with reported results from experimental studies. This could have been due to the functional role of CsgA. Functional amyloids are scarcely represented in reference datasets on which computational predictors of amyloids are based. Therefore, a statistical sequence profile of functional amyloids is most probably significantly different from that of a pathological amyloid. This conjecture comes from different structural details of the two classes of amyloids, their different temporal characteristics, and different stability and controllability by environmental conditions and interactions. However, currently available methods are not totally useless regarding functional amyloids; they are capable of guiding a more informed search. In our studies, bioinformatic method AmyloGram predicted a different amyloid propensity of R4 fragments from *E. coli* and *S. enterica*, indicating that non-amyloid R4 from *E. coli* changes its nature in *S. enterica*, where it may be

aggregating. The analysis of mutations in its gate-keeper residues supported this possibility. Although the R4 fragment of *E. coli* was previously reported as non-aggregating in similar conditions [9], there were no studies of R4 from *S. enterica*.

The clue from computational prediction was confirmed in our experiments—R4 from *S. enterica* turned out to be strongly aggregating. ThT kinetic studies showed very fast aggregation of R4 from *S. enterica*, in which we were not able to observe a lag phase. It contrasted with the results of the ThT study of R4 from *E. coli*, which did not show aggregation. The results were consistent with ATR-FTIR studies, showing more aggregates in R4 from *S. enterica* and the prevalence of random coils in *E. coli*. However, our TEM studies, taken after 7 days of incubation, showed that the initial lack of aggregates of R4 from *E. coli* did not reflect its true nature. R4 also forms amyloid fibrils, but the process is much slower than in *S. enterica*. Aggregation of R4 was also reported in [51], where its concentration was much higher than typically used in such studies, which indicated some amyloid propensity of the peptide. Nevertheless, after 30 days of incubation, the aggregates of R4 from *E. coli* disintegrated. The general structure was random coil, however, the presence of oligomers could not be excluded. Stability of R4 from *S. enterica* turned out differently—the aggregates were unchanged after 30 days, as indicated by ATR-FTIR and FT-Raman studies.

We note that, in aggregation prone peptides, charged residues were observed mostly in terminal parts. Additional arguments for the importance of the charge distribution can be found in the change of aggregation propensity of the R4 fragment, observed from our study. In its sequence derived from *E. coli*, two charged amino acids are located outside the terminal or linker region (6K and 17D). However, in its strongly aggregating counterpart from *S. enterica*, they are no longer present. The replacement of amino acids at positions 17 and 21 leads to the change in charge distribution. This, alongside the loss of charge at position 6, leads to the structure with all charged amino acids at the same side of the folded peptide, and outside the core of the predicted β -arch. Combined with the loss of gate-keeper residues, it could be the key reason for the different amyloid propensity.

The experimental techniques confirmed self-aggregation of R1, R3, and R5 fragments for both species, as previously reported for *E. coli*, and now also shown by ATR-spectra and TEM micrographs for *S. enterica*. We also observed a much weaker aggregation propensity of R3 than R1 and R5 fragments, as previously reported by other studies for *E. coli* [49]. The ratio of aggregation was significantly lower for fragments of *E. coli*, as obtained from CD spectra and ThT-measurements. However, disintegration of amyloid fibrils in *E. coli* proceeded faster, as shown by ATR-FTIR and FT-Raman techniques for the fragments after 30 days of incubation.

The multitude of techniques applied in our studies also revealed other more subtle details regarding the aggregation processes and secondary structures of the repeating fragments, indicating also the presence of more complex structures formed of some fragments, as well as their evolution over time. All spectroscopy techniques confirmed the presence of β -harpin structure of monomers in general confirmation of CsgA fragments. All fragments, except R2, exhibited signatures of turns and β -sheet structures in vibrational techniques. These results are in agreement with previously published computational models of CsgA [74,75] and ssNMR structure [76].

We also showed that FT-Raman can be used as a complementary technique to infrared spectroscopy in amyloid studies [77]. It provides information about secondary structure and, additionally, about tertiary structure—revealing exposure of aromatic amino acids to the external environment. The undoubted advantage of Raman spectroscopy is the fact that the analysis of the well-resolved Amide III band provides complementary structural information to the Amide I. Using the FT-Raman technique, we could observe that R3 fragment of *S. enterica* had more complex structures than other fragments of CsgA. More sub-bands were present in the second derivative of R3 spectra than in those of other fragments. TEM images confirmed that the fibrils of R3 were morphologically distinct. The morphological differences can be caused by location of charged amino acids in the

sequence. The R3 fragment contains additional positively charged amino acids in the linker region and negatively charged aspartic acid residue at position 17.

Our studies showed that amyloid structures of CsgA repeats are more rapidly formed and more stable in *S. enterica* than those in *E. coli*, which has not been demonstrated so far. This result seems to be in accordance with in vitro aggregation of different curli variants, where the self-assembly of CsgA from *E. coli* is slower than that from *S. typhimurium* [11]. This phenomenon might be related to the general lifestyle of the Salmonella genus, where biofilm formation seems to be an important long-term colonization strategy [78]. Quicker self-assembly of CsgA could provide an advantage towards the prolonged infection.

Although there are no relevant reports concerning human amyloid diseases associated with salmonellosis, some clues could be derived from animal studies. As reported in [79], bacterial infection with *Salmonella Typhimurium* of the brains of transgenic 5XFAD mice resulted in rapid seeding and accelerated β -amyloid deposition, which closely colocalized with the invading bacteria. This finding could support a hypothesis that β -amyloid may play an immuno-protective role against bacterial infections and drive amyloidosis as a side effect. However, another mechanism may also be in play—a cross-talk between amyloid curli in bacterial biofilm and β -amyloid peptides where interactions of human proteins with bacterial curli accelerate formation of pathological aggregates. Therefore, understanding of amyloid propensity of Salmonella curli could be instrumental in studying aspects of human amyloid diseases.

Another important novelty in our studies is simultaneous use of a combination of several different experimental techniques. This approach enabled comparing different aspects revealed by each of the methods. In particular, FT-Raman spectroscopy was applied, which is very infrequently used in amyloid research. We also studied temporal changes in amyloid characteristics, regarding the curli from both species, not reported so far.

Further studies are required to shed more light on the surprising efficiency of self-assembly of CsgA produced by *S. enterica*, especially, effects of the sequence variability on the whole protein characteristics in vivo. The results would contribute significantly to better understanding of the curli aggregation.

4. Materials and Methods

4.1. Sample Preparation

CsgA *S. enterica* and *E. coli* fragments sequences were provided by CASLO (CASLO ApS, Denmark) (Table S4 in SI). Additionally, fragments: R2, R5 and partially R3 of *E. coli* of strain K12, were synthesized “in-house”. The synthesis was carried out with an automated solid-phase peptide synthesizer (Liberty Blue, CEM) using rink amide AM resin (loading: 0.59 mmol/g) (Table S5, Figure S6 in SI). Fmoc deprotection was achieved using 20% piperidine in DMF for 1 min at 90 °C. A double-coupling procedure was performed with 0.5 M solution of DIC and 0.25 M solution of OXYMA (1:1) in DMF for 4 min at 90 °C. Cleavage of the peptides from the resin was accomplished with the mixture of TFA/TIS/H₂O (95:2.5:2.5) after 3 h of shaking. The crude peptide was precipitated with ice-cold Et₂O and centrifuged (9000 rpm, 15 min, 4 °C). Peptides were purified using preparative HPLC (Knauer Prep) with a C18 column (Thermo Scientific, Hypersil Gold 12 μ , 250 mm \times 20 mm) with water/ACN (0.05 TFA) eluent system. The purity of synthesized peptides was in the range between 95% and 99.6%. A sample of each peptide was dissolved in 490 μ L of 0.01 M NaOH and vortexed for one minute. Then, 450 μ L of phosphate-buffered saline (PBS) pH 7.2 was added, followed by 60 μ L of Milli-Q[®] (Merck & Co. Inc., USA) water, pH 6.9. The final concentration of the aliquot was about 500 μ M, pH 7.4. To obtain monomers, each sample was filtered through a 0.2 μ m PVDF syringe filter (Table S2 in SI, prepared based on the MIRRAGGE protocol [80]).

Initial monomerization of aggregates is a necessary step; however, it may affect the results with regard to their full validity when they are extrapolated to actual behavior of the protein. The lack of initial protofibrils, which constitute transient pre-fibrillar intermediates,

may affect the pathway of further aggregation process. The filtration method lowers the effective concentration of the peptides.

4.2. Bioinformatic Analysis

The aggregation propensity of studied peptides was assessed using nine bioinformatics methods: AmyloGram [24], PATH [25], Pasta2.0 [28], Waltz [29], AmylPred2 [30], FoldAmyloid [31], MetAmyl [32], Tango [33], and ArchCandy [19]. Each predictor was used with its default parameters. Pairwise alignments of corresponding regions in *E. coli* and *S. enterica* were visualized using Jalview software [81].

4.3. Circular Dichroism (CD)

CD spectra were recorded on JASCO J-815 at 20 °C between 250 and 190 nm in PBS buffer pH = 7.2 with the following parameters: 0.2 nm resolution, 1.0 nm bandwidth, 20 mdeg sensitivity, 0.25 s response, 50 nm/min scanning speed, and 0.02 cm cuvette path length. The sample concentration was 500 µM. The CD spectra of the solvent alone was recorded and subtracted from the raw data. The CD intensity is given as mean residue molar ellipticity (θ [deg × cm² × dmol⁻¹]). Spectra were smoothed and plotted using Origin 2020b software.

4.4. Attenuated Total Reflectance—Fourier-Transform Infrared (ATR-FTIR)

FTIR studies were performed using Nicolet 6700 FTIR spectrometer (Thermo Scientific, USA) with ATR accessory and heated diamond top-plate (PIKE Technologies), continuously purged with dry air. Each sample of 10 µL of peptide aqueous solution was dropped directly on the diamond surface and allowed to dry out. All ATR-FTIR spectra were obtained in the range of 3600–400 cm⁻¹. For each spectrum, 512 interferograms was co-added with 4 cm⁻¹ resolution at constant temperature 22 °C [71.6 F]. Directly before sampling, the background spectrum of diamond/air was recorded as a reference (512 scans, 4 cm⁻¹). We used 500 µM concentration, which was essential to obtain a good signal-to-noise ratio. The raw data are shown in Tables S7 and S8 in SI.

4.5. FT-Raman

Raman spectra were carried out using a Nicolet NXR 9650 FT-Raman spectrometer with MicroStage extension equipped with Nd:YVO₄ laser (1064 nm, 500 mW) as an excitation source and InGaAs detector. A drop of 10 µL of each sample was deposited on the gold surface and dried under laser irradiation. All FT-Raman spectra were acquired in the range of 3700–0 cm⁻¹ with 4 cm⁻¹ resolution by averaging 1024 scans. The raw data are shown in Table S9 in SI.

4.6. Spectral Analysis

The spectra were analyzed using OriginPro 2020b (OriginLab Corporation, USA). The analysis included spectra baseline correction, smoothing using the Savitzky–Golay filter (polynomial order 2, widow size 35, SG 35) [82], normalization of spectra relative to Amide I band (ATR-FTIR), or deformation vibrations of CH₂ group, at 1450 cm⁻¹ (FT-Raman).

PCA was performed on the second derivative of the Amide I region of the spectra (1725–1590 cm⁻¹) using Scikit-learn Python package [83]. Matplotlib [84] Python package was used for visualization.

4.7. Thioflavin T (ThT) Fluorescence Assay

The fluorescence of each well was read by a microplate reader CLARIOstar, as well as BMG LABTECH at 25 °C with 30 s shaking every 58.8 s during 244.85 min measurements. The samples containing 10 µL of 500 µM peptide and 90 µL of 500 µM ThT solution were mixed in a 96-well plate. The excitation wavelength was set at 440 nm and emission at 480 nm. Each group of experiments contained six parallel samples, and the data were averaged after measurements.

4.8. Transmission Electron Microscopy (TEM)

Imaging was performed using a transmission electron microscope Hitachi H-800 (Hitachi HighTech, Japan) on accelerating voltage of 150 kV. Negative stained samples were prepared by applying a 4 μ L drop of solution containing 0.5 μ M peptide in water on glow discharged carbon on copper grid (Agar S160, Agar Scientific Ltd, United Kingdom). After 1 min of adhesion, an excess of the material was blotted, and 2% uranyl acetate was applied for 1 min before blotting. The samples were allowed to dry under normal conditions for at least 1 h.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/ijms22105127/s1>, **SI:** Table S1: Amyloidogenicity prediction results. Where 0 denotes non-amyloid, 1 stands for amyloid., Table S2: MIRRAGGE, Figure S1: Scree plot of PCA from ATR-FTIR spectra., Figure S2: Loading plot of PCA analysis as is resulted from the ATR-FTIR data., Figure S3: Normalized FT-Raman spectra of CsgA protein fragments with the second derivatives spectra smoothed 2 times with SG 35 in the wavenumber range of 1375–1185 cm^{-1} (Amide III). (A) Spectra for *E. coli* fragments after 30 days of incubation at 37 $^{\circ}\text{C}$., (B) Spectra for *S. enterica* fragments after 30 days of incubation at 37 $^{\circ}\text{C}$., Table S3: Main band positions of Amide III in FT-Raman spectra of studied peptides in aqueous solution after 30 days of incubation at 37 $^{\circ}\text{C}$. Band positions (cm^{-1}) along with tentative assignments based on the minima of the second derivatives. Bold values indicate the most intensive local minima., Table S4: Full width at half maximum (FWHM) of Amide I in the FT-Raman., Table S5: Peptides analytical data purchased from CASLO., Table S6: Peptides analytical data synthesized “in house”., Figure S4: Analytical HPLC chromatograms of “in house” studied peptide., Table S7: Raw ATR-FTIR spectra of *E. coli* fragments in the range of 3600–900 cm^{-1} ., Table S8: Raw ATR-FTIR spectra of *S. enterica* fragments in the range of 3600–900 cm^{-1} ., Table S9: Raw FT-Raman spectra of *E. coli* fragments in the range of 3600–600 cm^{-1} ., Figure S5: ATR-FTIR second derivatives spectra of *E. coli* fragments in the wavenumber range of 1725–1590 cm^{-1} ., smoothed twice with SG 35 (see Methods). (A) on the day of the dissolving (B) after month incubation at 37 $^{\circ}\text{C}$. Peptide concentration was 500 μ M, Figure S6: ATR-FTIR second derivatives spectra of *S. enterica* fragments in the wavenumber range of 1725–1590 cm^{-1} ., smoothed twice with SG 35 (see Methods). (A) on the day of the dissolving (B) after month incubation at 37 $^{\circ}\text{C}$. Peptide concentration was 500 μ M, Figure S7: FT-Raman second derivatives spectra, smoothed twice with SG 35 (see Methods), in the wavenumber range of 1725–1575 cm^{-1} . (A) Spectra for *E. coli* fragments after 30 days of incubation at 37 $^{\circ}\text{C}$., (B) Spectra for *S. enterica* fragments after 30 days of incubation at 37 $^{\circ}\text{C}$. Peptide concentration was 500 μ M.

Author Contributions: Conceptualization, M.K. and M.B.; methodology, N.S., M.G.-G., J.W.W., M.S., A.M.Ż., M.K.; experiments investigation, N.S., M.G.-G., M.S., J.W.W., A.M.Ż.; formal analysis, N.S., M.G.-G., M.S., M.K.; writing—original draft preparation, N.S., M.G.-G., J.W.W., M.S., A.M.Ż., M.B., M.K.; writing—review and editing, N.S., M.G.-G., J.W.W., M.S., A.M.Ż., M.B., M.K.; visualization, N.S. and J.W.W.; supervision, M.K.; project administration, M.K.; funding acquisition, M.B. and M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the National Science Centre, Poland, Grant 2019/35/B/NZ2/03997(MK, MB, MGG, JW) and Grant No. 2017/26/D/ST5/00341 (MS), National Centre for Research and Development, Poland under POWR.03.02.00-00-I003/16 (NS), and Wrocław Center of Biotechnology program “The Leading National Research Center (KNOW) for years 2014–2018” (MB).

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shanmugam, N.; Baker, M.O.D.G.; Ball, S.R.; Steain, M.; Pham, C.L.L.; Sunde, M. Microbial functional amyloids serve diverse purposes for structure, adhesion and defence. *Biophys. Rev.* **2019**, *11*, 287–302. [[CrossRef](#)]
2. Otzen, D.; Riek, R. Functional amyloids. *Cold Spring Harb. Perspect. Biol.* **2019**, *11*, a033860. [[CrossRef](#)] [[PubMed](#)]
3. Flemming, H.C.; Wingender, J. The biofilm matrix. *Nat. Rev. Microbiol.* **2010**, *8*, 623–633. [[CrossRef](#)]
4. Flemming, H.C.; Wuertz, S. Bacteria and archaea on Earth and their abundance in biofilms. *Nat. Rev. Microbiol.* **2019**, *17*, 247–260. [[CrossRef](#)] [[PubMed](#)]

5. Schiebel, J.; Böhm, A.; Nitschke, J.; Burdukiewicz, M.; Weinreich, J.; Ali, A.; Roggenbuck, D.; Rödiger, S.; Schierack, P. Genotypic and phenotypic characteristics associated with biofilm formation by human clinical *Escherichia coli* isolates of different pathotypes. *Appl. Environ. Microbiol.* **2017**, *83*, 1660–1677. [[CrossRef](#)] [[PubMed](#)]
6. Chapman, M.R.; Robinson, L.S.; Pinkner, J.S.; Roth, R.; Heuser, J.; Hammar, M.; Normark, S.; Hultgren, S.J. Role of *Escherichia coli* curli operons in directing amyloid fiber formation. *Science* **2002**, *295*, 851–855. [[CrossRef](#)]
7. Robinson, L.S.; Ashman, E.M.; Hultgren, S.J.; Chapman, M.R. Secretion of curli fibre subunits is mediated by the outer membrane-localized CsgG protein. *Mol. Microbiol.* **2006**, *59*, 870–881. [[CrossRef](#)] [[PubMed](#)]
8. Hammer, N.D.; Schmidt, J.C.; Chapman, M.R. The curli nucleator protein, CsgB, contains an amyloidogenic domain that directs CsgA polymerization. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 12494–12499. [[CrossRef](#)] [[PubMed](#)]
9. Wang, X.; Chapman, M.R. Sequence determinants of bacterial amyloid formation. *J. Mol. Biol.* **2008**, *380*, 570–580. [[CrossRef](#)] [[PubMed](#)]
10. Sewell, L.; Stylianou, F.; Xu, Y.; Taylor, J.; Sefer, L.; Matthews, S. NMR insights into the pre-amyloid ensemble and secretion targeting of the curli subunit CsgA. *Sci. Rep.* **2020**, *10*, 1–10. [[CrossRef](#)] [[PubMed](#)]
11. Zhou, Y.; Smith, D.; Leong, B.J.; Brännström, K.; Almqvist, F.; Chapman, M.R. Promiscuous cross-seeding between bacterial amyloids promotes interspecies biofilms. *J. Biol. Chem.* **2012**, *287*, 35092–35103. [[CrossRef](#)] [[PubMed](#)]
12. Dueholm, M.S.; Albertsen, M.; Otzen, D.; Nielsen, P.H. Curli functional amyloid systems are phylogenetically widespread and display large diversity in operon and protein structure. *PLoS ONE* **2012**, *7*, e51274. [[CrossRef](#)] [[PubMed](#)]
13. Sampson, T.R.; Challis, C.; Jain, N.; Moiseyenko, A.; Ladinsky, M.S.; Shastri, G.G.; Thron, T.; Needham, B.D.; Horvath, I.; Debelius, J.W.; et al. A gut bacterial amyloid promotes α -synuclein aggregation and motor impairment in mice. *eLife* **2020**, *9*, e53111. [[CrossRef](#)] [[PubMed](#)]
14. Miller, A.L.; Bessho, S.; Grando, K.; Tükel, Ç. Microbiome or infections: Amyloid-Containing biofilms as a trigger for complex human diseases. *Front. Immunol.* **2021**, *12*, 638867. [[CrossRef](#)] [[PubMed](#)]
15. Friedland, R.P.; Chapman, M.R. The role of microbial amyloid in neurodegeneration. *PLoS Pathog.* **2017**, *13*, e1006654. [[CrossRef](#)] [[PubMed](#)]
16. Harrell, J.E.; Hahn, M.M.; D’Souza, S.J.; Vasicek, E.M.; Sandala, J.L.; Gunn, J.S.; McLachlan, J.B. Salmonella biofilm formation, chronic infection, and immunity within the intestine and hepatobiliary tract. *Front. Cell. Infect. Microbiol.* **2021**, *10*, 624622. [[CrossRef](#)]
17. Perov, S.; Lidor, O.; Salinas, N.; Golan, N.; Tayeb-Fligelman, E.; Deshmukh, M.; Willbold, D.; Landau, M. Structural insights into curli CsgA cross- β fibril architecture inspire repurposing of anti-amyloid compounds as anti-biofilm agents. *PLoS Pathog.* **2019**, *15*, e1007978. [[CrossRef](#)] [[PubMed](#)]
18. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [[CrossRef](#)]
19. Ahmed, A.B.; Znassi, N.; Château, M.-T.; Kajava, A.V. A structure-based approach to predict predisposition to amyloidosis. *Alzheimer’s Dement.* **2015**, *11*, 681–690. [[CrossRef](#)]
20. Gazit, E. A possible role for π -stacking in the self-assembly of amyloid fibrils. *FASEB J.* **2002**, *16*, 77–83. [[CrossRef](#)] [[PubMed](#)]
21. Guo, M.; Gorman, P.M.; Rico, M.; Chakrabartty, A.; Laurents, D.V. Charge substitution shows that repulsive electrostatic interactions impede the oligomerization of Alzheimer amyloid peptides. *FEBS Lett.* **2005**, *579*, 3574–3578. [[CrossRef](#)]
22. Santos, J.; Iglesias, V.; Santos-Suárez, J.; Mangiagalli, M.; Brocca, S.; Pallarès, I.; Ventura, S. pH-dependent aggregation in intrinsically disordered proteins is determined by charge and lipophilicity. *Cells* **2020**, *9*, 145. [[CrossRef](#)]
23. Dueholm, M.S.; Nielsen, S.B.; Hein, K.L.; Nissen, P.; Chapman, M.; Christiansen, G.; Nielsen, P.H.; Otzen, D.E. Fibrillation of the major curli subunit CsgA under a wide range of conditions implies a robust design of aggregation. *Biochemistry* **2011**, *50*, 8281–8290. [[CrossRef](#)] [[PubMed](#)]
24. Burdukiewicz, M.; Sobczyk, P.; Rödiger, S.; Duda-Madej, A.; MacKiewicz, P.; Kotulska, M. Amyloidogenic motifs revealed by n-gram analysis. *Sci. Rep.* **2017**, *7*. [[CrossRef](#)] [[PubMed](#)]
25. Wojciechowski, J.W.; Kotulska, M. PATH—Prediction of amyloidogenicity by threading and machine learning. *Sci. Rep.* **2020**, *10*, 1–9. [[CrossRef](#)] [[PubMed](#)]
26. Wozniak, P.P.; Kotulska, M. AmyLoad: Website dedicated to amyloidogenic protein fragments. *Bioinformatics* **2015**, *31*, 3395–3397. [[CrossRef](#)] [[PubMed](#)]
27. Louros, N.; Konstantoulea, K.; De Vleeschouwer, M.; Ramakers, M.; Schymkowitz, J.; Rousseau, F. WALTZ-DB 2.0: An updated database containing structural information of experimentally determined amyloid-forming peptides. *Nucleic Acids Res.* **2020**, *48*, D389–D393. [[CrossRef](#)] [[PubMed](#)]
28. Walsh, I.; Seno, F.; Tosatto, S.C.E.; Trovato, A. PASTA 2.0: An improved server for protein aggregation prediction. *Nucleic Acids Res.* **2014**, *42*, 301–307. [[CrossRef](#)]
29. Maurer-Stroh, S.; Debulpaep, M.; Kueemmerer, N.; De La Paz, M.L.; Martins, I.C.; Reumers, J.; Morris, K.L.; Copland, A.; Serpell, L.; Serrano, L.; et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods* **2010**, *7*, 237–242. [[CrossRef](#)] [[PubMed](#)]
30. Tsolis, A.C.; Papandreou, N.C.; Iconomidou, V.A.; Hamodrakas, S.J. A consensus method for the prediction of “Aggregation-prone” peptides in globular proteins. *PLoS ONE* **2013**, *8*, e54175. [[CrossRef](#)]

31. Garbuzynskiy, S.O.; Lobanov, M.Y.; Galzitskaya, O.V. FoldAmyloid: A method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* **2010**, *26*, 326–332. [[CrossRef](#)] [[PubMed](#)]
32. Emily, M.; Talvas, A.; Delamarche, C. MetAmyl: A METa-predictor for AMYloid proteins. *PLoS ONE* **2013**, *8*, e79722. [[CrossRef](#)] [[PubMed](#)]
33. Fernandez-Escamilla, A.M.; Rousseau, F.; Schymkowitz, J.; Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **2004**, *22*, 1302–1306. [[CrossRef](#)] [[PubMed](#)]
34. Jackson, M.P.; Hewitt, E.W. Why are functional amyloids non-toxic in humans? *Biomolecules* **2017**, *7*, 71. [[CrossRef](#)] [[PubMed](#)]
35. Roberts, R.G. Good amyloid, bad amyloid—What’s the difference? *PLoS Biol.* **2016**, *14*, e1002362. [[CrossRef](#)] [[PubMed](#)]
36. Sarroukh, R.; Goormaghtigh, E.; Ruyschaert, J.M.; Raussens, V. ATR-FTIR: A “rejuvenated” tool to investigate amyloid proteins. *Biochim. Biophys. Acta Biomembr.* **2013**, *1828*, 2328–2338. [[CrossRef](#)] [[PubMed](#)]
37. Shivu, B.; Seshadri, S.; Li, J.; Oberg, K.A.; Uversky, V.N.; Fink, A.L. Distinct β -sheet structure in protein aggregates determined by ATR-FTIR spectroscopy. *Biochemistry* **2013**, *52*, 5176–5183. [[CrossRef](#)] [[PubMed](#)]
38. Ettah, I.; Ashton, L. Engaging with raman spectroscopy to investigate antibody aggregation. *Antibodies* **2018**, *7*, 24. [[CrossRef](#)]
39. Ridgley, D.M.; Claunch, E.C.; Barone, J.R. Characterization of large amyloid fibers and tapes with Fourier transform infrared (FT-IR) and raman spectroscopy. *Appl. Spectrosc.* **2013**, *67*, 1417–1426. [[CrossRef](#)]
40. Ojha, B.; Fukui, N.; Hongo, K.; Mizobata, T.; Kawata, Y. Suppression of amyloid fibrils using the GroEL apical domain. *Sci. Rep.* **2016**, *6*, 1–13. [[CrossRef](#)]
41. Shu, Q.; Crick, S.L.; Pinkner, J.S.; Ford, B.; Hultgren, S.J.; Frieden, C. The *E. coli* CsgB nucleator of curli assembles to β -sheet oligomers that alter the CsgA fibrillization mechanism. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 6502–6507. [[CrossRef](#)] [[PubMed](#)]
42. Zou, Y.; Li, Y.; Hao, W.; Hu, X.; Ma, G. Parallel β -sheet fibril and antiparallel β -sheet oligomer: New insights into amyloid formation of hen egg white lysozyme under heat and acidic condition from FTIR spectroscopy. *J. Phys. Chem. B* **2013**, *117*, 4003–4013. [[CrossRef](#)] [[PubMed](#)]
43. Seo, J.; Hoffmann, W.; Warnke, S.; Huang, X.; Gewinner, S.; Schöllkopf, W.; Bowers, M.T.; Helden, G.V.; Pagel, K. An infrared spectroscopy approach to follow β -sheet formation in peptide amyloid assemblies. *Nat. Chem.* **2017**, *9*, 39–44. [[CrossRef](#)] [[PubMed](#)]
44. Sadat, A.; Joye, I.J. Peak fitting applied to Fourier transform infrared and Raman spectroscopic analysis of proteins. *Appl. Sci.* **2020**, *10*, 5918. [[CrossRef](#)]
45. Khurana, R.; Fink, A.L. Do parallel β -helix proteins have a unique Fourier transform infrared spectrum? *Biophys. J.* **2000**, *78*, 994–1000. [[CrossRef](#)]
46. Berthelot, K.; Ta, H.P.; Géan, J.; Lecomte, S.; Cullin, C. In vivo and in vitro analyses of toxic mutants of HET-s: FTIR antiparallel signature correlates with amyloid toxicity. *J. Mol. Biol.* **2011**, *412*, 137–152. [[CrossRef](#)] [[PubMed](#)]
47. Requena, J.R.; Wille, H. The structure of the infectious prion protein Experimental data and molecular models. *Prion* **2014**, *8*, 60–66. [[CrossRef](#)]
48. Kong, J.; Yu, S. Fourier transform infrared spectroscopic analysis of protein secondary structures. *Acta Biochim. Biophys. Sin.* **2007**, *39*, 549–559. [[CrossRef](#)]
49. Evans, M.L.; Chapman, M.R. Curli biogenesis: Order out of disorder. *Biochim. Biophys. Acta Mol. Cell Res.* **2014**, *1843*, 1551–1558. [[CrossRef](#)] [[PubMed](#)]
50. Van Gerven, N.; Klein, R.D.; Hultgren, S.J.; Remaut, H. Bacterial amyloid formation: Structural insights into curli biogenesis. *Trends Microbiol.* **2015**, *23*, 693–706. [[CrossRef](#)]
51. Wang, X.; Smith, D.R.; Jones, J.W.; Chapman, M.R. In vitro polymerization of a functional *Escherichia coli* amyloid protein. *J. Biol. Chem.* **2007**, *282*, 3713–3719. [[CrossRef](#)] [[PubMed](#)]
52. Cerf, E.; Sarroukh, R.; Tamamizu-Kato, S.; Breydo, L.; Derclaye, S.; Dufrière, Y.; Narayanaswami, V.; Goormaghtigh, E.; Ruyschaert, J.-M.; Raussens, V. Anti-parallel β -sheet—a signature structure of the oligomeric amyloid-beta peptide. *Biochem. J.* **2009**, *421*, 415–423. [[CrossRef](#)] [[PubMed](#)]
53. Svozil, J.; Baerenfaller, K. A cautionary tale on the inclusion of variable posttranslational modifications in database-dependent searches of mass spectrometry data. In *Methods in Enzymology*; Academic Press Inc.: Cambridge, MA, USA, 2017; pp. 433–452.
54. Wang, H.; Shu, Q.; Frieden, C.; Gross, M.L. Deamidation slows curli amyloid-protein aggregation. *Biochemistry* **2017**, *56*, 2865–2872. [[CrossRef](#)] [[PubMed](#)]
55. Robinson, A.B.; Mckerrow, J.H.; Cary, P. Controlled deamidation of peptides and proteins: An experimental hazard and a possible biological timer. *Proc. Natl. Acad. Sci. USA* **1970**, *66*, 753–757. [[CrossRef](#)] [[PubMed](#)]
56. Ruyschaert, J.M.; Raussens, V. ATR-FTIR analysis of amyloid proteins. In *Methods in Molecular Biology*; Humana Press Inc.: Totowa, NJ, USA, 2018; pp. 69–81.
57. Milošević, J.; Prodanović, R.; Polović, N. On the protein fibrillation pathway: Oligomer intermediates detection using ATR-FTIR spectroscopy. *Molecules* **2021**, *26*, 970. [[CrossRef](#)] [[PubMed](#)]
58. Cai, S.; Singh, B.R. A Distinct utility of the amide III infrared band for secondary structure estimation of aqueous protein solutions using partial least squares methods. *Biochemistry* **2004**, *43*, 2541–2549. [[CrossRef](#)] [[PubMed](#)]
59. Flynn, J.D.; McGlinchey, R.P.; Walker, R.L.; Lee, J.C. Structural features of-synuclein amyloid fibrils revealed by Raman spectroscopy. *J. Biol. Chem.* **2018**, *293*, 767–776. [[CrossRef](#)] [[PubMed](#)]

60. Ngarize, S.; Herman, H.; Adams, A.; Howell, N. Comparison of changes in the secondary structure of unheated, heated, and high-pressure-treated β -lactoglobulin and ovalbumin proteins using Fourier transform Raman spectroscopy and self-deconvolution. *J. Agric. Food Chem.* **2004**, *52*, 6470–6477. [[CrossRef](#)] [[PubMed](#)]
61. Devitt, G.; Rice, W.; Crisford, A.; Nandhakumar, I.; Mudher, A.; Mahajan, S. Conformational evolution of molecular signatures during amyloidogenic protein aggregation. *ACS Chem. Neurosci.* **2019**, *10*, 4593–4611. [[CrossRef](#)]
62. Signorelli, S.; Cannistraro, S.; Bizzarri, A.R. Structural characterization of the intrinsically disordered protein p53 using Raman spectroscopy. *Appl. Spectrosc.* **2017**, *71*, 823–832. [[CrossRef](#)]
63. Celis, F.; Garcia, M.; Diaz-Fleming, G.; Campos-Vallette, M. A review of Raman, SURFACE-enhanced Raman scattering (SERS) and related spectroscopic techniques applied to biomolecules in biomaterials. *J. Chil. Chem. Soc.* **2017**, *62*, 3627–3632. [[CrossRef](#)]
64. Takeuchi, H.; Watanabe, N.; Satoh, Y.; Harada, I. Effects of hydrogen bonding on the tyrosine Raman bands in the 1300–1150 cm⁻¹ region. *J. Raman Spectrosc.* **1989**, *20*, 233–237. [[CrossRef](#)]
65. Schwenk, N.; Mizaikoff, B.; Cárdenas, S.; López-Lorente, Á.I. Gold-nanostar-based SERS substrates for studying protein aggregation processes. *Analyst* **2018**, *143*, 5103–5111. [[CrossRef](#)] [[PubMed](#)]
66. Ji, R.D.; Balakrishnan, G.; Hu, Y.; Spiro, T.G. Intermediacy of poly(L-proline) II and beta-strand conformations in poly(L-lysine) beta-sheet formation probed by temperature-jump/UV resonance Raman spectroscopy. *Biochemistry* **2006**, *45*, 34–41. [[CrossRef](#)]
67. Ji, M.; Arbel, M.; Zhang, L.; Freudiger, C.W.; Hou, S.S.; Lin, D.; Yang, X.; Bacskai, B.J.; Xie, X.S. Label-Free imaging of amyloid plaques in Alzheimer's disease with stimulated raman scattering microscopy. *Sci. Adv.* **2018**, *4*. [[CrossRef](#)]
68. Kurouski, D.; Van Duyn, R.P.; Lednev, I.K. Exploring the structure and formation mechanism of amyloid fibrils by Raman spectroscopy: A review. *Analyst* **2015**, *140*, 4967–4980. [[CrossRef](#)]
69. Dolui, S.; Mondal, A.; Roy, A.; Pal, U.; Das, S.; Saha, A.; Maiti, N.C. Order, disorder, and reorder state of lysozyme: Aggregation mechanism by raman spectroscopy. *J. Phys. Chem. B* **2020**, *124*, 50–60. [[CrossRef](#)] [[PubMed](#)]
70. Gras, S.L.; Waddington, L.J.; Goldie, K.N. Transmission electron microscopy of amyloid fibrils. *Methods Mol. Biol.* **2011**, *752*, 197–214.
71. Selivanova, O.M.; Galzitskaya, O.V. Structural polymorphism and possible pathways of amyloid fibril formation on the example of insulin protein. *Biochemistry* **2012**, *77*, 1237–1247. [[CrossRef](#)]
72. Gade Malmos, K.; Blancas-Mejia, L.M.; Weber, B.; Buchner, J.; Ramirez-Alvarado, M.; Naiki, H.; Otzen, D. ThT 101: A primer on the use of thioflavin T to investigate amyloid formation. *Amyloid* **2017**, *24*, 1–16. [[CrossRef](#)]
73. Manno, M.; Craparo, E.F.; Martorana, V.; Bulone, D.; San Biagio, P.L. Kinetics of insulin aggregation: Disentanglement of amyloid fibrillation from large-size cluster formation. *Biophys. J.* **2006**, *90*, 4585–4591. [[CrossRef](#)] [[PubMed](#)]
74. Debenedictis, E.P.; Ma, D.; Keten, S. Structural predictions for curly amyloid fibril subunits CsgA and CsgB. *RSC Adv.* **2017**, *7*, 48102–48112. [[CrossRef](#)]
75. Debenedictis, E.P.; Keten, S. Mechanical unfolding of alpha-and beta-helical protein motifs. *Soft Matter* **2019**, *15*, 1243–1252. [[CrossRef](#)] [[PubMed](#)]
76. Tian, P.; Boomsma, W.; Wang, Y.; Otzen, D.E.; Jensen, M.H.; Lindorff-Larsen, K. Structure of a functional amyloid protein subunit computed using sequence variation. *J. Am. Chem. Soc.* **2015**, *137*, 22–25. [[CrossRef](#)] [[PubMed](#)]
77. Wilkosz, N.; Czaja, M.; Seweryn, S.; Skirlińska-Nosek, K.; Szymonski, M.; Lipiec, E.; Sofińska, K. Molecular Spectroscopic markers of abnormal protein aggregation. *Molecules* **2020**, *25*, 2498. [[CrossRef](#)]
78. Desai, S.K.; Padmanabhan, A.; Harshe, S.; Zaidel-Bar, R.; Kenney, L.J. Salmonella biofilms program innate immunity for persistence in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 12462–12467. [[CrossRef](#)]
79. Kumar, D.K.; Choi, S.H.; Washicosky, K.J.; Eimer, W.A.; Tucker, S.; Ghofrani, J.; Lefkowitz, A.; McColl, G.; Goldstein, L.E.; Tanzi, R.E.; et al. Amyloid- β peptide protects against microbial infection in mouse and worm models of Alzheimer's disease. *Sci. Transl. Med.* **2016**, *8*, 340ra72. [[CrossRef](#)] [[PubMed](#)]
80. Martins, P.M.; Navarro, S.; Silva, A.; Pinto, M.F.; Sárkány, Z.; Figueiredo, F.; Barbosa Pereira, P.J.; Pinheiro, F.; Bednarikova, Z.; Burdukiewicz, M.; et al. MIRRAGGE—Minimum information required for reproducible AGGregation experiments. *Front. Mol. Neurosci.* **2020**, *13*, 222. [[CrossRef](#)] [[PubMed](#)]
81. Waterhouse, A.M.; Procter, J.B.; Martin, D.M.A.; Clamp, M.; Barton, G.J. Jalview version 2—A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **2009**, *25*, 1189–1191. [[CrossRef](#)]
82. Savitzky, A.; Golay, M.J.E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36*, 1627–1639. [[CrossRef](#)]
83. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; et al. Scikit-Learn: Machine learning in python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
84. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [[CrossRef](#)]

A spatiotemporal reconstruction of the *C. elegans* pharyngeal cuticle reveals a structure rich in phase-separating proteins

Muntasir Kamal^{1,2†}, Levon Tokmakjian^{2,3†}, Jessica Knox^{1,2†}, Peter Mastrangelo^{1,2}, Jingxiu Ji^{1,2}, Hao Cai⁴, Jakub W Wojciechowski⁵, Michael P Hughes⁶, Kristóf Takács⁷, Xiaoquan Chu⁸, Jianfeng Pei⁹, Vince Grolmusz⁷, Malgorzata Kotulska⁵, Julie Deborah Forman-Kay^{4,10}, Peter J Roy^{1,2,3*}

¹Department of Molecular Genetics, University of Toronto, Toronto, Canada; ²The Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Canada; ³Department of Pharmacology and Toxicology, University of Toronto, Toronto, Canada; ⁴Molecular Medicine Program, The Hospital for Sick Children, Toronto, Canada; ⁵Wroclaw University of Science and Technology, Faculty of Fundamental Problems of Technology, Department of Biomedical Engineering, Wroclaw, Poland; ⁶Department of Cell and Molecular Biology, St. Jude Children's Research Hospital, Memphis, United States; ⁷PIT Bioinformatics Group, Institute of Mathematics, Eötvös University, Budapest, Hungary; ⁸Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China; ⁹Department of Computer Science and Technology, Tsinghua University, Beijing, China; ¹⁰Department of Biochemistry, University of Toronto, Toronto, Canada

*For correspondence: peter.roy@utoronto.ca

†These authors contributed equally to this work

Competing interest: The authors declare that no competing interests exist.

Funding: See page 29

Preprinted: 14 March 2022

Received: 11 April 2022

Accepted: 11 October 2022

Published: 19 October 2022

Reviewing Editor: Luisa Cochella, Johns Hopkins University School of Medicine, United States

© Copyright Kamal, Tokmakjian, Knox et al. This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

Abstract How the cuticles of the roughly 4.5 million species of ecdysozoan animals are constructed is not well understood. Here, we systematically mine gene expression datasets to uncover the spatiotemporal blueprint for how the chitin-based pharyngeal cuticle of the nematode *Caenorhabditis elegans* is built. We demonstrate that the blueprint correctly predicts expression patterns and functional relevance to cuticle development. We find that as larvae prepare to molt, catabolic enzymes are upregulated and the genes that encode chitin synthase, chitin cross-linkers, and homologs of amyloid regulators subsequently peak in expression. Forty-eight percent of the gene products secreted during the molt are predicted to be intrinsically disordered proteins (IDPs), many of which belong to four distinct families whose transcripts are expressed in overlapping waves. These include the IDPAs, IDPBs, and IDPCs, which are introduced for the first time here. All four families have sequence properties that drive phase separation and we demonstrate phase separation for one exemplar in vitro. This systematic analysis represents the first blueprint for cuticle construction and highlights the massive contribution that phase-separating materials make to the structure.

Editor's evaluation

Cuticles are specialized extracellular matrices that cover the bodies of ecdysozoans, which make up 85% of all animals, and how cuticles are formed is very poorly understood, in particular in light of the fact that cuticles are shed and regrown as animals grow. The authors present a comprehensively and carefully curated resource of the components of the pharyngeal cuticle of *C. elegans* and provide a

spatiotemporal framework to understand cuticle assembly. In doing so, the authors propose a function for a large class of intrinsically disordered proteins (IDPs). The significance of this work is high because our understanding of both cuticle formation and of IDPs is poor.

Introduction

Over 85% of living animal species belong to the superphylum ecdysozoa. This group includes nematodes, arthropods, tardigrades, and five other phyla (Telford *et al.*, 2008; Aguinaldo *et al.*, 1997). They are defined by having a common ancestor and a specialized extracellular matrix that covers their body called the cuticle. The ecdysozoan cuticle is shed and regrown to accommodate juvenile growth in a process called ecdysis or molting.

Cuticle shape is patterned by the tissue beneath it, but also takes on additional diversity beyond the underlying tissue shape. One example of this structural diversity is the mouthparts of nematodes. Many carnivorous nematodes and nematode parasites of animals have cuticle-based teeth that bite into their prey or host (Sieriebriennikov and Sommer, 2018; John and Petri, 2006). Nematode parasites of plants have needle-like cuticle stylets that pierce plants and act as a syringe to deposit effectors and suck out vital nutrients (Mejias *et al.*, 2019). Bacterivorous nematodes, like the model nematode *Caenorhabditis elegans*, have cuticle grinders that pulverize bacteria into digestible bits (Sparacio *et al.*, 2020). These specialized mouthparts are variations of the cuticle that lines the anterior alimentary tract. Despite this diversity in form and the importance of the cuticle to most animals, a spatiotemporal blueprint for cuticle construction is lacking. Here, we provide such a blueprint by mining published datasets of *C. elegans* gene expression.

All epithelia in *C. elegans* that would otherwise be exposed to the environment, except the intestine, are protected by a cuticle. These include the body cuticle that protects the hypodermis (aka epidermis), the anterior alimentary cuticle that reinforces the lumen of the buccal cavity and pharynx, and other cuticles that protect the rectum, vulva, and excretory pore tissues (Altun and Hall, 2020). Here, we will refer to the anterior alimentary cuticle as the pharyngeal cuticle.

The non-chitinous body cuticle has multiple layers that include an outer carbohydrate-rich glycocalyx, a lipid-rich epicuticle, and multiple inner collagenous layers (Altun and Hall, 2020; Page and Johnstone, 2007; Cox *et al.*, 1981). By contrast, the pharyngeal cuticle is not collagenous (Altun and Hall, 2020; Cox *et al.*, 1981) and instead contains a chitin-chitosan matrix that likely helps maintain luminal integrity (Zhang *et al.*, 2005; Heustis *et al.*, 2012). The pharyngeal cuticle is layered (Sparacio *et al.*, 2020; Wright and Thomson, 1981), but the molecular composition of the different layers is unknown. Like other ecdysozoans, *C. elegans* sheds its cuticles at the end of each larval stage. As the old cuticle is being shed, a new cuticle is built underneath, and the next developmental stage ensues (Sparacio *et al.*, 2020; Lazetic and Fay, 2017). *C. elegans* adults do not molt.

In addition to chitin, the pharyngeal cuticle contains a group of largely disordered proteins called the APPGs (also known as the ABU/PQN Paralog Group) (George-Raizen *et al.*, 2014). The APPGs are low complexity (i.e., they have a biased composition involving a limited set of amino acids) and have been described as prion-like (Michelitsch and Weissman, 2000) and potentially amyloidogenic (George-Raizen *et al.*, 2014). An examination of the expression pattern of five APPGs showed that all five are expressed in cells that surround the pharyngeal cuticle and that APPG::GFP fusion proteins are incorporated into the pharyngeal cuticle (George-Raizen *et al.*, 2014). The disruption of two of these genes exhibits feeding phenotypes consistent with disruption of this cuticle (George-Raizen *et al.*, 2014). In this study, we find the APPGs to be one of several groups of proteins dominated by large intrinsically disordered regions (IDRs) with low-complexity sequences that are likely secreted into the developing pharyngeal cuticle.

IDRs are defined here as a 30 or more continuous residues whose primary sequence fails to form a stereotypical stable tertiary structure and instead rapidly interconverts between heterogeneous conformations (van der Lee *et al.*, 2014). Despite lacking ordered structure, IDRs can interact with other IDRs through local areas of hydrophobicity, complementary charge, hydrogen-bond formation, and pi-stacking interactions along the respective peptide chains (Vernon and Forman-Kay, 2019). IDRs often harbor repeating sequence features that can facilitate the formation of multivalent interaction networks with multiple binding partners (Vernon and Forman-Kay, 2019). Depending on the local environment, multivalent IDRs, and particularly low-complexity IDRs, can phase separate to

form liquid–liquid phase-separated droplets (LLPS) (i.e., liquid condensates) or gels, which can then transition to more solid structures, including fibers (Mittag and Parker, 2018; Banani et al., 2017). LLPS has been shown to be an important first step in the self-assembly of IDR-rich proteins into the extracellular matrices of insects, arachnids, and molluscs (reviewed in Muiznieks et al., 2018). For example, IDR-rich proteins that form liquid condensates fill a porous chitin-based matrix in a key step of squid beak development (Tan et al., 2015). Given that the affinity of any one interaction along an IDR is relatively weak, the ability of IDRs to form these phase-separated networks is easily modulated by a variety of factors, including pH, ions, temperature, protein concentration, and post-translational modifications (Murray et al., 2017).

Here, we describe the spatiotemporal logic of pharyngeal cuticle construction that we have uncovered by mining published mRNA expression datasets and canonical amyloid and chitin-binding dyes. We identify six families of low-complexity proteins that are likely secreted into the developing cuticle, including the IDPAs, IDPBs, and IDPCs, each of which are described for the first time here, and the APPGs, NSPBs, and the FIPRs. These six families peak in expression level in successive waves over the course of each larval stage. Computational analyses predict that the IDPA, IDPB, IDPC, and APPG families, and 12 other singletons are IDR-rich proteins capable of phase separation. We speculate that the malleable properties of the disordered phase-separating proteins are especially suited to a flexible cuticle that must be rapidly destroyed and reconstructed during molting.

Results

Validating fluorescent dyes as probes of pharyngeal cuticle structure

Earlier transmission electron microscopy of the *C. elegans* pharynx cuticle revealed it to be a complex structure that changes in character along its anterior–posterior axis (Sparacio et al., 2020; Wright and Thomson, 1981; White et al., 1986; Figure 1). To further characterize its structure, we first sought to validate dyes as probes of the cuticle. Congo Red (CR) fluoresces red and binds to amyloid oligomers, protofibrils, and fibrils (Bennhold, 1922; Wu et al., 2012) and has been previously shown to stain the cuticular grinder of the pharynx (George-Raizen et al., 2014). Thioflavin S (ThS) increases in blue fluorescence emission upon binding amyloid structures (Vassar and Culling, 1959). Calcofluor white (CFW) fluoresces deep blue and is used as a chitin probe in other systems (Roncero et al., 1988). Eosin Y (EY) is a yellow-red fluorescent dye that binds chitosan, which is the deacetylated form of chitin (Baker et al., 2007).

We confirmed that the four dyes specifically bind components within the pharyngeal cuticle in two ways. First, we performed pulse-chase experiments with the dyes to determine whether the dye's fluorescent signal would be lost as the larvae shed their old cuticle during their transition to the next developmental stage (see 'Materials and methods' for details). After the 18 hr chase, very few animals who were initially L3s had CFW, EY, CR, or ThS signal (Figure 2, Figure 2—figure supplement 1). By contrast, the dyes' signal persisted in animals that were initially young adults (Figure 2). The loss of the four dyes from the larvae but not adults in the pulse-chase experiments indicates that the dyes bind the pharyngeal cuticle.

Second, we tested whether the dyes bind the pharyngeal cuticle after the cuticle has separated from the animal, the attachment of which persists in *mlt-9(RNAi)* mutants (Frاند et al., 2005). We found that all four dyes bind the exterior pharyngeal cuticle of *mlt-9(RNAi)* animals (Figure 2S–X). As a positive control, we find that GFP-tagged ABU-14 is retained in the shed pharyngeal cuticle (Figure 2Y). These data establish CR, ThS, CFW, and EY as specific probes of the pharyngeal cuticle.

Cuticle dyes stain distinct structures within the pharyngeal cuticle

We examined the colocalization of the four dyes in wildtype animals and correlated the resulting patterns to the ultrastructural features observed in a series of unpublished TEM images by Kenneth A. Wright and Nicole Thomson (Wright and Thomson, 1981; Figure 3). These TEM images show that the cuticle of the buccal cavity and the channels is a mixture of electron-light and electron-dense (dark) material, with the dark material forming circumferential ribs (white arrows) and 'flaps' (yellow arrows).

Two features suggest that the chitin-binding dyes may bind components within the electron-light material. First, the expansive electron-light material at the anterior half of the buccal cuticle correlates

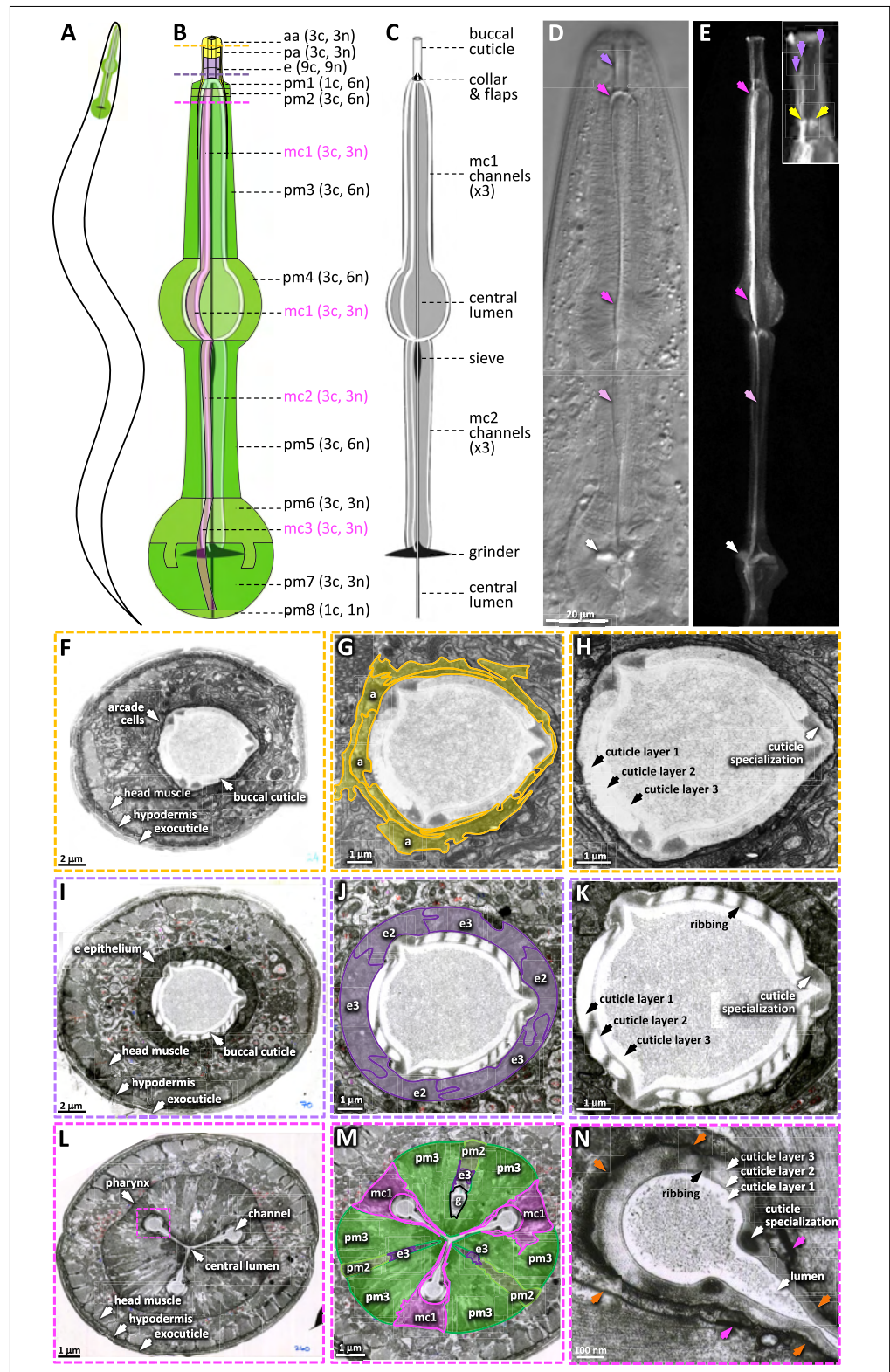


Figure 1. The pharyngeal cuticle and surrounding cells. **(A)** A schematic of the relative position of the *C. elegans* pharynx (green). **(B)** A schematic of the pharynx. The image of the outer cells is transparent, revealing the pharyngeal cuticle underneath. The 5 cells of the gland and the 20 pharynx-associated neurons are not shown. Each of the cell types are labeled followed by individual number of cells (c) and nuclei (n). aa, anterior arcade
Figure 1 continued on next page

Figure 1 continued

cells; pa, posterior arcade cells; e, pharynx epithelium; pm1-8, pharynx muscle; mc1-3, marginal cells. The yellow, purple, and pink dashed lines represent the area of the cross sections in (F–H), (I–K), and (L–N), respectively. (C) A schematic of the pharyngeal cuticle. Black and gray is cuticle; white is the lumen of the buccal cavity, central lumen, and channels. (D, E) Micrographs of the head of young adults expressing ABU-14::sfGFP. Differential interference contrast (DIC) is on the left and GFP of a similarly staged animal, taken with confocal microscopy, is on the right. Purple arrows show the buccal cuticle. The three purple arrows in the inset mark regions of ABU-14::GFP enrichment that likely correspond to the cuticle specializations noted in (H) and (K). Yellow arrows, the metastomal flaps; dark pink arrows, mc1 channel; light pink arrows, mc2 channel; white arrows, grinder. (F–N) TEM images taken from the *White et al., 1986* N2T series, stored on the WormAtlas EM archives. (F–H) show a cross section of the anterior buccal cavity; the surrounding arcade cells are highlighted in yellow in (G). (I–K) show a cross section of the posterior buccal cavity; the surrounding e epithelial cells are highlighted in purple in (J). (L–N) show a cross section of the procorpus posterior to the buccal cavity. In (M), the mc1 marginal cells associated with the channels are highlighted in pink; the pharyngeal muscles pm2 and pm3 are highlighted in green and 'g' indicates the gland. The pink box in (L) indicates the magnified area in (N). Orange arrows in (N) indicate the pm3-mc1 plasma membrane interface; the pink arrows indicate the adherens junctions.

with the expanded CFW and EY signal (orange arrows in **Figure 3A and E**). Second, CFW and EY brightly stain a prominent collar at the base of the buccal cavity (green arrows in **Figure 3A and E**). The amyloid-binding dyes stain the collar less (**Figure 3B and C**), and ABU-14::GFP fails to mark the collar (**Figure 3D**). In the TEM images, this collar is composed of light material. Hence, the electron-light material is likely enriched with chitin.

The CR dye and the ABU-14::GFP localize to the cuticle flaps (yellow arrows in **Figures 1E and 3B and D**), which are composed of the darker electron-dense material in the TEM (**Figure 3E**). The dark material of the flaps is contiguous with the dark ribbing of the buccal cuticle and the luminal-facing coating of the cuticle, all of which encapsulate the less electron-dense material (**Figure 3E**). An analogous organization is present in the cuticle that lines the channels (**Figure 3E**). Together, these observations suggest that the electron-dense material may be enriched in amyloid-like proteins and establish CR, ThS, CFW, and EY as useful markers of pharyngeal cuticle structure.

Mining expression datasets yields a spatiotemporal map of pharyngeal cuticle development

To better understand pharynx cuticle construction, we built a spatiotemporal map of cuticle-centric gene expression by combining four published datasets (see **Figure 4—source data 1**). First, we anchored the map using a dataset that tracked gene expression levels in synchronized animals every hour for 16 hr from the mid L3-stage to adulthood at 25°C (*Hendriks et al., 2014*). This study identified 2718 genes whose expression oscillates during larval development with a peak in expression every 8 hr ($p < 0.001$); this period corresponds to the 8 hr duration of the third and fourth larval stages at 25°C. Two of these 2718 genes have been retired due to reannotation. The 2716 genes can be grouped into bins of genes that peak at different larval development phases. For example, some genes peak during the first and ninth hour, others peak during second and tenth hour etc., such that there are successive waves of genes that oscillate through time (see **Figure 1e** of *Hendriks et al., 2014*). We present the 2716 genes from this dataset in the temporal order in which the genes peak in their expression over the 8 hr cycle (**Figure 4A**). We note that since we initiated our study an additional temporally resolved dataset has been published (*Meeuse et al., 2020*).

Second, we defined the interval on the map that corresponds to the molt by overlaying a dataset of genes that are upregulated during the L4 molt ($p < 0.001$) (*George-Raizen et al., 2014*). The overlay indicates that molting peaks in the sixth hour on the map (**Figure 4A and B**, **Figure 4—source data 1**). The fact that the genes that are upregulated during the L4 molt are clustered on the map provides reciprocal validation for both datasets (*George-Raizen et al., 2014; Hendriks et al., 2014*). We herein routinely refer to hour 6 as the reference peak molting hour.

Third, we identified the genes on the temporal map whose expression is enriched in the cells surrounding the pharyngeal cuticle relative to all other tissues. We did this by overlaying single-cell expression data from cells isolated from L2-staged animals (*Cao et al., 2017*). We found 367 'pharynx'-enriched transcripts (≥ 1.5 -fold enriched in the pharynx relative to all other tissues and at least 25 transcripts per 1 million reads) that oscillate over time (**Figure 4A**, **Figure 4—figure supplement 1**,

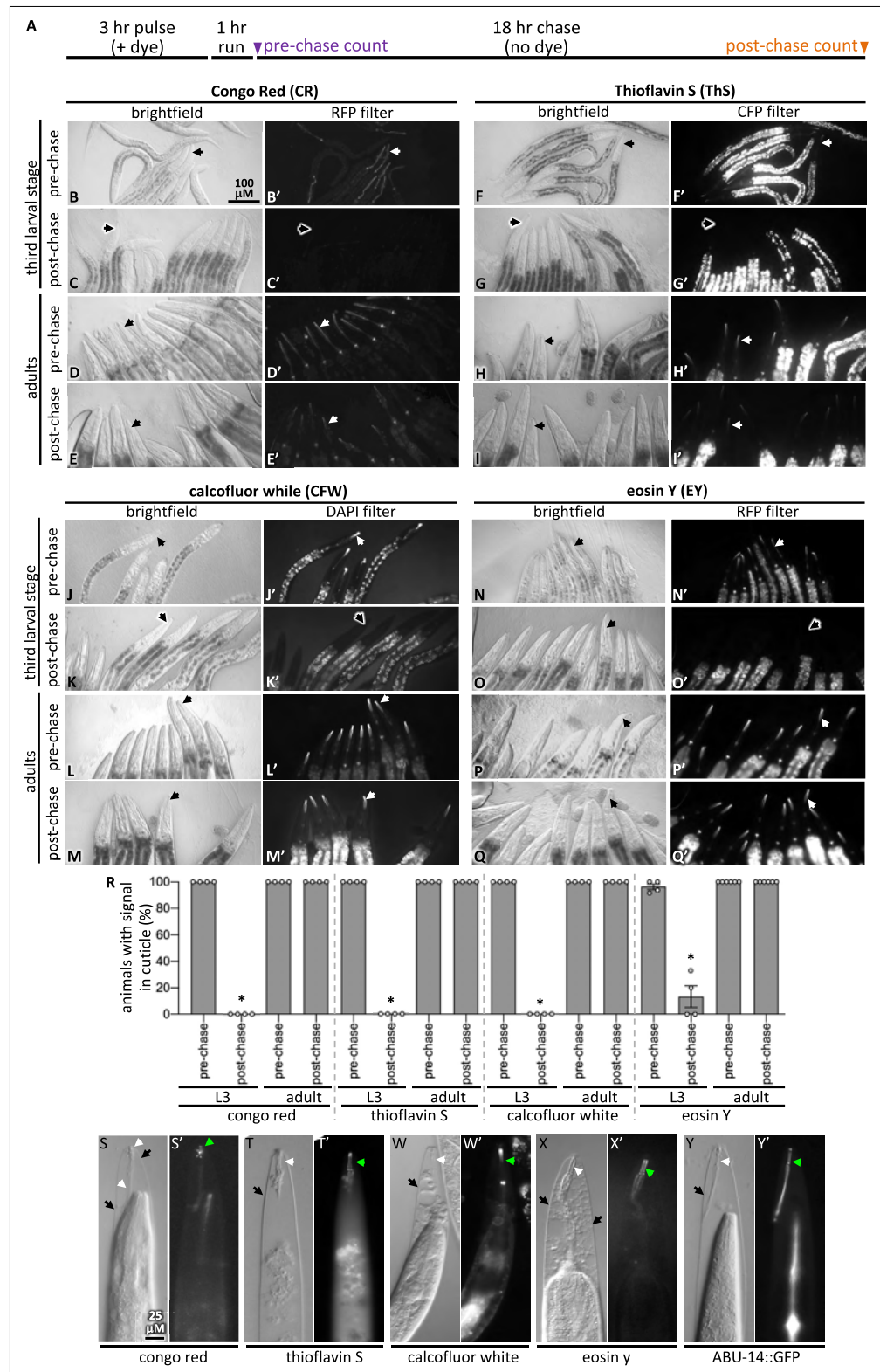


Figure 2. Pulse-chase and cuticle mutant experiments show dye association with the cuticle. **(A)** Schematic showing the pulse-chase assay. Synchronized populations of L3 or adult worms were incubated with a dye for 3 hr (the ‘pulse’), after which worms were washed with M9 and run on normal plates with food for 1 hr. Worms are transferred to fresh plates and the presence of the dye was scored (see ‘Materials and methods’ for details).

Figure 2 continued on next page

Figure 2 continued

Then, 18 hr later (i.e., after the chase), worms were again scored for the presence of the dye. **(B–Q)** In each of the four groups of eight micrographs with the dye indicated in the header, the top two rows show the pulse-chase experiment done starting with L3s, and the bottom two rows show the pulse-chase experiment done with adults. The filter used to visualize the dyes is indicated at the header of the rightmost column in each of the four panel sets. In all panels, white arrows highlight the presence of the dye in the cuticle and black arrows show cuticle without dye signal. The scale bar is indicated. **(R)** The fraction of worms with stained cuticle before and after the chase for each dye is shown; a minimum of four repeats **(N)** were done with a sample size of 7–34 animals (average = 13) **(n)** per repeat. Asterisk denotes statistically significant difference relative to the pre-chase values ($p < 0.05$). Standard error of the mean is shown. **(S–X)** Wildtype animals treated with *mlt-9(RNAi)* that are incubated with the indicated dye for 3 hr. The brightfield differential interference contrast (DIC) image and the corresponding fluorescent image are shown for each treatment. **(Y)** An animal expressing transgenic ABU-14::GFP treated with *mlt-9(RNAi)* but without dye stain. The scale in **(S)** applies to all panels.

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. The fluorescence and filter controls for dye staining.

Figure 4—source data 1. This set of genes includes those enriched in expression within the pharyngeal epithelium, muscles, and gland cells, but not pharyngeal-associated neurons.

Fourth, we determined the likelihood of gene products being secreted using Signal P (v4.1) predictions extracted from the WormBase Parasite database to identify signal peptides (with scores of 0.45 or more) genome-wide (Hertz-Fowler and Hall, 2004). We recognize that while this approach is systematic, Signal P does not identify all secreted or plasma membrane-associated transmembrane proteins. The oscillating pharynx-enriched set contained 226 genes (62%) that encode a signal peptide (Figure 4A, Figure 4—source data 1). By comparison, only 39% of the remaining oscillating gene set ($n = 2349$) and only 17% of the entire non-oscillating genes of the genome ($n = 17,614$) encode a signal peptide (Figure 4—source data 1). The temporal map shows a concentration of genes that peak in expression from the pharynx and are secreted at the time of molting (Figure 4A).

We investigated the change in transcript abundance in the pharynx over the cyclical 8 hr window of larval development for the oscillating genes. We found a nearly 30-fold increase in transcript abundance for those gene products predicted to be secreted relative to the global average of pharynx gene expression during the peak molting hour (Figure 4B). There is a shoulder of peak expression at hour 7 for those non-secreted gene products (Figure 4B) that may correspond to the increase in tissue growth after the molt. Cao et al., 2017 further dissected their single-cell sequencing data into tissue subtypes. We find that the expression of predicted secreted products from the pharynx epithelial cells peaks dramatically during the peak molting hour, whereas pharynx gland transcription peaks in the preceding hour (Figure 4). Non-secreted epithelial and muscle products peak in expression during hour 7 (Figure 4C and E). Given that mRNA expression levels are positively correlated with protein abundance in invertebrate systems (Ho et al., 2018; Schrimpf et al., 2009), we conclude that there is a likely a burst of proteins secreted in preparation for the molt.

Orthogonal data validate the spatiotemporal map

We explored the validity of the spatiotemporal map in four ways. First, previous work established that the molting of the body cuticle precedes that of the pharyngeal cuticle (Wright and Thomson, 1981). We therefore expected a peak in gene expression from the hypodermis that precedes that of the pharynx, which is what we observe (Figure 5A).

Second, we systematically investigated published reports of expression (not including the datasets used to build the spatiotemporal map) for the 226 oscillating pharynx secretome genes. In this analysis, we also included the 17 additional genes of special interest called out in Figure 4A that include *myo-1*, *myo-2*, and *myo-5* for example (see Supplementary file 1 for details). We surveyed Yuji Kohara's whole-mount RNA in situ database (Motohashi et al., 2006) and literature reports of transgene and sequencing-based expression patterns curated by WormBase to determine whether there is additional evidence that these 243 genes are enriched in expression within the pharynx (Supplementary file 1). 83 (34%) of the 243 genes lacked reported expression patterns in the Kohara and WormBase databases. Of the remaining 160, 152 (95%) demonstrate a clear enrichment of expression within the pharynx (Figure 5B; Supplementary file 1).

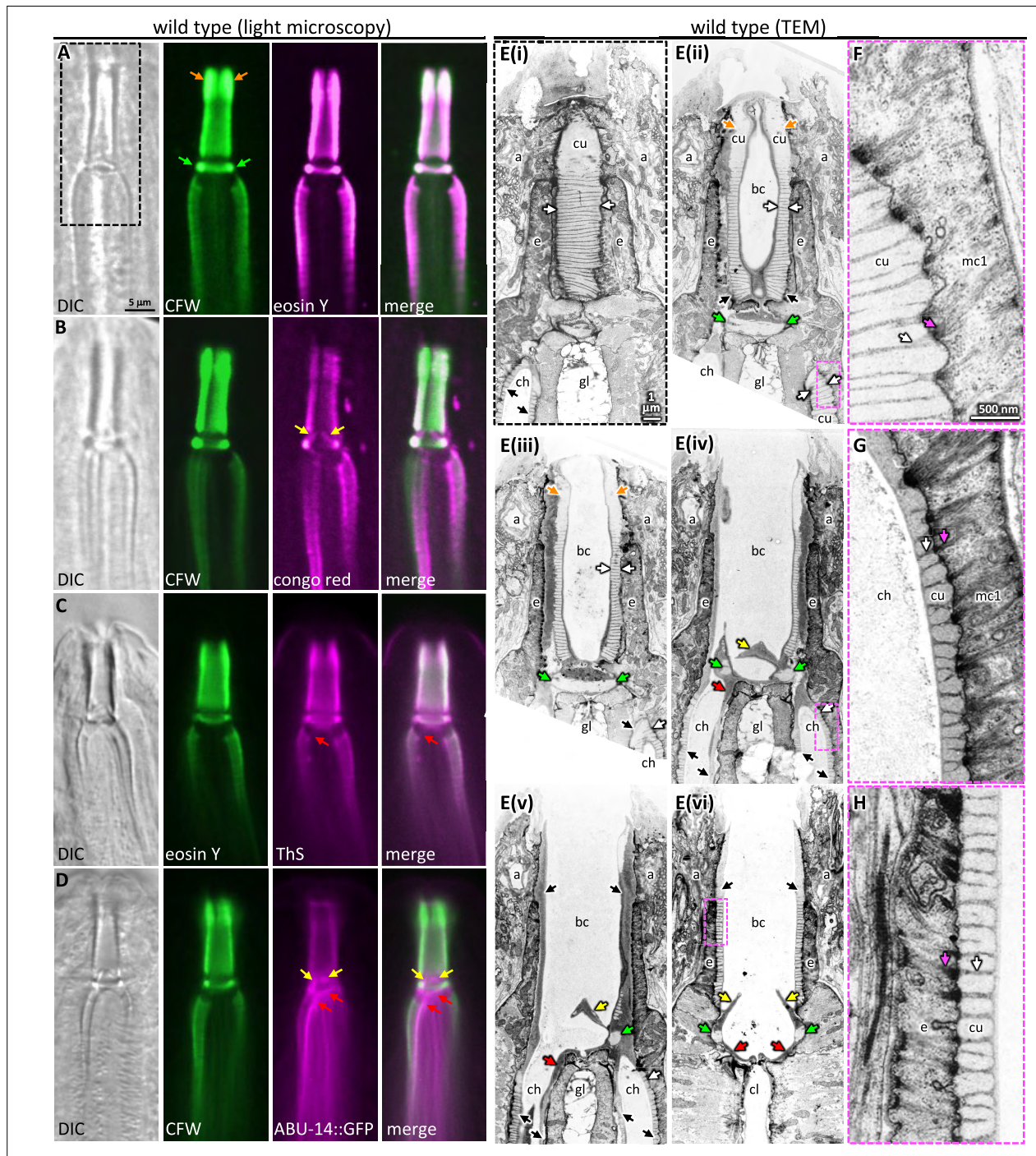


Figure 3. Probing pharyngeal cuticle composition with characterized dyes. (A–D) Images of the buccal and mc1 channel cuticles and surrounding cells. The dyes or GFP-fusion protein examined is indicated. DIC, differential interference contrast; CFW, calcofluor white; ThS, thioflavin S. The scale shown in (A) is the same for (B–D). (E) Serial coronal sections of unpublished transmission electron micrographs taken by *Wright and Thomson, 1981*. The scale in (E(i)) applies to all images in the E series. (F–H) Magnifications of the boxed areas highlighted in the images to the left. (G) represents a slightly different plane than that depicted in (E(iv)) and was chosen because of the clearly visible filaments. The scale in (F) is the same as that for (G) and (H). For all panels: a, arcade cells; ch, mc1 channel; cu, cuticle; e, e epithelium; gl, gland cell; bc, buccal cavity; cl central lumen; orange arrows, the anterior enlargement of the buccal cuticle; green arrows, the prominent ring at the base of the buccal cavity; yellow arrows, the electron-dense flaps at the base of the buccal cavity; red arrows, the electron-dense material at the anterior end of the channel cuticle; black arrows, pharyngeal cuticle when too small to be labeled with ‘cu’; white arrows, the ribbing of the pharyngeal cuticle; pink arrows, the cytoplasmic filaments that correspond to the abutment of the ribbing.

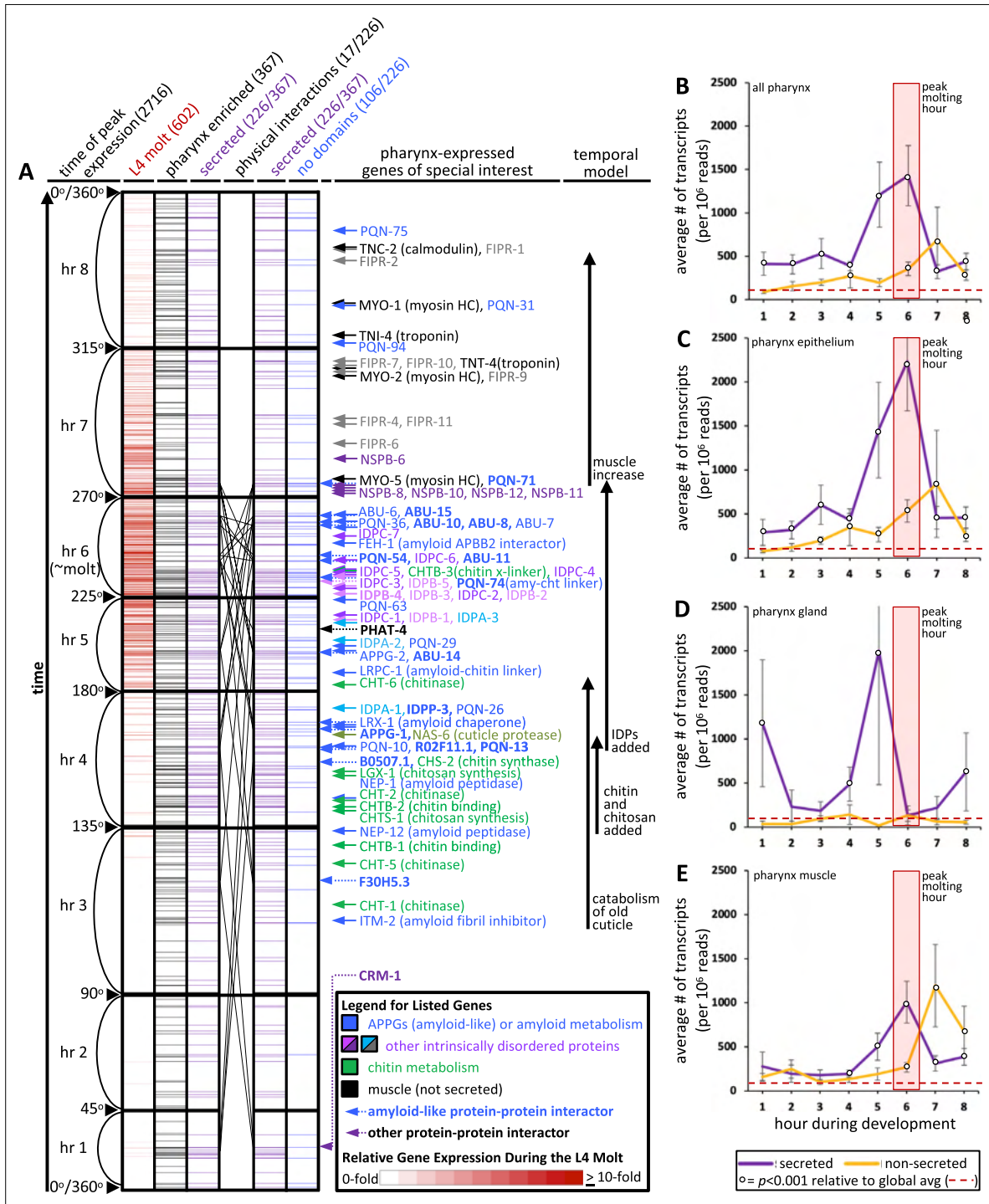


Figure 4. An informatic reconstruction of the pharyngeal cuticle. **(A)** A chart of 2716 genes whose expression oscillates over larval development with a periodicity that corresponds to larval stages. See text for details. Each row represents a single gene. Rows are arranged along the y-axis in order of the time at which each gene reaches its peak expression level with those earliest in the time course at the bottom and those latest in the period at the top. Because the periodicity is a continuum during larval development, *Hendriks et al., 2014* represented time as degrees of a circle. That concept is preserved here, and the degree is indicated along the y-axis and divided into bins of time relative to the molting period. The first data column (red) represents the 602 oscillating genes that also were found to be upregulated in expression during the L4 lethargus (molting) period (see Supplemental Table 1 in *George-Raizen et al., 2014*); the scale of the relative expression level from this independent study (*George-Raizen et al., 2014*) is indicated in the legend. The second data column (black) represents the 367 genes from the set of 2716 that are enriched in expression in the pharynx (data from *Figure 4 continued on next page*

Figure 4 continued

Cao et al., 2017; see Figure 4—source data 1 for enrichment). The purple columns show the 226 genes (of the 367 pharynx-enriched set) that are predicted to be secreted. They are duplicated to show the 26 protein–protein interactions (PPI) among the 17 oscillating pharynx-secreted proteins identified through Genemania (see Figure 5C, Figure 4—source data 1, and Figure 5—source data 1 for the details of which protein pairs interact). The identity of the interacting proteins is indicated with bold lettering and a dotted arrow on the right of the graph. The last column (blue) represents those pharynx-enriched genes that lack an obvious domain as predicted by WormBase, PFAM, and SMART databases (see text for details). 78 pharynx-expressed genes of special interest are indicated with arrows to the right of the graph. The color of the arrows and text corresponds to broad categories indicated in the legend. (B–E) The average number of transcripts produced by genes whose expression is enriched in the indicated tissue as a function of developmental time. In all graphs, results are binned according to the hours indicated in (A), the global average transcript number (49.33) is indicated by the red dotted line. Statistical differences were measured using a Student's t-test against the global average of gene expression levels in the pharynx. Standard error of the mean is shown in all graphs. The peak molting hour in (B–E) is highlighted by the transparent red box.

The online version of this article includes the following source data and figure supplement(s) for figure 4:

Source data 1. This is the master file with all relevant data for the spatiotemporal map.

Figure supplement 1. Tissue-enriched expression levels of tissue-enriched classes of genes.

Third, we reasoned that the pharynx secretome might be rich in protein–protein interactions (PPIs) because many of the secreted proteins likely interact to form a matrix. We explored PPIs systematically using Genemania, which is an online tool that facilitates the analysis of experimentally derived interaction data curated from the literature (Franz et al., 2018). To analyze each tissue's secretome, we returned to the Cao et al., 2017 single-cell sequence data to parse the proteome into proteins that are enriched in the major tissues using the same criteria described above for the pharynx (Figure 4—source data 1). These tissues included the pharynx (470 proteins), body wall muscles (BWMs) (326 proteins), glia (426 proteins), gonad (832 proteins), hypodermis (411 proteins), intestine (781 proteins), and neurons (965 proteins) (Figure 4—figure supplement 1). We separated out the 166 collagens from the proteome because of their unique sequence properties. The remaining 15,892 proteins are binned into a non-specific group. For each of these groups, we parsed them into those encoding a signal peptide, and those without. Genemania reports multiple lines of evidence for 36 PPIs among a network of 20 proteins within the pharynx secretome (Figure 5C). This interaction network is denser than that from most other secretomes (Figure 5—figure supplement 1, Figure 5—source data 1).

Fourth, literature searches reveal that the spatiotemporal map includes many genes with known roles in pharynx development (*feh-1*, *myo-1*, *myo-2*, *nep-1*, *pqn-75*, *sms-5*, *tnc-2*, and *tmi-4*) and the few genes known to play roles in pharynx cuticle formation (*abu-6*, *abu-14*, *chs-2*, and *nas-6*) (Supplementary file 1). We further investigated the functional relevance of the map by conducting a survey of publicly available mutants of genes predicted to contribute to the pharyngeal cuticle. Light microscopy revealed obvious cuticle defects in the pharynx of animals harboring disruptions of *feh-1*, *idpa-3*, *idpc-1*, *lrpc-1*, and the positive control *nas-6* (Figure 5D; Supplementary file 1), bringing the total number of genes with known pharynx cuticle defects to 7 of the 243 genes listed in Supplementary file 1. The pattern of amyloid and chitin dyes is unaligned in the *feh-1*, *idpa-3*, *idpc-1*, and *lrpc-1* mutants (Figure 5D). This not only provides insight into the proteins' importance in cuticle structure, but reinforces the idea that the two dyes recognize distinct components within the cuticle.

Finally, we further confirmed the map's ability to predict spatial expression patterns by inserting green fluorescent protein coding sequence in frame with five poorly characterized gene products, namely, IDPA-3, IDPB-3, IDPC-1, FIPR-4, and NSPB-12 (Figure 6). We also included the previously characterized ABU-14::GFP (Figure 6A). We counterstained the resulting transgenic animals with CFW to interrogate the spatial overlap of the tagged proteins with the chitinous cuticle. As predicted, we found that all five reporters are expressed exclusively in association with the pharynx and overlap in their localization with the pharynx cuticle. Briefly, tagged IDPA-3 was enriched in the grinder, overlapping the CFW-stained component and lining of the terminal bulb cuticle. In addition, we observed enrichment of tagged IDPA-3 in the presumptive ECM that lies between the terminal bulb and the intestinal valve (white arrow in Figure 6B). Tagged IDPB-3 was expressed weakly and localized exclusively to the pm6 cells and material surrounding the CFW-stained grinder (Figure 6C). Tagged IDPC-1 had a similar pattern to that of tagged ABU-14; associating with both the anterior and posterior components of the pharyngeal cuticle. However, tagged ABU-14 appears to localize adjacent to CFW-stained components whereas tagged IDPC-1 overlaps CFW-stained components (Figure 6A and D, Figure 6—figure supplement 1). Tagged NSPB-12 localized to the anterior pharynx cuticle

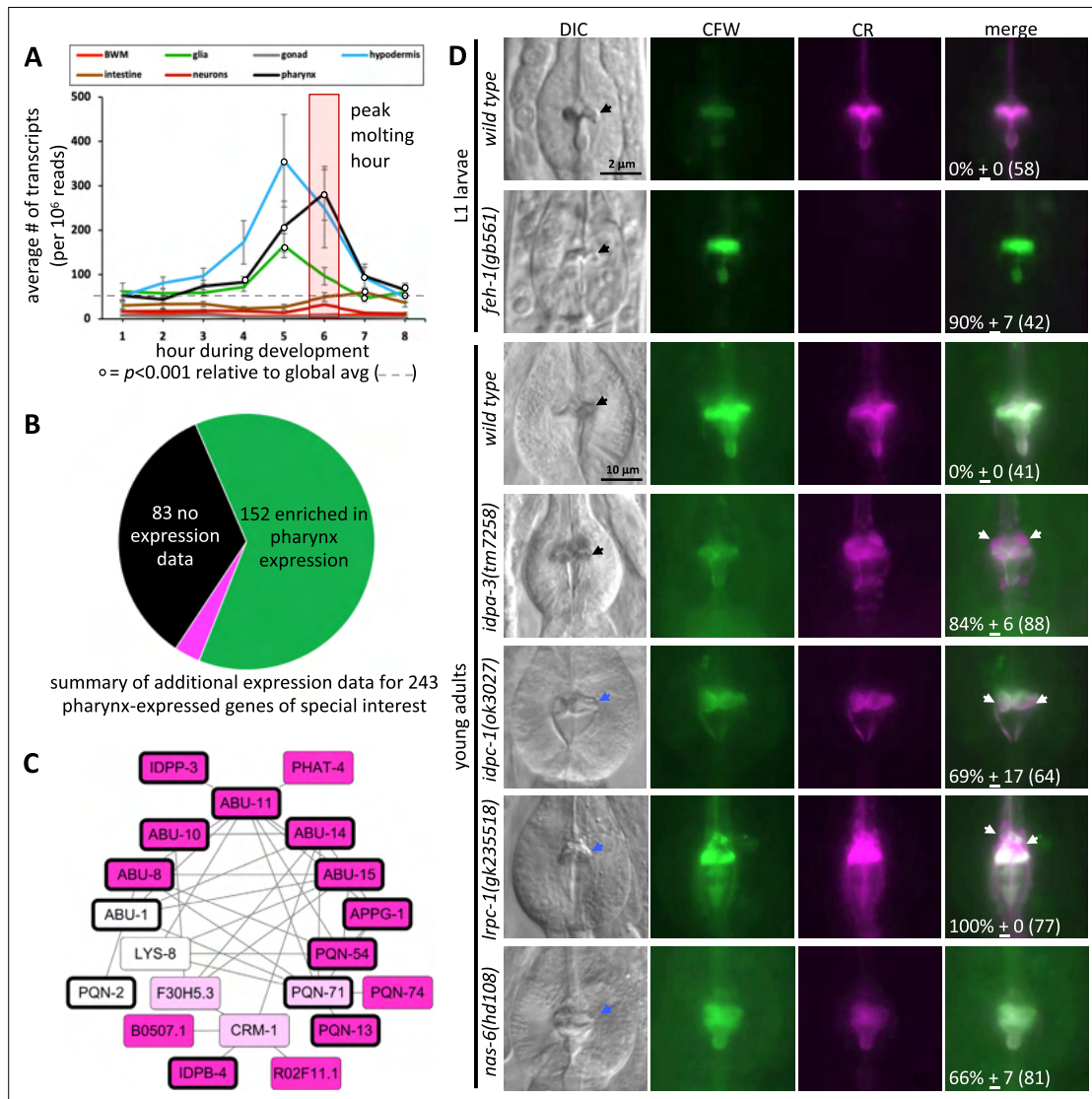


Figure 5. The spatiotemporal map has predictive power. **(A)** Average gene expression in each of the indicated tissue types plotted as a function of developmental time. In the first hour of the time course, for example, 219 genes peak in expression and the average expression of each of these 219 genes in each of the indicated tissues is plotted for hour 1 on the graph. Standard error of the mean is shown. The peak pharynx molting hour is highlighted by the transparent red box. Significant differences relative to the global mean is calculated with a Student's *t*-test. **(B)** A pie chart summarizing the search of publicly available information on previously documented expression patterns of the 226 oscillating pharynx secretome genes and 17 other genes of interest (which are part of the 78 genes highlighted in **Figure 4A**). Published expression patterns could be found for 160 of the 243 genes. Of the 160, the expression pattern of only 8 genes (indicated in fuchsia) did not support clear enrichment in the pharynx. See **Supplementary file 1** for details. **(C)** Protein–protein interactions within the pharynx secretome. Dark pink nodes are those genes that peak in expression during hours 4, 5, or 6 on the spatiotemporal map. Light pink nodes peak in expression outside of hours 4, 5, or 6. White nodes represent genes that do not oscillate. Nodes outlined in bold are those proteins composed of >75% intrinsically disordered regions (IDRs). **(D)** A survey of mutants for obvious pharynx cuticle defects. Each of the indicated backgrounds are stained with calcofluor white (CFW) and Congo Red (CR). The mean percentage of animals showing defects, together with the standard error of the mean ($N = 3$ independent trials with more than eight animals each trial). The total number of animals surveyed is indicated in brackets. The scale for L1 and adult animals is shown. DIC, differential interference contrast, black arrows indicate a normal terminal bulb grinder, blue arrows indicate a dysmorphic grinder, and white arrows indicate discordant CR staining.

The online version of this article includes the following source data and figure supplement(s) for figure 5:

Source data 1. Supporting information for the network diagram in **Figure 5C** and related insights.

Figure supplement 1. The pharynx secretome has a dense protein–protein interaction (PPIs) network relative to other secretomes.

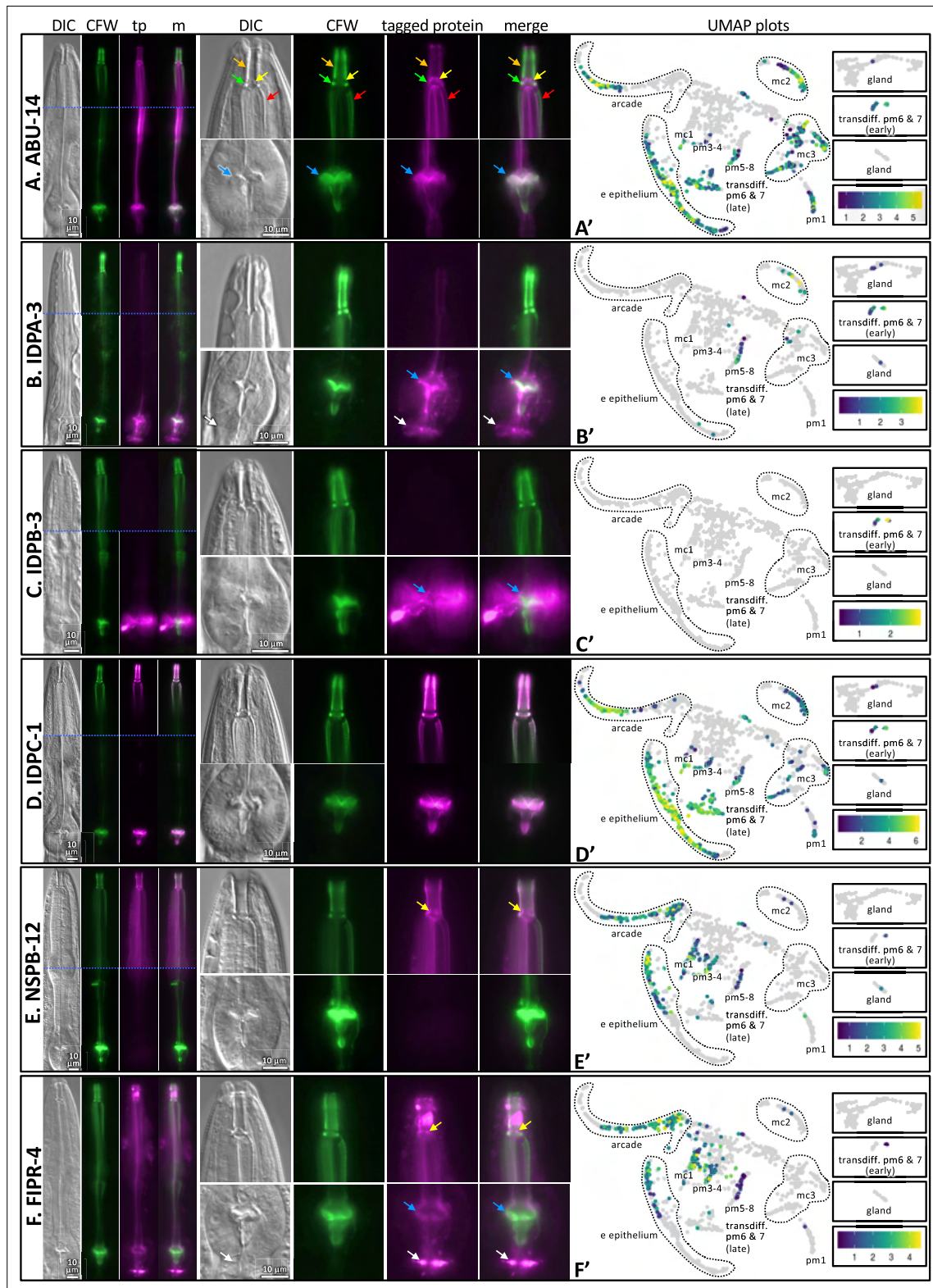


Figure 6. The localization of six fluorescently tagged pharynx cuticle components. Each of the six large horizontal boxes contain data about the six predicted gene products indicated on the left. In each box, the four images on the left are of the head of a single worm, imaged first with differential interference contrast (DIC), then with calcofluor white (CFW) in green, then the fluorescently tagged protein (tp) protein of interest in fuchsia, followed by a merged (m) image as indicated at the top of the columns. A blue horizontal line indicates the intersection of two cropped images to show

Figure 6 continued on next page

Figure 6 continued

different relevant focal planes of the same animal. The scale is indicated. The middle set of eight images correspond to magnified buccal cavity and channels (top) and terminal bulb and grinder (bottom). The scale is indicated. Colored arrows are used for reference in (A) and used to draw attention to particular features in (B–F): Orange, buccal cavity; yellow, flaps; green, collar; red, anterior channels; blue, grinder; white, presumptive ECM between the terminal bulb and intestinal valve. The graph on the right is a UMAP plot of the pharynx mRNA expression pattern for the respective gene (see text for details). The relative expression level is indicated. (A) An RP3439 animal harboring the *trls113[Pabu-14:abu-14:superfolderGFP; rol-6(d); unc-119(+)]* integrated array. (B) An RP3519 animal harboring the *Ex[idpa-3p::IDPA-3::mNeonGreen; myo-2p::mCherry]* extrachromosomal array. (C) An RP3498 animal harboring the *Ex[idpb-3p::IDPB-3::mNeonGreen; myo-2p::mCherry]* extrachromosomal array. (D) An RP3497 animal with genomic *idpc-1* fused in-frame to the coding sequence for mGreenLantern. (E) An RP3499 animal harboring the *Ex[nsfb-12p::NSPB-12::mNeonGreen; myo-2p::mCherry]* extrachromosomal array. (F) An RP3514 animal harboring the *Ex[fipr-4p::FIPR-4::mNeonGreen; myo-2p::mCherry]* extrachromosomal array. All animals are counterstained with the calcofluor white (CFW) chitin stain. The expression patterns shown are typical of the population that are positive for the transgene.

The online version of this article includes the following figure supplement(s) for figure 6:

Figure supplement 1. A comparison of the tagged ABU-14 and IDPC-1 localization patterns.

components exclusively, including that of the buccal cavity, flaps, and anterior channels (**Figure 6E**). Tagged FIPR-4 localized to both anterior and posterior pharynx cuticle components (but not the grinder teeth proper) and the presumptive pharynx-intestinal valve ECM (**Figure 6F**). Together, these analyses provide confidence in the predictive value of the spatiotemporal map.

The pharynx secretome is enriched in proteins with high predictions of phase separation

To better understand the types of proteins that are secreted by the pharynx, we manually curated the domain organization of all 367 oscillating pharynx-enriched gene products as reported by the WormBase, SMART, and PFAM protein databases (*Letunic and Bork, 2018; El-Gebali et al., 2019; Figure 4—source data 1*). We found that 106 of the 226 secreted proteins (47%) lacked any defined domain (last column of the chart in **Figure 4A, Figure 4—source data 1**). This prompted a systematic investigation of low-complexity sequence within the pharynx secretome using NCBI's SEG algorithm (*Wootton and Federhen, 1993*). Indeed, we found the pharynx secretome to be greatly enriched with low-complexity regions (LCRs) ($p=1E-69$) (**Figure 7A**). Given that low complexity is tightly associated with intrinsic disorder, we used the Spot-Disorder algorithm (*Hanson et al., 2017*) to systematically analyze whether the pharynx secretome is also enriched for IDRs and found that it is ($p=8E-10$) (**Figure 7B**).

Low-complexity intrinsically disordered protein regions often provide multivalency that can enable a protein to transition from being soluble to becoming a phase-separated liquid, gel, stable polymeric matrix, or an insoluble amyloid (*Muiznieks et al., 2018*). We explored the potential of the different protein sets to phase separate using three different predictive algorithms, including PSPredictor (*Chu et al., 2022*), PLAAC (*Lancaster et al., 2014*), and LLPhyScore (*Cai et al., 2022*). PLAAC was originally designed to scan for prion-like sequences, but has been retrospectively used as a reliable tool to predict phase separation (*Vernon and Forman-Kay, 2019*). Each algorithm reveals that the pharynx secretome is enriched in proteins with phase separation capability ($p=2E-46$, $p=2E-52$, and $p=2E-31$, respectively) (**Figure 7C–E**).

We also examined low-complexity, intrinsic disorder and phase-separation propensity as a function of developmental time. The peak molting hour corresponds to a clear peak in low-complexity and intrinsic disorder of secreted products (**Figure 7A' and B'**). The other three predictors also show significant peaks in phase separation propensity of secreted products during the peak molting hour, but variably show peaks at other time points as well (**Figure 7C'–E'**). To better understand the relative abundance of gene products with the specific sequence features highlighted in **Figure 7A'–E'**, we multiplied the trait value for each gene with the relative number of transcripts for each respective gene. In this light, we see a striking peak of all trends at the peak molting hour (**Figure 7A''–E''**). This analysis suggests that the pharyngeal cuticle is likely flooded with low-complexity, intrinsically disordered proteins with phase separation potential during the peak molting hour.

Finally, we tested these predictions by asking whether IDPC-2 can phase separate. Upon cleaving off the MBP affinity tag from the in vitro-expressed proteins, we see that IDPC-2 and the positive control FUS can form phase-separated droplets (**Figure 8A and B**). In these experiments, we use a

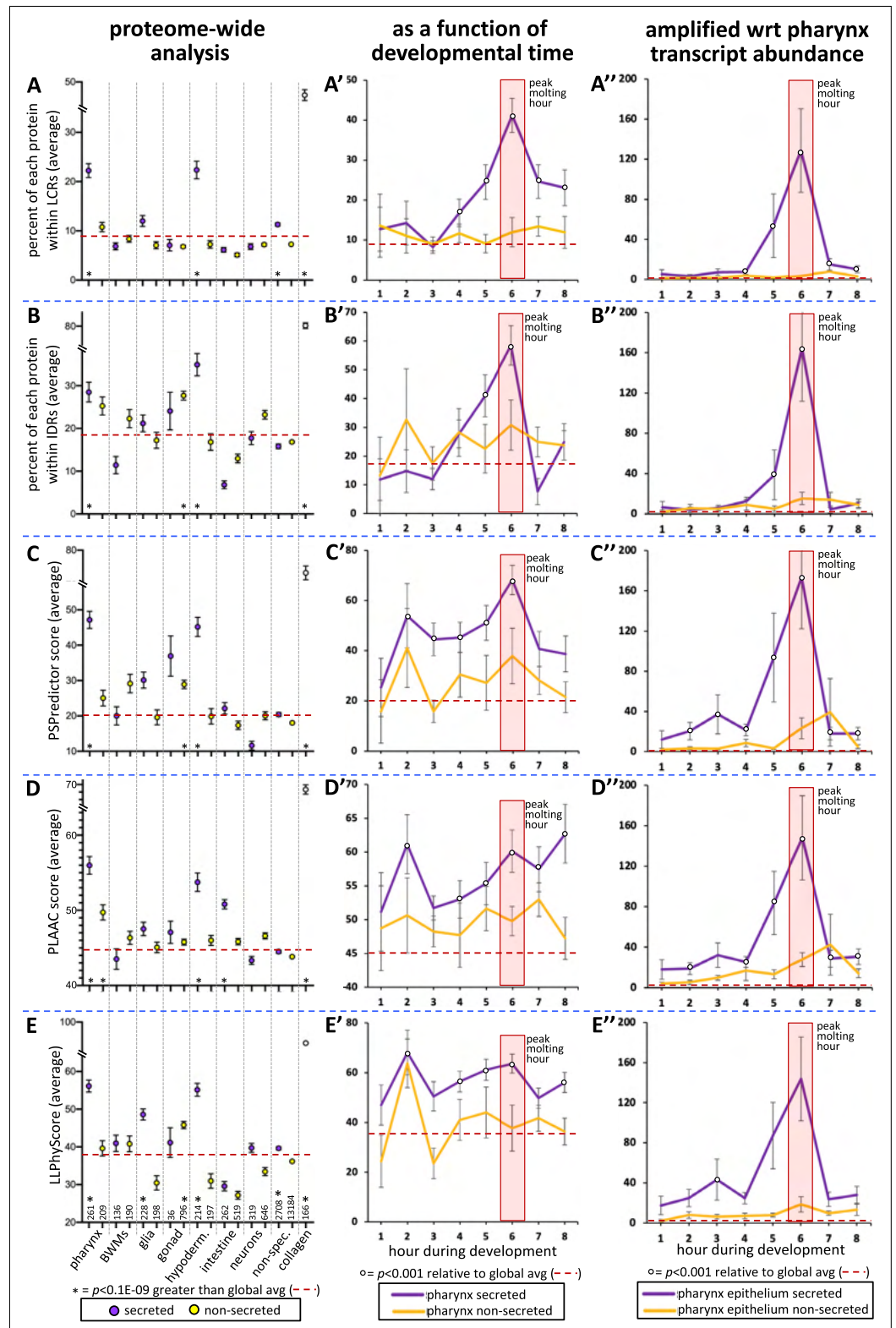


Figure 7. The pharynx secretome is enriched with intrinsically disordered proteins with phase separation capability. (A–E) An analysis of the entire proteome for the indicated properties. The tissue type examined, as well as the number of genes in each bin, is indicated at the bottom of the graph in (E) (hypoderm., hypodermis; non-spec., non-specific). Statistical differences, indicated with an asterisk at the bottom of each graph, were measured

Figure 7 continued on next page

Figure 7 continued

using a Student's *t*-test against the global average (indicated with a red hatched line for each property). (A'–E') An examination of the same properties as (A–E), but with a focus on genes whose expression is enriched in the pharynx over developmental time. (A''–E'') An examination of the same properties as (A'–E'), but normalized with respect to each gene's transcript abundance within the pharyngeal epithelium. For each gene, the number of transcripts was multiplied by the value of gene products property (i.e., % within low-complexity region [LCR], % within intrinsically disordered regions [IDRs], or PSPredictor score, etc.), and the average for that temporal bin was calculated. The Y-axis in (A''–E'') reports numbers in the thousands. Statistical differences were measured using a Student's *t*-test against the global average. In all graphs, standard error of the mean is shown. Because the PLAAC algorithm can report negative scores up to –60, 60 was added to the PLAAC scores of all gene products for the sake of clarity. The peak molting hour is highlighted by the transparent red box.

molecular crowding reagent (Ficoll) to mimic *in vivo* molecular crowding (André and Spruijt, 2020). These data support the informatic analyses that predict that many of the proteins incorporated into the cuticle may be capable of phase separation.

The pharynx secretome is not enriched with amyloidogenic proteins

We investigated the propensity of pharynx secretome proteins to form filaments. We first used the LARKS algorithm that predicts kinked β -structure, which can drive proto-filament assembly and reversible fiber formation (Hughes *et al.*, 2018). Indeed, we find a significant enrichment in LARKS scores within the pharynx secretome (Figure 8C). This prediction is corroborated by the LLPhyScore predictor of kinked β -structure (Figure 8D). We also investigated whether the pharynx secretome is enriched in amyloidogenic proteins. Both the Budapest (Keresztes *et al.*, 2021) and AmyloGram (Burdukiewicz *et al.*, 2017) machine-learning predictors, as well as the structure-based PATH predictor (Wojciechowski and Kotulska, 2020), fail to show any enrichment within the pharynx secretome of amyloidogenic proteins (Figure 8E–G).

We further probed the ability of the pharynx secretome to form amyloid fibers using CR dye. CR has long been used as a diagnostic tool to identify rigid amyloid fibrils because of its special property of emitting apple green birefringence upon binding the ordered fibril array in the presence of polarized white light (Divry, M, 1927). This is in sharp contrast to the colorless birefringence of the crystalizing compounds (Figure 8H). While CR specifically stains the pharynx cuticle, we found that CR-stained cuticles do not emit apple green birefringence ($n > 30$) (Figure 8I and J). We are confident that our imaging system is capable of detecting CR-derived apple green birefringence because of a serendipitous observation. We found that when CR is co-incubated with a small molecule (called wact-190) that forms crystals in the pharyngeal cuticle (Kamal *et al.*, 2019), the resulting crystals exhibit apple green birefringence (Figure 8K). We infer that this happens because CR likely becomes incorporated into a regular array, that is, the wact-190 crystal. Together, these results indicate that it is unlikely that the cuticle harbors rigid amyloid fibrils, which is consistent with both the flexible nature of the pharynx cuticle (Huang *et al.*, 2008; Avery, 1993) and the absence of any detectable amyloid-like fibers in previous transmission electron micrographs of the pharynx cuticle (Wright and Thomson, 1981; White *et al.*, 1986). We conclude that the pharynx secretome is likely enriched in proteins with intrinsic disorder, phase separation capability, and proto-filament formation capability, but not enriched with proteins that form rigid amyloid fibrils.

The transcripts encoding secreted IDR protein families peak in expression in overlapping waves during cuticle construction

Given the enrichment in low-complexity sequence within the pharynx secretome, we were curious to know whether it has any global bias in amino acid residue distribution relative to other protein sets. We found a significant enrichment of nine residues with a strong bias against charged and hydrophobic residues (at least $p < 2E-05$; Figure 9A). Upon considering relative abundance of amino acid residues as a function of time, we see that proteins rich in cysteine, proline, and glutamine peak in expression during new cuticle construction (Figure 9B).

We used the Clustal Omega clustering tool (Sievers *et al.*, 2011) to determine whether there were families of proteins with similar sequence within the 106 proteins that lacked domains within the pharynx secretome. We found six distinct families of low-complexity proteins through this analysis

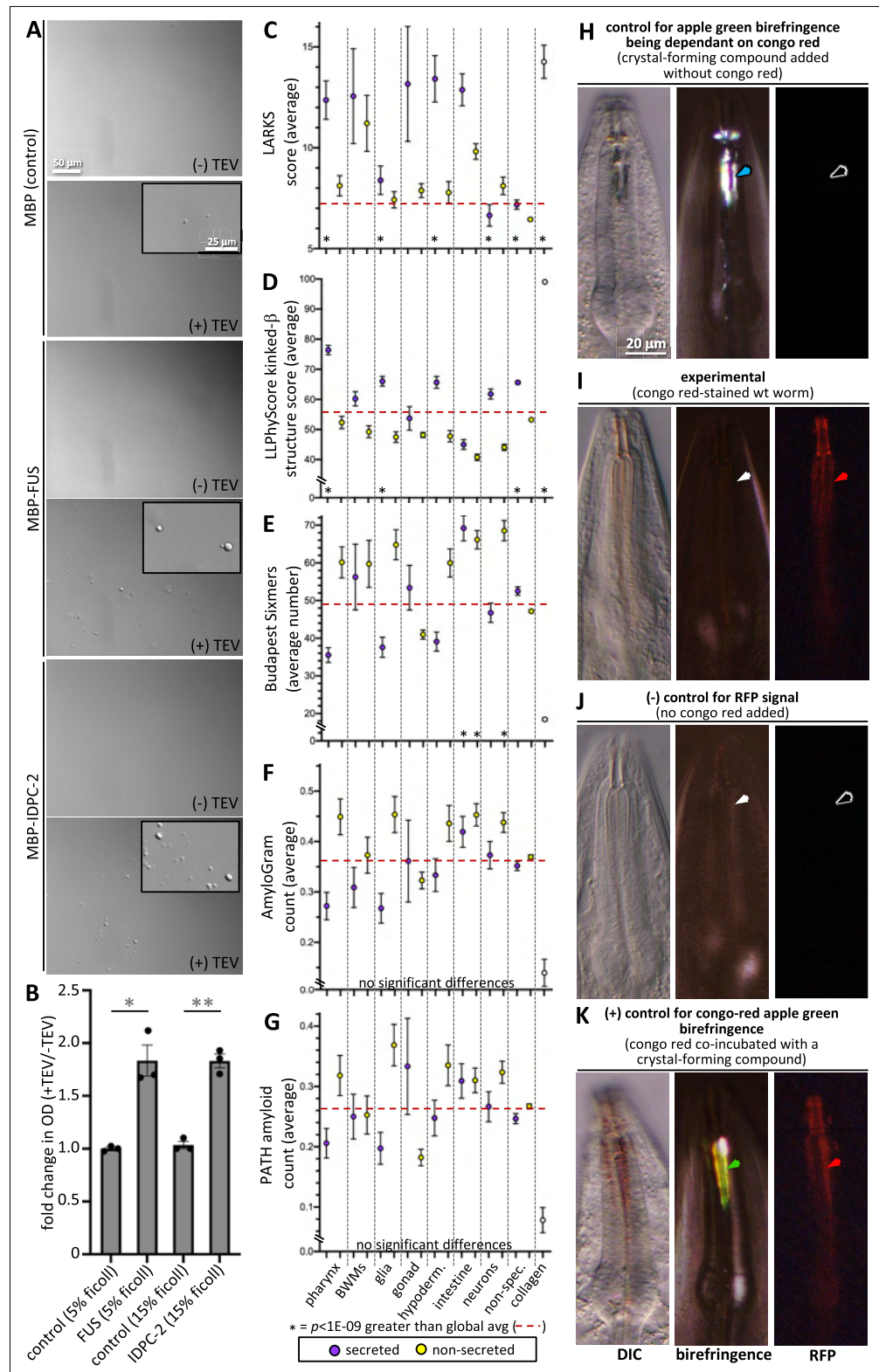


Figure 8. Cuticle proteins can likely phase separate and are enriched with protofilament but not amyloidogenic sequence. **(A)** In vitro purified maltose-binding protein (MBP) control (15% Ficoll) or fusions with the FUS-positive control (5% Ficoll) and IDPC-2 (15% Ficoll) phase separate into spheres upon cleaving off the MBP tag with TEV protease, while the MBP-only negative control does not. The inset is a magnification of the corresponding area.

Figure 8 continued on next page

Figure 8 continued

The scales for all insets and larger images are respectively the same. **(B)** Quantification of the fold change in optical density (OD; 395 nm) of indicated samples after 1 hr of treatment with TEV relative to the OD without the addition of TEV. * $p < 0.01$ and ** $p < 0.001$, respectively, using a Student's *t*-test. In **(A, B)**, all proteins are at a concentration of 1 mg/mL, except for FUS, which is at 1.5 mg/mL. The MBP-only control is therefore a vast molar excess. **(C–G)** An analysis of the entire proteome for the indicated properties. The details are the same as that indicated for **Figure 7**. **(H)** Control for the dependence of the apple green color on Congo Red (CR). Wildtype animals are incubated in wact-190 as previously described (*Kamal et al., 2019*) and yield birefringent crystals that lack notable apple green color (blue arrowhead). **(I)** Wildtype adult worms incubated with CR exhibit red fluorescent pharyngeal cuticle (red arrowhead; left column), but no apple green birefringence (white arrowhead; middle column). Differential interference contrast (DIC) is shown in the left column. Zero out of 30 animals exhibited apple green birefringence. **(J)** Control for the CR RFP signal. Wildtype animals are incubated without CR present. No birefringence (white arrowhead) or CR signal (black arrowhead) results. **(K)** Control for the ability to detect CR apple green birefringence. The wildtype animal was incubated simultaneously in CR and wact-190, a small molecule that crystalizes in the pharyngeal cuticle. The apple green birefringence (green arrowhead) manifests under these conditions because CR likely incorporates into the regular crystal lattice of the wact-190-derived crystals. The scale in **(H)** is representative of all panels.

(*Figure 9C, Figure 9—source data 1*). Members of each family share an enrichment of particular residues (*Figure 9D*), contain regions of high percentage positional sequence identity (*Figure 9E, Figure 9—figure supplement 1*), and are expressed at similar times as one another (*Figures 4A and 9F*). These six families include three new families of IDR-rich proteins, which we have named IDPA, IDPB, and IDPC, a subgroup of APPGs (*George-Raizen et al., 2014; Figure 9E*), and the relatively short NSPBs and FIPRs about which little is known. See *Supplementary file 1* for all newly named genes presented in this study and *Supplementary file 2* for all members of the six families described here. Systematic searches relying on positional alignment reveal no obvious homologs of these six families in any group beyond Nematoda (WormBase). Furthermore, a comparison of the consensus sequence from these families (*Figure 9E, Figure 9—figure supplement 1*) to the cuticle proteins of other Ecdysozoans (*Willis, 2010*) reveals no obvious similarity in the pattern or amino acid sequence biases.

The transcription of the six families of low-complexity proteins peaks in expression in successive overlapping waves, with five of the waves concentrated around the peak molting hour (*Figure 9F*). The combined use of the three different predictors of phase separation suggests that the IDPAs, IDPBs, IDPCs, and the APPGs may be able to phase separate (*Figure 9F*). The FIPRs and NSPBs are also likely to phase separate but fail to score high with the SpotDisorder algorithm because of their small size. The IDPAs and IDPBs are predicted to form protofilaments (as measured by LARKS), the IDPAs and APPGs score especially high with the prion sequence evaluator (PLAAC), and five members of the APPGs (ABU-6, ABU-7, ABU-8, ABU-15, and PQN-54) are predicted to be amyloidogenic (as measured by AmyloGram and PATH) (*Figure 9F*). These results further support the idea that a large proportion of the proteins secreted by the pharynx during cuticle construction are IDR-rich with phase-separating capability.

Epithelial and transdifferentiated cells secrete abundant products during the molt

We sought increased spatial resolution of peak gene expression that is associated with pharyngeal cuticle construction over the course of the temporal map. We therefore returned to the Cao et al. single-cell sequencing dataset (*Cao et al., 2017; Packer et al., 2019*) to systematically visualize the expression patterns of pharynx secretome components. *Cao et al., 2017* and *Packer et al., 2019* identified 1675 sequenced cells that belong to the pharynx. When grouped according to similar expression profiles, the pharynx cells form subclusters on a Uniform Manifold Approximation and Projection (UMAP) created by Packer et al. (see https://cello.shinyapps.io/celegans_L2/) that represent cells of a similar type (*Packer et al., 2019; Figure 10A*). Based on the expression of some characterized reporter transgenes and their single-cell sequence analysis of the embryo, Packer et al. made tentative cell assignments for most subclusters of the L2 pharynx (see Supplemental Table 12 in *Packer et al., 2019*).

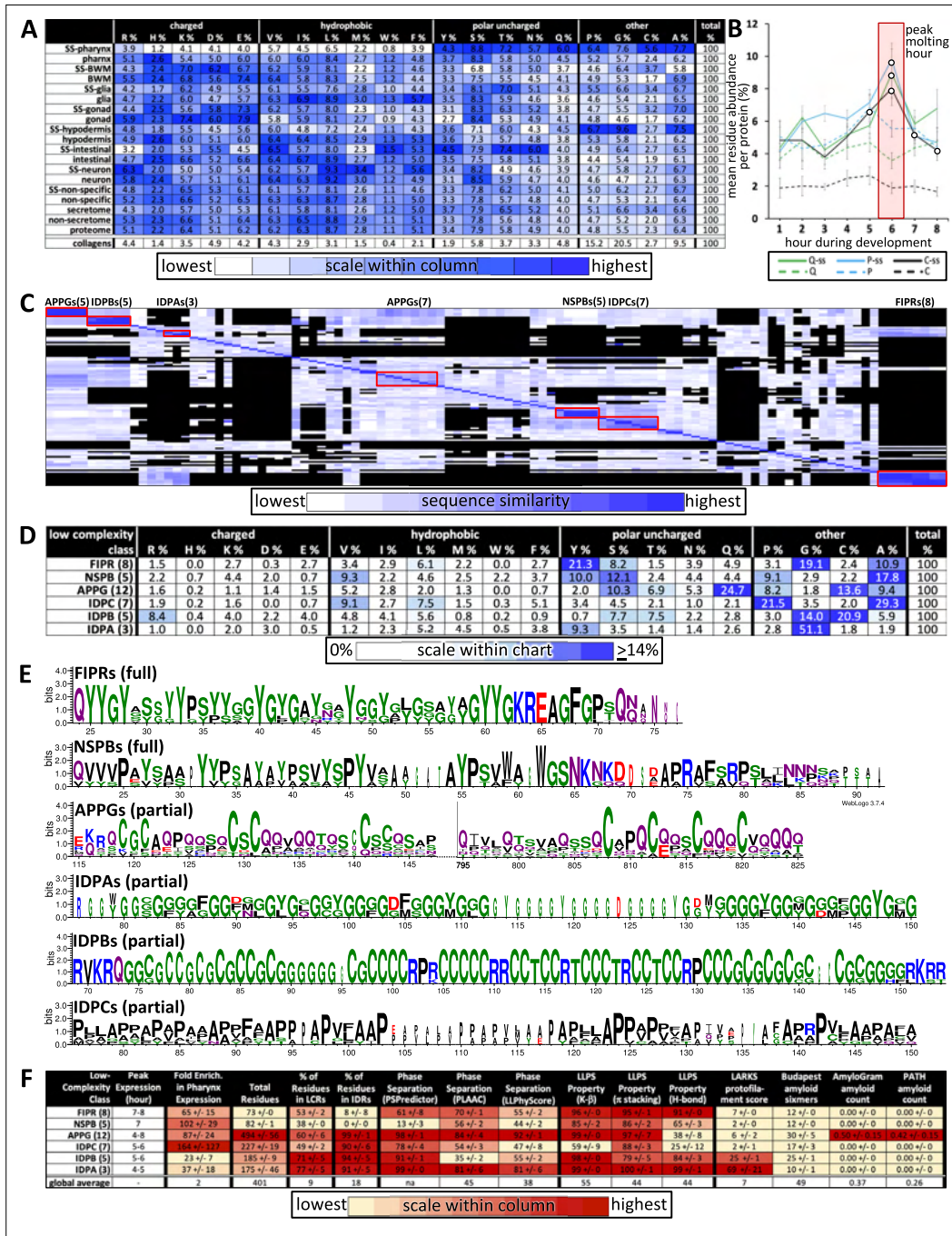


Figure 9. Properties of the low-complexity protein families that are likely secreted into the developing cuticle. (A) Average percent amino acid composition of the proteins within the indicated tissue type. The percentages along a single row sum to 100. The color scale indicates the range of values within a single column so as to compare the relative abundance of the indicated residue among the different protein sets. The collagens are not included in the color scale comparison. SS, secreted proteins based on harboring a signal sequence; BWM, body wall muscles. All of the mean residue percentages from the set of proteins secreted from the pharynx cells are significantly different compared to that of the remaining proteome (Student's t-test; $p < 2E-05$). (B) A plot of the average percentage cysteine, proline, and glutamine composition of each protein as a function of developmental time. Secreted (ss) and non-secreted proteins are represented by solid lines and dashed lines, respectively. Open circles indicate significant differences relative to the non-secreted class ($p < 0.05$). (C) Clustal Omega pairwise comparisons of all 106 low-complexity proteins in the pharynx secretome. Both X and Y axis have the same 106 proteins in the same order. Families with high sequence identity are outlined with a red box. (D) Similar to (A), except that residue composition is restricted to the indicated low-complexity family and that the color scale compares percentages across the entire chart. (E) Consensus sequence logos for the indicated protein families. The full consensus sequence (without the signal peptide) of the FIPRs and NSPBs is

Figure 9 continued on next page

Figure 9 continued

shown. The full consensus sequence of the remaining groups is given in Figure S5. (F) A chart of properties for the six low-complexity families. Because the PLAAC algorithm can report negative scores up to -60 , 60 was added to the PLAAC scores of all gene products for the sake of clarity. All values show means \pm standard error of the mean.

The online version of this article includes the following source data and figure supplement(s) for figure 9:

Source data 1. Supporting information for the chart diagram in **Figure 9C**.

Figure supplement 1. Consensus sequence for select low-complexity families.

We searched the literature for additional GFP reporter transgenes that are expressed in the postembryonic pharynx to help refine the identities of many of the L2 pharynx subclusters (**Figure 10—figure supplements 1 and 2**). We then transformed the Cao and Packer et al. L2 pharynx subcluster data into transcript summaries (see 'Materials and methods') and examined the expression level of oscillating pharynx-enriched transcripts in each of the subclusters (**Figure 10B and C**; see **Figure 1** for the relative location of each cell type).

During hours 3, 4, and 5, abundant products are secreted by the e epithelial cells, the mc3 marginal cells and presumptive pm6 and 7 transdifferentiated cells (see below). The identity of these transcripts (see **Figure 4** and **Figure 4—source data 1**) suggests that the cells are accumulating stores for the catabolism of the old cuticle and construction of the new one at the onset of the molt. Despite being confident in our assignment of cluster 11 as pm1 (**Figure 10—figure supplements 1 and 2**), the expression profile of cluster 11 is more like the arcade, e epithelial cells, and mc3 marginal cells than muscle, suggesting that pm1 may also play a role in the catabolism of the old cuticle. This is consistent with the correlation between the pharynx UMAP plot for ABU-14 and what we observe in animals with fluorescently tagged ABU-14 (**Figure 6A and A'**).

During hours 5 and 6 (which is the peak molting hour), the arcade and e epithelial cells produce abundant secreted components, consistent with the construction of a new buccal cuticle (**Figure 10B and D**). The mc1 and mc2 marginal cells also secrete abundant product (**Figure 10B and E**), again consistent with the construction of the channel cuticles and sieve (see **Figures 1, 6A and A'**).

Conspicuously absent from the expression profiles of confidently assigned subclusters is abundant secretion from the cells that surround the grinder in the posterior bulb (i.e., pm6 and pm7). Subcluster 22, which is confidently identified as pm5, pm6, pm7, and pm8 muscle, express only low levels of secreted proteins during the peak molting hour. Previous work has shown that the pm6 and pm7 cells transdifferentiate from muscle into highly secretory cells during the molting period to build a larger grinder (**Sparacio et al., 2020**). Based on the expression of a combination of markers (**Figure 10—figure supplements 1 and 2**) and the abundant expression of secreted products, we infer that subclusters 1 and 5 represent transdifferentiated pm6 and pm7 that secrete many of the same components used in the anterior pharynx epithelia to build the grinder (**Figure 10B and F**). We find that the IDPAs and IDPBs are expressed in the early transdifferentiating pm6 and pm7 cells (**Figure 10B** and **Supplementary file 2**), and therefore likely contribute to grinder formation. This prediction is consistent with our finding that disruption of IDPA-3, which localizes to the grinder (**Figure 6B and B'**), results in obvious grinder defects (**Figure 5D**). This prediction is also supported by the exclusive localization of tagged IDPB-3 to the grinder and pm6 cells (**Figure 6C and C'**). Finally, *idpb-1* and *idpp-3* are two genes belonging to subcluster 1 (**Figure 10B**, hours 4 and 5) and Yuji Kohara's mRNA in situ expression database reveals robust and specific expression of these two genes in only the posterior bulb cells (**Motohashi et al., 2006; Supplementary file 1**). Together, these observations are consistent with the assignment of subclusters 1 and 5 to the transdifferentiating pm6 and pm7 cells.

During the peak molting hour 6, IDPCs and the APPGs are expressed in most cells that contribute to the pharyngeal cuticle. Again, Kohara's mRNA in situ database confirms this interpretation with robust and specific pharynx expression patterns for *abu-6*, *abu-14*, *appg-2*, *idpc-1*, *idpc-3*, and *idpc-5*, and *pqn-13* (**Supplementary file 1**). The localization of tagged ABU-14 and IDPC-1 also supports this conclusion (**Figure 6A, A', D and D'**).

During hours 7 and 8, NSPB and FIPR expression is more restricted to the arcade, e epithelial cells, and the mc1 cells (**Figure 10B** and **Supplementary file 2**). Tagged NSPB-12 supports this prediction (**Figure 6E and E'**). Tagged FIPR-4, while localizing to the anterior cuticle, is also present in the posterior cuticle, suggesting that secreted FIPR-4 may be able to diffuse extensively (**Figure 6F and F'**).

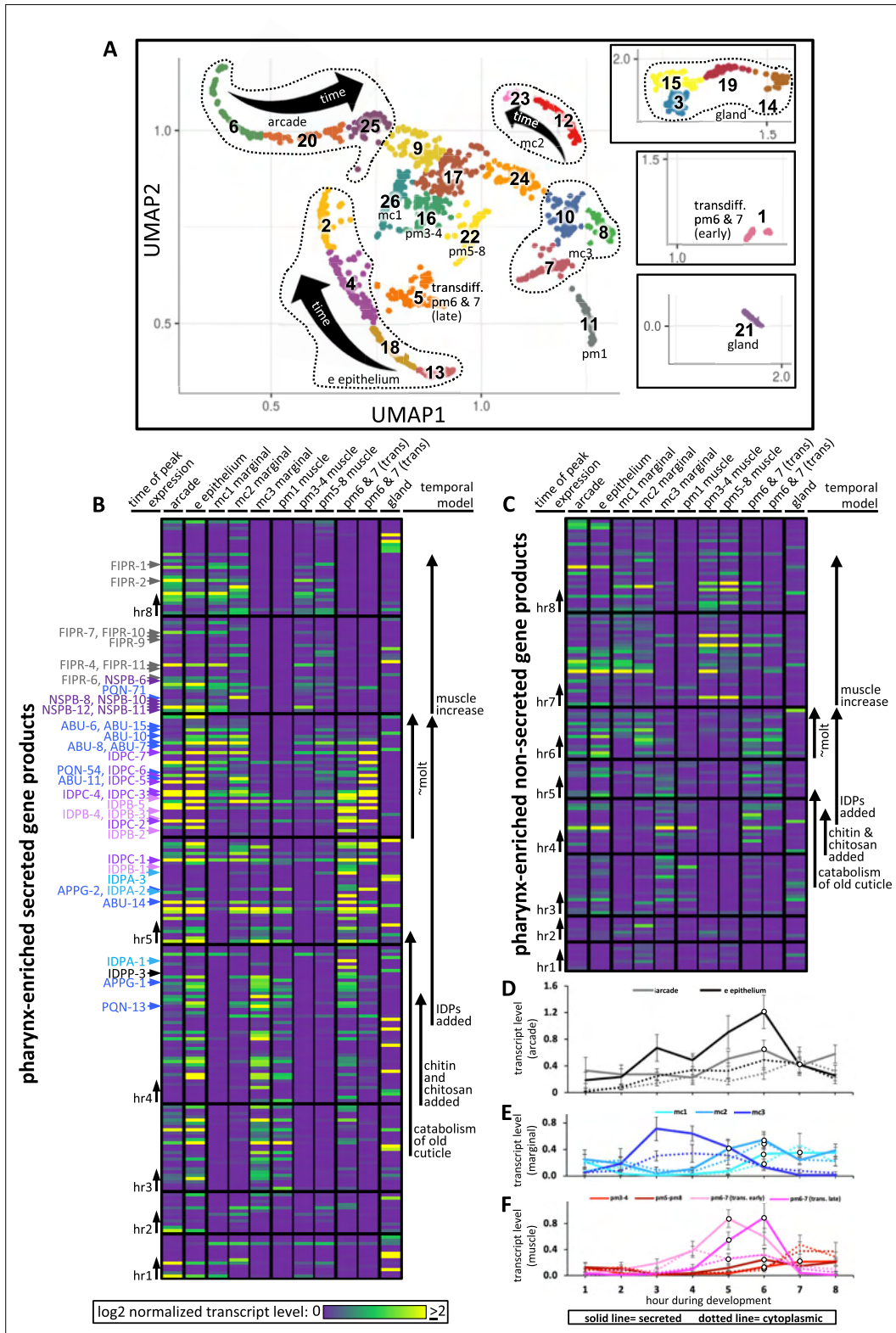


Figure 10. Expression of pharynx-enriched genes in distinct cell types. (A) A UMAP of 1675 pharynx cells modified with permission from *Packer et al., 2019*'s online tool. The clusters are numbered according to *Packer et al., 2019*. The cell type identities are partially based on those from *Packer et al., 2019* (see *Figure 10—figure supplements 1 and 2* for details). Due to space constraints, three cluster groups from the map are shown as insets. (B, C) The expression level of the pharynx-enriched gene set in the indicated tissue type. The graph notation and the order of genes in rows is preserved *Figure 10 continued on next page*

Figure 10 continued

from **Figure 4A**. Genes encoding a signal peptide are shown in **(B)** and those without a signal sequence are shown in **(C)**. The mc1, pm3-4, and pm5-8 values represent the average gene expression of the cells within the respective clusters (26, 16, and 22). The values corresponding to the other cell types represent the highest average from among the group of clusters that constitute that cell type. For example, the arcade cells are represented by clusters 6, 20, and 25, but the expression level from each of these clusters is distinguished by time, not space, and averaging signal from all three would dilute the expression level that represents that cell type. All members of the six low-complexity families are indicated on the left of **(B)** and the color code is the same as that present in **Figure 4A**. **(D–F)** The average transcript level of all genes within the indicated cell type as a function of binned time. Open white circles represent a significantly greater value ($p < 0.01$) compared to the bin 2 hr previous.

The online version of this article includes the following figure supplement(s) for figure 10:

Figure supplement 1. Identity assignment of the pharynx UMAP clusters.

Figure supplement 2. UMAP plots of the gold standard genes used to assign identity to the pharynx UMAP reference cluster.

Cytoplasmic components involved in muscle development peak in expression during hours 7 and 8 (**Figure 10C and F**).

The number of genes expressed from the gland cells is not obviously enriched in any one temporal interval (**Figure 10B and C**), yet the overall abundance of gland transcripts peak in hour 5 (**Figure 4D**). This apparent contradiction is due to the two most abundantly expressed genes from the gland, *phat-2* and *phat-4*, peaking in expression during hour 5 (**Figure 10B**, **Figure 4—source data 1**). PHAT-2 and PHAT-4 are paralogous mucin-like proteins (*Ghai et al., 2012*; *Smit et al., 2008*) whose timing of peak expression suggests that they may play a role in cuticle structure or function. PHAT-2 and PHAT-4 notwithstanding, the overall temporal pattern of expression from the gland suggests that its products do not play a large role in cuticle turnover during the molt.

Discussion

A model of pharyngeal cuticle construction

Here, we have mined published resources to bioinformatically reconstruct the *C. elegans* pharynx cuticle. This map provides unprecedented insight into the spatiotemporal progression of cuticle construction. During hours 3 and 4, genes that encode homologs of chitin and amyloid catabolic enzymes peak in their expression. These include the predicted chitinases CHT-1, CHT-2, CHT-5, CHT-6, two predicted amyloid peptidases (NEP-1 and NEP-12) (*Iwata et al., 2001*), and the NAS-6 protease that helps degrade pharyngeal cuticle (*Sparacio et al., 2020*; *Park et al., 2010*). The predicted amyloid-fibril inhibitor ITM-2 (*Cohen et al., 2015*) also peaks in expression during this interval, perhaps to prevent aggregation during disassembly. The expression profile at this interval is consistent with preparation for apolysis (the detachment of the old cuticle).

During hours 4, 5, and 6, anabolic enzymes and constructive components peak in expression. These include the characterized chitin synthase CHS-2 (*Zhang et al., 2005*), putative chitosan synthases LGX-1 and CHTS-1 that deacetylates chitin to produce chitosan (*Heustis et al., 2012*), and putative chitin binders and cross-linkers CHTB-1, CHTB-2, and CHTB-3. In this interval, components implicated in amyloid metabolism also peak in expression. These include a predicted amyloid chaperone LRX-1 (*Cam et al., 2004*), two predicted amyloid-chitin linkers LRPC-1 and PQN-74 (*Brodeur et al., 2012*), and a predicted amyloid precursor protein interactor FEH-1 (*McLoughlin and Miller, 2008*).

During hours 5 and 6, a massive increase in gene expression of the pharynx secretome occurs. The period coincides with the upregulation of secreted intrinsically disordered proteins from the pharynx epithelium and includes successive waves of peak transcript expression encoding four of the intrinsically disordered families, IDPA, IDPB, IDPC, and APPG members that have been previously implicated in cuticle development (*George-Raizen et al., 2014*).

During hours 5 and 6, the gene products that peak in expression are rich in PPIs compared to the proteins secreted by other tissues. The protein interactors within the pharynx secretome network are highly enriched in low-complexity sequences predicted to phase separate.

Finally, during hours 7 and 8, genes that encode muscle contraction components are upregulated, which likely corresponds to a period of tissue growth at the tail end of molting. We also see the peak expression of the low-complexity families NSPB and FIPR, which are likely added to the cuticle in its final phase of maturation. Together, these observations illustrate the utility of the spatiotemporal map in revealing the logic by which a cuticle is assembled.

The pharynx cuticle is unlikely to harbor amyloid fibrils

Despite the pharynx secretome not being enriched for amyloidogenic proteins, multiple pharynx cuticle proteins are predicted to nevertheless be amyloidogenic. In addition, multiple predicted amyloid regulators are upregulated during pharyngeal cuticle development. Yet, evidence argues against the presence of amyloid fibrils within the pharynx cuticle. We speculate that fibril formation may not occur within the pharyngeal cuticle because of the heterogeneous mixture of the IDR-rich proteins within the structure. In other words, the relatively low concentration of any one protein species within the cuticle mixture may preclude the assembly of long fibrils with birefringent properties. Indeed, the presence of other IDRs antagonizes A β 42 fibril formation (Ikeda et al., 2020). A second factor that may antagonize fibril formation is the presence of a chitin matrix. During the formation of the squid beak, IDR-rich proteins form phase-separated coacervates that infiltrate a chitin matrix (Tan et al., 2015), which may limit amyloid fibril formation. It is unknown whether similar dynamics take place during pharyngeal cuticle development. Third, the pharynx secretome is enriched with kinked β -structure that can support liquid-phase separation and may facilitate protofilament formation but otherwise antagonizes extensive fibril growth (Hughes et al., 2018). Notably, many well-characterized proteins with amyloidogenic propensity only form fibrils when associated with pathogenesis (Patel et al., 2015; Cremades et al., 2012).

The idea that the pharyngeal cuticle contains a non-rigid network of IDRs is appealing because the pharyngeal cuticle must be sufficiently flexible to accommodate pharynx movements along the dorsal–ventral (Huang et al., 2008) and anterior–posterior (Avery, 1993) axes. Indeed, others have suggested that IDR-rich proteins within chitin-based cuticles might add elastic properties to what might otherwise be an inflexible chitin-based material (Andersen, 2011). An elastic cuticle might also aid in returning the open and extended lumen (which results from pharynx muscle contraction) to the relaxed ground state position.

Potential contributions of IDPs to the cycles of cuticle formation and destruction

A key feature of phase-separating IDRs is their potential to reversibly transition between different states of matter depending on local conditions and post-translational modifications (Murray et al., 2017; Deiana et al., 2019), including liquids and gel-like biomaterials. The pharyngeal cuticle must soften, be shed, and be reconstructed about every 8 hr during larval development (Lazetic and Fay, 2017). The notion that a network of IDR-rich proteins is not locked into a rigid state but may instead be regulated to increase or decrease intermolecular interactions and change material properties as needed during the molting cycle is an appealing idea that requires further investigation.

Both the APPGs and the IDPBs are highly enriched with cysteines and contribute heavily to an increase in the relative abundance of cysteines that is likely deposited into the developing cuticle as the animal prepares to molt. Other work has shown that the *C. elegans* cuticle is indeed rich in disulfides during the intermolt period and becomes reduced to facilitate apolysis (Stenvall et al., 2011). Furthermore, exogenously supplied reducing agent can induce pharyngeal cuticle apolysis during the intermolt period (Stenvall et al., 2011). Manipulating the redox state of cysteines can alter the ability of IDR-rich proteins to phase separate or further condense (Reed and Hammer, 2018; Zhang et al., 2020; Kato et al., 2019). Whether the abundant cysteines within the pharyngeal cuticle are key to phase separation and yield a network of variably dynamic cross-linked proteins remains to be determined.

The spatiotemporal map suggests that many different types of IDPs likely contribute to the pharyngeal cuticle. Previous studies have shown that coexisting condensed protein phases, each with distinct protein compositions, can yield complex biomaterials with layers and other non-uniform properties (Mountain and Keating, 2020; Lu and Spruijt, 2020; Lin et al., 2018). The distinct compositions of the six families uncovered by the spatiotemporal map are suggestive of the potential immiscibility of their condensed phases and of physical mechanisms for building the cuticle, particularly when combined with varying temporal expression, similar to what is observed during cuticle formation of the mussel byssus (Jehle et al., 2020). What is becoming clearer is how evolution has repeatedly capitalized on biomolecular condensates to make complex protective structures.

The molecular composition of cuticles may be evolutionarily plastic

The extent to which the blueprint of *C. elegans* pharyngeal cuticle development is conserved among other phyla within Ecdysozoa is unknown. The incorporation of chitin and chitosan within Ecdysozoan cuticles is firmly established (Moussian, 2010; Muthukrishnan et al., 2019). Mounting evidence also indicates that the arthropod cuticle has abundant IDR-rich proteins (Andersen, 2011) with amyloid-like folds (Sviben et al., 2020). However, of the 12 families of known arthropod cuticle proteins, only CPAP1 and CPAP3 have recognizable conservation with nematodes (Willis, 2010; Muthukrishnan et al., 2019). CPAP1/3 are defined by the ChtBD2 chitin-binding domain that is also harbored in the pharyngeal cuticle proteins CHTB-2, LRPC-1, and PQN-74. CPR is the only other arthropod cuticle family protein beyond the CPAPs that is well-characterized to bind chitin; the function of the remaining families remains obscure (Willis, 2010; Muthukrishnan et al., 2019). Furthermore, homologs of the six low-complexity families found within the pharyngeal cuticle cannot be found beyond Nematoda. It is not clear whether the IDR-rich proteins of arthropod and nematode cuticles are of distinct evolutionary origin or have simply diverged beyond recognition because of reduced primary sequence constraints. Regardless, the IDR-chitin combination clearly provides an effective barrier that is evolutionarily malleable to provide diverse form for millions of species.

The spatiotemporal map is a foundation for future investigation

The spatiotemporal map provides a starting point to investigate many important questions. First, what is the mechanism by which the temporal unfurling of gene expression is coordinated? While the global oscillatory pattern of *C. elegans* gene expression has been modeled in detail (Meeuse et al., 2020; Hutchison et al., 2020), how the oscillatory pattern of each gene becomes temporally offset from other oscillating genes is not understood. One candidate regulator of oscillation is the *C. elegans* period ortholog LIN-42. LIN-42 is a known regulator of developmental timing in the worm (Jeon et al., 1999; McCulloch and Rougvie, 2014), is expressed in the pharynx and other tissues (Monsalve et al., 2011), and alters the timing of molting when disrupted (Monsalve et al., 2011). Temporally uncoordinated gene expression would almost certainly be lethal, yet *lin-42* null mutants are viable (Edelman et al., 2016), suggesting that other key regulators are involved. Investigating the relationship between tissue-restricted transcription factors and their targets as a function of developmental time may provide insight into the coordinated temporal regulation of gene expression (Roy, 2022).

Second, how are catabolic and anabolic processes separated and regulated? The process of molting leaves animals vulnerable and must occur rapidly. In that light, it is perhaps not surprising that we observe a temporal overlap of expression of catabolic and anabolic components. Previous work on the ultrastructure of the grinder cuticle and molt indicates that dense core vesicles (DCVs) lie in wait until the new cuticle is assembled, at which point the DCVs likely fuse with the plasma membrane and dump their contents (Sparacio et al., 2020). Based on the timing of the peak expression of secreted components with respect to the timing of the molt itself, we surmise that (1) there is a temporal lag between the period of peak expression for a given gene and when protein abundance peaks, and (2) unknown mechanisms regulate the timing at which catabolic and anabolic components, perhaps within distinct DCVs, are released into the ECM. In this way, it might be possible to have temporal overlap in the peak expression of genes that encode catabolic and anabolic components. Exactly how the secretion of catabolic and anabolic components is regulated remains to be determined.

Finally, how are patterns within the pharyngeal cuticle established? Cuticle lumen shape and size are likely patterned by the underlying cells, but this simply extends the question. How is the patterning of the electron-dense cuticle ribbing established? Is the information that governs pattern of the flaps, which is seemingly independent of the shape of nearby cells, contained within the flaps' protein components? Do the successive waves of expression of low-complexity protein families contribute to the layering of the cuticle seen in the electron micrograph cross sections? How might coexisting condensed phases of these proteins establish layering and other complexities of the cuticle structure? The spatiotemporal map of pharyngeal cuticle construction presented here may serve as the foundation for answering these and other questions in the future.

Materials and methods

Methods

C. elegans culture, microscopy, and synchronization

C. elegans strains were cultured as previously described (Kamal et al., 2019). Unless otherwise noted, the wildtype N2 Bristol strain was used. Worms are prepared for imaging by washing them three times in M9 buffer and resuspended in a paralytic solution of either 50 mM levamisole or 50 mM sodium azide. The resuspended worms are then mounted on a 3% agarose pad on a glass slide and a coverslip for all brightfield and fluorescent microscopic analyses and photography. Unless otherwise noted, a Leica DMRA compound microscope with a Qimaging Retiga 1300 monochrome camera was used for routine analyses. Confocal imaging was performed using the Zeiss LSM 880 attached to an inverted epifluorescent microscope with a $\times 63$ (numerical aperture 1.4) oil immersion objective. Worms expressing GFP were excited using an argon laser operating at 488 nm. Confocal images were obtained using digital detectors with an observation window of 490–607 nm (green). Pseudo-transmission images were obtained by illuminating with the 488 nm laser and detected with the transmission photomultiplier tube and converted to digital images. Birefringent analyses were done with the Leica DMRA with the polarizer and analyzer polarized filters at right angles to one another. Colored birefringence images were captured using a Leica Flexacam C1 colour camera.

Synchronized populations of worms were obtained by first washing off a population of worms rich with gravid adults on plates with M9 buffer, collecting the sample in 15 mL conical tubes, and centrifuging the samples at $800 \times g$ to concentrate worms. The supernatant is then removed via aspiration and additional washes with M9 buffer are done until all bacteria are removed. 1.5 mL of suspended worms are then left in each tube and in rapid succession, 1 mL of 10% hypochlorite solution (Sigma) is added followed by 2.5 mL of 1 M sodium hydroxide solution and 1 mL double-distilled water. The mixture is incubated on a nutator for ~ 3.5 min. The tubes are then vortexed for 10 s with two 5 s pulses and visually inspected for near-complete digestion of post-embryonic worms. M9 buffer is then added to 12 mL. The tube is spun at 2000 rpm for 1 min, supernatant removed, fresh M9 buffer added to ~ 12 mL, and the tube is vigorously shaken. This is repeated two more times. After the final wash, the tube is incubated overnight on a nutator at 20°C to allow egg-hatching. The next day, the sample is checked for synchronized L1s. To obtain other synchronized stages, the synchronize L1s are plated on solid agar substrate with *Escherichia coli* food and allowed to progress to the desired stage before processing.

C. elegans transgenes

NQ824 *qnEx443[Pabu-14:abu-14:sfGFP; rol-6(d); unc-119(+)]* was a kind gift from David Raizen. We chromosomally integrated the *qnEx443* extra-chromosomal array using previously described methodology (Mello and Fire, 1995), resulting in the RP3439 *trIs113[Pabu-14:abu-14:sfGFP; rol-6(d); unc-119(+)]* strain. Tagged IDPC-1 was generated by InVivoBiosystems (Eugene, USA) by using CRISPR/Cas9-based mGreenLantern knock-in at the C-terminus of the Y47D3B.6 native locus. Two guide RNAs, sgRNA1 (5'-AGCTCCTGGGACACAGGCTG-3') and sgRNA2 (5'-GCTGGAGTCTGCCAGTGCGC-3'), were designed to target the C-terminus of Y47D3B.6. The single-stranded donor homology DNA included 35 bp homology arms flanking a GGGSGGG linker and the mGreenLantern sequence. Insertion of the mGreenLantern sequence was identified by PCR and confirmed by sequencing.

IDPA-3, IDPB-3, FIPR-4, and NSPB-12 were tagged C-terminally with mNeonGreen. The mNeonGreen coding sequence was PCR-amplified from the *C. elegans* strain WD835 (a kind gift from Brent Derry) using the following primers: 5-mNeon (5'-GTCAGACCGGTGGCGGTGGATCAGTCTC CAAGGGAGAGGAGGACAACATGG-3') and 3-mNeon (5'-TTACGGAATTCTCACCTTGTAGAGCTC GTCCATTCCCATG-3'). The 5-mNeon primer introduced a flexible GGGGS linker sequence to the epitope tag. The resulting PCR product was purified, digested with AgeI and EcoRI, and the 728 bp fragment was ligated to the 5 kb AgeI/EcoRI digested pPRGS762 (*unc-6p::YFP*) vector backbone to generate pPRJK1199 (*unc-6p-mNeonGreen-unc-54* 3'UTR). The coding and upstream promoter sequences (up to the end of the upstream gene) of IDPA-3, IDPB-3, FIPR-4, and NSPB-12 were amplified from wildtype *C. elegans* N2 genomic DNA template using the following primer pairs: 5-IDPA-3 (5'-CCGTACTGCAGAGCATCTCTAGAACTGACCATCTGACC-3') and 3-IDPA-3 (5'-GTTAGACCGGTG TTTGGCATTGGTGGCCATCCTCCTTG-3'); 5-IDPB-3 (5'-CAGTACTGCAGAGCAGATGATCTCACTA

GTGCAACC-3') and 3-IDPB-3 (5'-GTTAGACCGGTGCACTTGTCTCCTCCCTGGCTGG-3'); 5-FIPR-4 (5'-CCGTACTGCAGCATGTGTTGGTTTTGTCATAGAACTGTCG-3') and 3-FIPR-4 (5'-GTTAGACCGGTGTTCTGAATAGGTCCAAATCCAGC-3'); 5-NSPB-12 (5'-CCGTAATGCATTGCTGGCGTATTGTCTAAACCTTGC-3') and 3-NSPB-12 (5'-GTTAGACCGGTAGCGGTGGTTGGCTTCTGATTGTTAAG-3'). The PCR products were purified, digested with PstI and AgeI (IDPA-3, IDPB-3, FIPR-4) or NsiI and AgeI (NSPB-12), and ligated to the 4.2 kb fragment of the PstI/AgeI digested pPRJK1199 vector to generate pPRJK1213 (idpa-3p::IDPA-3::mNeonGreen [1232 bp of sequence upstream of the ATG]), pPRJK1202 (idpb-3p::IDPB-3::mNeonGreen [334 bp of sequence upstream of the ATG]), pPRJK1212 (fipr-4p::FIPR-4::mNeonGreen [1360 bp of sequence upstream of the ATG]), and pPRJK1203 (nspb-12p::NSPB-12::mNeonGreen [1973 bp of sequence upstream of the ATG]), respectively. All constructs were verified by sequencing. Wildtype *C. elegans* N2 worms were injected with each of the constructs described above along with the pPRGS382 (*myo-2p::mCherry*) co-injection marker at the following concentrations for expression analysis: pPRJK1213 (10 ng/μL) + pPRGS382 (2 ng/μL) + pKS (88 ng/μL); pPRJK1202 (10 ng/μL) + pPRGS382 (2 ng/μL) + pKS (88 ng/μL); pPRJK1212 (10 ng/μL) + pPRGS382 (2 ng/μL) + pKS (88 ng/μL); pPRJK1203 (10 ng/μL) + pPRGS382 (2 ng/μL) + pKS (88 ng/μL).

Pulse-chase analyses

Synchronized wildtype L1 worms are plated on 10 cm plates at 7000 L1s/plate seeded with OP50 *E. coli* strain. Plates with worms destined for pulse-chase analyses of larvae or adults are grown at 16°C or 25°C, respectively. Then, 72 hr after plating, the 'L3' samples and the 'adult' samples are washed with M9 to remove bacteria. The concentrations and solvents for all dyes are described in the relevant methods section. In all cases, 50 μL of packed worms from centrifugation are used per tube in the dye incubation. Note that the number of worms should not exceed 1000 because adding more worms reduces stain intensity. Also, siliconized tips are used with the ends cut with flame-sterilized scissors to avoid injuring the worms. The tubes with worms and dye are then incubated on a nutator for 3 hr in the dark at room temperature. After incubation, the 1.5 mL tubes are spun at 5000 rpm for 1 min and the concentrated pellet is carefully transferred to 15 mL falcon tube and washed with 8 mL of M9 buffer to remove excess dye. The tubes are inverted gently and spun at 2000 rpm for 1 min. The supernatant is removed and the concentrated washed worms are spotted onto the clear (agar) surface of 6 cm plates seeded with OP50. Then, 30 min later, 20–30 worms are picked onto a second plate lightly seeded with OP50. The staining of the cuticle for each is then semi-quantitatively assessed on an epifluorescent microscope. These data represent the pre-chase counts. The scoring system was as follows: animals exhibiting robust staining in the buccal cavity and anterior channels = 3; animals exhibiting moderate staining in the buccal cavity and anterior channels = 2; animals showing faint staining in the buccal cavity and anterior channels = 1; animals showing no detectable staining in any part of the pharynx cuticle = 0. The remaining animals on the original 6 cm plate are incubated for a total of 18 hr at 20°C, after which dye staining of the cuticle is quantified. These data represent the post-chase counts.

Generating *mlt-9*(RNAi) Cuticle Defects

mlt-9 RNAi was carried out as described previously (Fränd et al., 2005) with some modifications. Briefly, a bacterial culture expressing dsRNA of *mlt-9* (referred to here as *mlt-9*(RNAi)) (Kamath et al., 2003) was started from a single colony in 30 mL LB broth containing 100 μg/mL ampicillin for 18 hr at 37°C at 200 rpm. The cells were pelleted by centrifuging at 3200 rpm for 15 min, after which the cells were concentrated tenfold. Then, 1 mL of the pelleted cells was added to 10 cm NGM agar plates containing 8 mM IPTG and 40 μg/mL carbenicillin and left to dry overnight at room temperature in the dark. The next day (day 0), 6500 synchronized L1s were plated onto each RNAi plate, after which the plates were stored at 16°C in the dark. Ninety hours later, the worms were inspected for *mlt-9* RNAi phenotypes. Approximately 50% of *mlt-9*(RNAi)-treated worms exhibit the expected cuticle defects. Performing mock RNAi with the empty L4440 plasmid failed to yield worms with obvious cuticle defects.

Dye staining of wildtype and *mlt-9*(RNAi) animals

Congo Red (CR) staining

Synchronized wildtype adult worms were washed and incubated with 0.02% CR from a 1% stock (w/v, dissolved in DMSO; Fisher chemical C580-25; CAS 573-58-0) in 500 μ L of liquid NGM for 3 hr in the dark. Worms are then prepped for microscopic analysis as described above.

Thioflavin S (ThS) staining

Synchronized wildtype adult worms were washed and incubated with 0.1% ThS from a 10% stock (w/v, dissolved in DMSO; ThS; SIGMA, T1892-25G) in 500 μ L of liquid NGM for 3 hr in the dark. Worms are then prepped for microscopic analysis as described above. The concentration chosen for ThS staining of *C. elegans* pharynx was based on a published protocol (Wu et al., 2006). ThS is a complex mixture of molecules with two major species of 377.1 and 510.1 MW and several other minor species (Enthammer et al., 2013). Given that the ratio of molecules is unknown, we used an average MW of 443.6 for ThS in our calculations.

Eosin Y (EY) staining

EY staining was performed as described (Heustis et al., 2012). Briefly, synchronized wildtype adult worms were washed and incubated with 0.15 mg/mL from a 5 mg/mL stock (dissolved in 70% ethanol; Eosin Y; Sigma-Aldrich, E4009) in 500 μ L of liquid NGM for 3 hr in the dark. Worms are then prepped for microscopic analysis as described above. Note that eosin Y stock should be stored at -20°C and before its use it should be incubated at 55°C for ~ 2 min and vigorously vortexed to ensure its solvation.

Calcofluor white (CFW) staining

Synchronized wildtype adult worms were washed and incubated with 0.005% CFW from a 1% stock (w/v, dissolved in DMSO; Fluorescent Brightener 28, Sigma-Aldrich, CAS 4404-43-7) in 500 μ L NGM for 3 hr in the dark. Worms are then prepped for microscopic analysis as described above. Note that the CFW stock should be placed in boiling water for ~ 2 min and then vigorously vortexed to ensure solvation of the dye.

Calculations of low-complexity and intrinsic disorder

LCRs in the amino acid sequences of each protein within the *C. elegans* proteome (WormBase release WS274) were identified using the SEG algorithm with default stringency parameters set (i.e., WINDOW = 12, LOWcut = 2.2, HIGHcut = 2.5) (Wootton and Federhen, 1993). Percentage sequence in LCRs was calculated for each protein based on the total number of residues found within LCRs returned by SEG relative to protein length. The intrinsic disorder of each protein within the *C. elegans* proteome (obtained from WormBase version WS274) was analyzed using the Spot-Disorder script (Hanson et al., 2017). The computational analysis was conducted using the Niagara supercomputer at the SciNet HPC Consortium. The GNU 'parallel' package was used to perform the computational analysis in parallel. The individual protein SPOT-Disorder output data were then computationally analyzed using Python for IDRs (defined as any string of 30 or more disordered residues), total number of disordered residues, and percentage of amino acid residues within intrinsically disordered regions.

LLPhyScore calculations

The LLPhyScore phase separation score of each protein was calculated using the LLPhyScore algorithm (Cai et al., 2022). The LLPhyScore algorithm is a machine learning-based interpretable predictive algorithm that is based on the idea that a combination of multiple different physical interactions drives protein liquid-liquid phase separation. A protein's LLPhyScore is a weighted combination of eight sub-scores, each representing one physical feature that is inferred from the input sequence. These physical features include protein-water interactions, hydrogen bonds, pi-pi interactions, disorder, kinked-beta structure, and electrostatics. The scores are optimized via training with 500+ experimentally known phase-separating protein sequences against selected negative sequences. More details about this algorithm can be found in the manuscript in preparation.

AmyloGram and path analyses

AmyloGram (*Burdukiewicz et al., 2017*) is a method based on machine learning, trained on hexapeptides experimentally tested for their amyloidogenic propensities (*Wozniak and Kotulska, 2015*). Amino acids are represented by the alphabet that best encoded amyloidogenicity of peptides modeled by n-grams, and it was optimized by a random forest classifier. Classification of a protein amyloidogenicity included calculating its profile with a hexapeptide window shifting along the protein chain. Proteins with amyloid propensity were identified on the basis of an appearance of at least one amyloidogenic fragment. To avoid an excessive number of false positives, non-default specificity values were used: 0.95 and 0.99.

PATH (*Wojciechowski and Kotulska, 2020*) uses molecular modeling and machine learning. It is a computational pipeline based on Python and bash scripts, using Modeller (*Sali and Blundell, 1993*) and PyRosetta (*Chaudhury et al., 2010*). A potentially amyloidogenic query sequence of a hexapeptide was threaded on seven representative amyloid templates. Comparative structure modeling provided evaluation of the models with statistics and physics-based functions. Next, the scores were used by the logistic regression classifier. The analyses with PATH were carried out in two stages. The first scan along the protein chain was done by AmyloGram with the specificity threshold at 0.99, which was then followed by structural modeling and classification using PATH. The second stage was only applied to amyloid-positive regions found by AmyloGram.

LARKS analyses

LARKS predictions were done on a proteome downloaded from WormBase on October 18, 2021. Sequences not completely comprised of the 20 canonical amino acids were rejected from analysis. Each protein from the filtered proteome set of 20,042 proteins was then submitted for LARKS predictions. First, the sequence was separated into a series of overlapping hexapeptide segments (each segment overlapped with five residues from the segment before it; a 150 amino acid sequence contains 145 hexapeptides). The sidechains for each residue in a hexapeptide are computationally grafted onto a fibril model for each of three different LARKS structures (FUS-SYSGYS, FUS-STGGYG, and hnRNPA1-GYNGFG; PDB IDs: 6BWZ, 6BZP, and 6BXX). Energy minimization is done using a Rosetta energy score as a readout, and if the final energy is below a backbone-dependent threshold, then hexapeptide segment is considered a LARKS. Proteins' LARKS content was determined by the number of favorable LARKS segments divided by the length of the protein.

In vitro expression and analysis of IDPs

Expression vectors and constructs

All protein expression vectors generated for this work were derivatives of the pMBP-FUS-FL-WT (a gift from Nicolas Fawzi [Addgene plasmid # 98651; <http://n2t.net/addgene:98651>; RRID:Addgene_98651; *Burke et al., 2015*], which was modified to remove the FUS1 coding region and to have two cloning sites BamHI and NotI for facile cloning of new proteins in phase with the HIS-tagged Maltose Binding Protein (MBP) at the N-terminus followed by a TEV protease cleavage site (TEVcs) to generate pPRRH1197. The coding region of proteins of interest (minus signal sequences) was codon optimized for expression in *E. coli*, synthesized with appropriate linkers, and subcloned into frame with MBP (GenScript), resulting in pPRPM1191 (HIS::MBP::TEVcs::IDPC-2).

Protein preparation and purification

Proteins were expressed in *E. coli* BL21DE3 RIPL in LB with kanamycin and chloramphenicol. Cells were grown to OD₆₀₀ of 0.5, induced with 0.5 mM IPTG, and grown overnight at 18°C. The next day cultures were centrifuged at 5000 × g at 4°C for 10 min. Pellets were frozen at -80°C then thawed and resuspended in lysis buffer (2.5 mM Tris pH 7.5, 500 mM NaCl, 20 mM imidazole, 2 mM DTT and 1x Protease inhibitor cocktail; Sigma, P8849). This suspension was sonicated to lyse *E. coli* and clarified by centrifugation at 39,000 × g for 45 min at 4°C. The cleared supernatant was added directly to a pre-equilibrated nickel column. Optimal wash and elution conditions had to be determined empirically for each protein. Purified fractions were then dialyzed with 2.5 mM Tris pH 7.5, 150 mM NaCl, 2 mM DTT to remove excess salts and imidazole and protein concentration determined with Bradford assay.

Phase separation assays

Proteins were incubated in 2.5 mM Tris pH 7.5, 150 mM NaCl, 2 mM DTT with either 5% Ficoll (Sigma, F2637) for MBP::FUS1 or 15% Ficoll for MBP::IDPC-2 for 1 hr at 30°C with or without TEV protease (10 units in a 50 μ L reaction). The optimal percent Ficoll was determined empirically. Turbidity was measured at 395 nm with a Clariostar plate reader (Mandel). 10 μ L of each reaction was spotted onto slides with coverslips then condensates visualized with DIC using a Leica DMRA2 microscope at \times 63 magnification.

Protein sequence analysis and logo generation

We used Clustal Omega (*Sievers et al., 2011*) to align the 110 low-complexity protein sequences and generate a percent identity matrix based on the multiple sequence alignment. For those low-complexity proteins with a predicted signal peptide, the first 20 amino acids were removed from the protein sequence before alignment.

To generate sequence logos, full-length protein sequences from each of the low-complexity protein families identified by the percent identity matrix were aligned using ClustalW (*Thompson et al., 1994*). Sequence logos were constructed based on these alignments using WebLogo 3.7.4; (<https://weblogo.berkeley.edu/>; *Crooks et al., 2004*; *Schneider and Stephens, 1990*). Amino acid residues were colored according to their chemical properties: polar (G,S,T,Y,C) in green, neutral (Q,N) in purple, basic (K,R,H) in blue, acidic (D,E) in red, and hydrophobic (A,V,L,I,P,W,F,M) in black. The height of the symbol within each stack indicates the relative frequency of that amino acid in that position. Stack widths are scaled by the fraction of symbols in that position (positions with many gaps are narrow). Details of protein sequences used can be found in **Figure 4—source data 1**.

Statistics and graphs

Except where indicated, statistical differences were measured using a two-tailed Student's *t*-test. Plots were either generated using Prism 8 graphing software or Excel.

Materials availability statement

The *C. elegans* strains expressing the fluorescently tagged fusion proteins will be made available at the *C. elegans* Genetic Center.

Acknowledgements

We are grateful to David Raizen and Fred Keeley for helpful conversations, and for the work of JG White, E Southgate, JN Thomson, and S Brenner, who generated the serial sections that we show in Figure 1, and KA Wright, JN Thomson, who generated the transverse images that we show in Figure 3. We thank John White and Jonathan Hodgkin for allowing MRC/LMB archival TEM images to be sent to WormAtlas (David Hall) at Albert Einstein College of Medicine for long-term curation. We thank David Hall and Zeynep Altun for helpful advice and for sharing unpublished images via WormImage and WormAtlas (funded by an NIH grant [OD 010943] to DH Hall). We also thank Iva Pritisanac for mentorship of intrinsic disorder calculations, the staff at Wormbase for assembling and communicating proteome files to us, and Tim Schedl for guidance on new gene assignments. For mutant strains, we are grateful to Mei Zhen and Wesley Hung, Harald Hutter, the *C. elegans* Gene Knockout Consortium, Don Moerman's and Bob Waterston's million mutation project, Shohei Mitani, and the *C. elegans* Genetics Centre. We thank Brent Derry and Matthew Eroglu for the codon-optimized mNeon-Green sequence. PJR dedicates this article to the recently retired Don Moerman – many things would have been more difficult without you. Funding was from NKFI grant 127909 (KT, VG), National Science Foundation Grant 1616265 (MPH), a grant (2019/35/B/NZ2/03997) from the National Science Centre, Poland (MK) NSERC Alexander Graham Bell Canada Graduate Scholarship (JK) Canadian Institutes of Health Research grants 376634 and 313296 (PJR) Canadian Research Chair grant (PJR).

Additional information

Funding

| Funder | Grant reference number | Author |
|--|------------------------|---------------------|
| NKFI | 127909 | Kristóf Takács |
| National Science Foundation | 1616265 | Michael P Hughes |
| National Science Centre, Poland | 2019/35/B/NZ2/03997 | Malgorzata Kotulska |
| Canadian Institutes of Health Research | 376634 | Peter J Roy |
| Canadian Institutes of Health Research | 313296 | Peter J Roy |
| National Science and Engineering Council of Canada | | Jessica Knox |
| Canada Research Chairs | | Peter J Roy |

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Muntasir Kamal, Investigation, Visualization, Methodology, Writing - original draft, Writing – review and editing; Levon Tokmakjian, Investigation, Visualization, Methodology, Writing – review and editing; Jessica Knox, Data curation, Investigation, Visualization, Methodology, Writing – review and editing; Peter Mastrangelo, Investigation, Methodology, Writing – review and editing; Jingxiu Ji, Investigation, Visualization; Hao Cai, Jakub W Wojciechowski, Kristóf Takács, Xiaoquan Chu, Formal analysis; Michael P Hughes, Formal analysis, Methodology, Writing – review and editing; Jianfeng Pei, Vince Grolmusz, Malgorzata Kotulska, Formal analysis, Supervision; Julie Deborah Forman-Kay, Conceptualization, Supervision, Methodology, Writing – review and editing; Peter J Roy, Conceptualization, Data curation, Formal analysis, Supervision, Funding acquisition, Investigation, Visualization, Methodology, Writing - original draft, Project administration, Writing – review and editing

Author ORCIDs

Jessica Knox <http://orcid.org/0000-0003-1465-5852>

Jingxiu Ji <http://orcid.org/0000-0003-4121-7719>

Jakub W Wojciechowski <http://orcid.org/0000-0001-5289-653X>

Vince Grolmusz <http://orcid.org/0000-0001-9456-8876>

Malgorzata Kotulska <http://orcid.org/0000-0002-2015-5339>

Julie Deborah Forman-Kay <http://orcid.org/0000-0001-8265-972X>

Peter J Roy <http://orcid.org/0000-0003-2959-2276>

Ethics

We (the authors) affirm that we have complied with all relevant ethical regulations for animal testing and research. Given that our experiments focused exclusively on the invertebrate nematode worm *C. elegans*, no ethical approval was required for any of the presented work.

Decision letter and Author response

Decision letter <https://doi.org/10.7554/eLife.79396.sa1>

Author response <https://doi.org/10.7554/eLife.79396.sa2>

Additional files

Supplementary files

- Supplementary file 1. Genes of special interest and evidence of pharynx expression. Note that the data within this table is also available in the **Figure 4—source data 1** file so as to put it in

context with other data for each gene. Also note that the column indicators below are named after the column in Supplementary File 1. (C) All 78 gene products called out in **Figure 4A** are shown, along with all 226 genes represented in **Figure 4A**, in addition to 17 genes that are of special interest (including additional members of the *idpp* gene class referred to elsewhere in the text). (B) The APPGs that have higher sequence similarity to one another and have more similar temporal expression patterns are described as APPG family (#1) members to distinguish them from more divergent APPGs. (E) The Name Status indicates the 41 new WormBase-approved gene assignments. (H) The indicated hour and degree is with respect to **Figure 4A**. (J) In some cases, the updated Signal P algorithm will identify a signal peptide when ParaSite did not, as indicated with a 'no, but likley SS.' (L) The spatial expression patterns of the indicated clones can be inspected at <http://nematode.lab.nig.ac.jp/dbest/srchbyclone.html>. A green color indicates confirmation of the expected expression pattern (enriched in pharynx); 'no signal' indicates little to no signal anywhere in photo micrographs. In two cases indicated in pink, signal could be observed in the animal, but the pharynx lacked signal. (M, N) The PubMed ID number (PMID) is shown for the publication that provides additional spatial expression information for the gene. The nature of the data is either from a transgene (transg), an antibody (Ab), or is sequence-based (seq). A green color indicates confirmation of the expected expression pattern (enriched in pharynx).

- Supplementary file 2. Transcript levels of six low-complexity protein families within pharynx cells. The data is extracted from the data presented in **Figure 10B**.
- MDAR checklist

Data availability

All source data for the spatiotemporal reconstruction is in the Source data files.

References

- Aguinaldo AM**, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**:489–493. DOI: <https://doi.org/10.1038/387489a0>, PMID: 9168109
- Altun ZF**, Hall DH. 2020. WormAtlas. WormAtlas.
- Andersen SO**. 2011. Are structural proteins in insect cuticles dominated by intrinsically disordered regions? *Insect Biochemistry and Molecular Biology* **41**:620–627. DOI: <https://doi.org/10.1016/j.ibmb.2011.03.015>, PMID: 21477652
- André AAM**, Spruijt E. 2020. Liquid-Liquid phase separation in crowded environments. *International Journal of Molecular Sciences* **21**:E5908. DOI: <https://doi.org/10.3390/ijms21165908>, PMID: 32824618
- Avery L**. 1993. Motor neuron M3 controls pharyngeal muscle relaxation timing in *Caenorhabditis elegans*. *The Journal of Experimental Biology* **175**:283–297. DOI: <https://doi.org/10.1242/jeb.175.1.283>, PMID: 8440973
- Baker LG**, Specht CA, Donlin MJ, Lodge JK. 2007. Chitosan, the deacetylated form of chitin, is necessary for cell wall integrity in *Cryptococcus neoformans*. *Eukaryotic Cell* **6**:855–867. DOI: <https://doi.org/10.1128/EC.00399-06>, PMID: 17400891
- Banani SF**, Lee HO, Hyman AA, Rosen MK. 2017. Biomolecular condensates: organizers of cellular biochemistry. *Nature Reviews. Molecular Cell Biology* **18**:285–298. DOI: <https://doi.org/10.1038/nrm.2017.7>, PMID: 28225081
- Bennhold H**. 1922. Specific staining of amyloid by congo red. *Muenchen. Med. Wochenschr* **69**:1537–1538.
- Brodeur J**, Thériault C, Lessard-Beaudoin M, Marcil A, Dahan S, Lavoie C. 2012. LDLR-related protein 10 (LRP10) regulates amyloid precursor protein (APP) trafficking and processing: evidence for a role in Alzheimer's disease. *Molecular Neurodegeneration* **7**:31. DOI: <https://doi.org/10.1186/1750-1326-7-31>, PMID: 22734645
- Burdukiewicz M**, Sobczyk P, Rödiger S, Duda-Madej A, Mackiewicz P, Kotulska M. 2017. Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports* **7**:12961. DOI: <https://doi.org/10.1038/s41598-017-13210-9>, PMID: 29021608
- Burke KA**, Janke AM, Rhine CL, Fawzi NL. 2015. Residue-by-residue view of in vitro FUS granules that bind the C-terminal domain of RNA polymerase II. *Molecular Cell* **60**:231–241. DOI: <https://doi.org/10.1016/j.molcel.2015.09.006>, PMID: 26455390
- Cai H**, Vernon RM, Forman-Kay JD. 2022. An interpretable machine-learning algorithm to predict disordered protein phase separation based on biophysical interactions. *Biomolecules* **12**:1131. DOI: <https://doi.org/10.3390/biom12081131>, PMID: 36009025
- Cam JA**, Zerbinatti CV, Knisely JM, Hecimovic S, Li Y, Bu G. 2004. The low density lipoprotein receptor-related protein 1B retains beta-amyloid precursor protein at the cell surface and reduces amyloid-beta peptide production. *The Journal of Biological Chemistry* **279**:29639–29646. DOI: <https://doi.org/10.1074/jbc.M313893200>, PMID: 15126508
- Cao J**, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, Adey A, Waterston RH, Trapnell C, Shendure J. 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**:661–667. DOI: <https://doi.org/10.1126/science.aam8940>, PMID: 28818938

- Chaudhury S**, Lyskov S, Gray JJ. 2010. PyRosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. *Bioinformatics* **26**:689–691. DOI: <https://doi.org/10.1093/bioinformatics/btq007>, PMID: 20061306
- Chu X**, Sun T, Li Q, Xu Y, Zhang Z, Lai L, Pei J. 2022. Prediction of liquid-liquid phase separating proteins using machine learning. *BMC Bioinformatics* **23**:72. DOI: <https://doi.org/10.1186/s12859-022-04599-w>, PMID: 35168563
- Cohen SIA**, Arosio P, Presto J, Kurudenkandy FR, Biverstal H, Dolfe L, Dunning C, Yang X, Frohm B, Vendruscolo M, Johansson J, Dobson CM, Fisahn A, Knowles TPJ, Linse S. 2015. A molecular chaperone breaks the catalytic cycle that generates toxic A β oligomers. *Nature Structural & Molecular Biology* **22**:207–213. DOI: <https://doi.org/10.1038/nsmb.2971>, PMID: 25686087
- Cox GN**, Kusch M, Edgar RS. 1981. Cuticle of *Caenorhabditis elegans*: its isolation and partial characterization. *The Journal of Cell Biology* **90**:7–17. DOI: <https://doi.org/10.1083/jcb.90.1.7>, PMID: 7251677
- Cremades N**, Cohen SIA, Deas E, Abramov AY, Chen AY, Orte A, Sandal M, Clarke RW, Dunne P, Aprile FA, Bertocini CW, Wood NW, Knowles TPJ, Dobson CM, Klenerman D. 2012. Direct observation of the interconversion of normal and toxic forms of α -synuclein. *Cell* **149**:1048–1059. DOI: <https://doi.org/10.1016/j.cell.2012.03.037>, PMID: 22632969
- Crooks GE**, Hon G, Chandonia JM, Brenner SE. 2004. Weblogo: a sequence logo generator: Figure 1. *Genome Research* **14**:1188–1190. DOI: <https://doi.org/10.1101/gr.849004>, PMID: 15173120
- Deiana A**, Forcelloni S, Porrello A, Giansanti A. 2019. Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. *PLOS ONE* **14**:e0217889. DOI: <https://doi.org/10.1371/journal.pone.0217889>, PMID: 31425549
- Divry M PF**. 1927. Sur les propriétés optiques de l'amyloïde. *C. R. Soc. Biol* **97**:1808–1810.
- Edelman TLB**, McCulloch KA, Barr A, Frøkjær-Jensen C, Jorgensen EM, Rougvie AE. 2016. Analysis of a lin-42/period null allele implicates all three isoforms in regulation of *Caenorhabditis elegans* molting and developmental timing. *G3: Genes, Genomes, Genetics* **6**:4077–4086. DOI: <https://doi.org/10.1534/g3.116.034165>, PMID: 27729432
- El-Gebali S**, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The pfam protein families database in 2019. *Nucleic Acids Research* **47**:D427–D432. DOI: <https://doi.org/10.1093/nar/gky995>, PMID: 30357350
- Enthammer M**, Papadakis ES, Salomé Gachet M, Deutsch M, Schwaiger S, Koziel K, Ashraf MI, Khalid S, Wolber G, Packham G, Cutress RI, Stuppner H, Troppmair J. 2013. Isolation of a novel thioflavin S-derived compound that inhibits Bag-1-mediated protein interactions and targets BRAF inhibitor-resistant cell lines. *Molecular Cancer Therapeutics* **12**:2400–2414. DOI: <https://doi.org/10.1158/1535-7163.MCT-13-0142>, PMID: 24048738
- Frand AR**, Russel S, Ruvkun G. 2005. Functional genomic analysis of *C. elegans* molting. *PLOS Biology* **3**:e312. DOI: <https://doi.org/10.1371/journal.pbio.0030312>, PMID: 16122351
- Franz M**, Rodriguez H, Lopes C, Zuberi K, Montojo J, Bader GD, Morris Q. 2018. GeneMANIA update 2018. *Nucleic Acids Research* **46**:W60–W64. DOI: <https://doi.org/10.1093/nar/gky311>, PMID: 29912392
- George-Raizen JB**, Shockley KR, Trojanowski NF, Lamb AL, Raizen DM. 2014. Dynamically-expressed prion-like proteins form a cuticle in the pharynx of *Caenorhabditis elegans*. *Biology Open* **3**:1139–1149. DOI: <https://doi.org/10.1242/bio.20147500>, PMID: 25361578
- Ghai V**, Smit RB, Gaudet J. 2012. Transcriptional regulation of HLH-6-independent and subtype-specific genes expressed in the *Caenorhabditis elegans* pharyngeal glands. *Mechanisms of Development* **129**:284–297. DOI: <https://doi.org/10.1016/j.mod.2012.06.005>, PMID: 22759833
- Hanson J**, Yang Y, Paliwal K, Zhou Y. 2017. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **33**:685–692. DOI: <https://doi.org/10.1093/bioinformatics/btw678>, PMID: 28011771
- Hendriks G-J**, Gaidatzis D, Aeschmann F, Großhans H. 2014. Extensive oscillatory gene expression during *C. elegans* larval development. *Molecular Cell* **53**:380–392. DOI: <https://doi.org/10.1016/j.molcel.2013.12.013>, PMID: 24440504
- Hertz-Fowler C**, Hall N. 2004. Parasite genome databases and web-based resources. *Methods in Molecular Biology* **270**:45–74. DOI: <https://doi.org/10.1385/1-59259-793-9-045>, PMID: 15153622
- Heustis RJ**, Ng HK, Brand KJ, Rogers MC, Le LT, Specht CA, Fuhrman JA. 2012. Pharyngeal polysaccharide deacetylases affect development in the nematode *C. elegans* and deacetylate chitin in vitro. *PLOS ONE* **7**:e40426. DOI: <https://doi.org/10.1371/journal.pone.0040426>, PMID: 22808160
- Ho B**, Baryshnikova A, Brown GW. 2018. Unification of protein abundance datasets yields a quantitative *Saccharomyces cerevisiae* proteome. *Cell Systems* **6**:192–205. DOI: <https://doi.org/10.1016/j.cels.2017.12.004>, PMID: 29361465
- Huang KM**, Cosman P, Schafer WR. 2008. Automated detection and analysis of foraging behavior in *Caenorhabditis elegans*. *Journal of Neuroscience Methods* **171**:153–164. DOI: <https://doi.org/10.1016/j.jneumeth.2008.01.027>, PMID: 18342950
- Hughes MP**, Sawaya MR, Boyer DR, Goldschmidt L, Rodriguez JA, Cascio D, Chong L, Gonen T, Eisenberg DS. 2018. Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks. *Science* **359**:698–701. DOI: <https://doi.org/10.1126/science.aan6398>, PMID: 29439243

- Hutchison LAD**, Berger B, Kohane IS. 2020. Meta-analysis of *Caenorhabditis elegans* single-cell developmental data reveals multi-frequency oscillation in gene activation. *Bioinformatics* **36**:4047–4057. DOI: <https://doi.org/10.1093/bioinformatics/btz864>, PMID: 31860066
- Ikedo K**, Suzuki S, Shigemitsu Y, Tenno T, Goda N, Oshima A, Hiroaki H. 2020. Presence of intrinsically disordered proteins can inhibit the nucleation phase of amyloid fibril formation of A β (1–42) in amino acid sequence independent manner. *Scientific Reports* **10**:12334. DOI: <https://doi.org/10.1038/s41598-020-69129-1>, PMID: 32703978
- Iwata N**, Tsubuki S, Takaki Y, Shirotani K, Lu B, Gerard NP, Gerard C, Hama E, Lee HJ, Saido TC. 2001. Metabolic regulation of brain Abeta by neprilysin. *Science* **292**:1550–1552. DOI: <https://doi.org/10.1126/science.1059946>, PMID: 11375493
- Jehle F**, Macías-Sánchez E, Sviben S, Fratzl P, Bertinetti L, Harrington MJ. 2020. Hierarchically-structured metalloprotein composite coatings biofabricated from co-existing condensed liquid phases. *Nature Communications* **11**:862. DOI: <https://doi.org/10.1038/s41467-020-14709-y>, PMID: 32054841
- Jeon M**, Gardner HF, Miller EA, Deshler J, Rougvié AE. 1999. Similarity of the *C. elegans* developmental timing protein LIN-42 to circadian rhythm proteins. *Science* **286**:1141–1146. DOI: <https://doi.org/10.1126/science.286.5442.1141>, PMID: 10550049
- John DT**, Petri W. 2006. Medical Parasitology. Saunders Elsevier, St. Louis.
- Kamal M**, Moshiri H, Magomedova L, Han D, Nguyen KCQ, Yeo M, Knox J, Bagg R, Won AM, Szlapa K, Yip CM, Cummins CL, Hall DH, Roy PJ. 2019. The marginal cells of the *Caenorhabditis elegans* pharynx scavenge cholesterol and other hydrophobic small molecules. *Nature Communications* **10**:3938. DOI: <https://doi.org/10.1038/s41467-019-11908-0>, PMID: 31477732
- Kamath RS**, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J. 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**:231–237. DOI: <https://doi.org/10.1038/nature01278>, PMID: 12529635
- Kato M**, Yang YS, Sutter BM, Wang Y, McKnight SL, Tu BP. 2019. Redox state controls phase separation of the yeast ataxin-2 protein via reversible oxidation of its methionine-rich low-complexity domain. *Cell* **177**:711–721. DOI: <https://doi.org/10.1016/j.cell.2019.02.044>, PMID: 30982603
- Keresztes L**, Szögi E, Varga B, Farkas V, Perczel A, Grolmusz V. 2021. The Budapest amyloid predictor and its applications. *Biomolecules* **11**:500. DOI: <https://doi.org/10.3390/biom11040500>, PMID: 33810341
- Lancaster AK**, Nutter-Upham A, Lindquist S, King OD. 2014. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics* **30**:2501–2502. DOI: <https://doi.org/10.1093/bioinformatics/btu310>, PMID: 24825614
- Lazetic V**, Fay DS. 2017. Molting in *C. elegans*. *Worm* **6**:e1330246. DOI: <https://doi.org/10.1080/21624054.2017.1330246>
- Letunic I**, Bork P. 2018. 20 years of the smart protein domain annotation resource. *Nucleic Acids Research* **46**:D493–D496. DOI: <https://doi.org/10.1093/nar/gkx922>, PMID: 29040681
- Lin YH**, Forman-Kay JD, Chan HS. 2018. Theories for sequence-dependent phase behaviors of biomolecular condensates. *Biochemistry* **57**:2499–2508. DOI: <https://doi.org/10.1021/acs.biochem.8b00058>, PMID: 29509422
- Lu T**, Spruijt E. 2020. Multiphase complex coacervate droplets. *Journal of the American Chemical Society* **142**:2905–2914. DOI: <https://doi.org/10.1021/jacs.9b11468>, PMID: 31958956
- McCulloch KA**, Rougvié AE. 2014. *Caenorhabditis elegans* period homolog lin-42 regulates the timing of heterochronic miRNA expression. *PNAS* **111**:15450–15455. DOI: <https://doi.org/10.1073/pnas.1414856111>, PMID: 25319259
- McLoughlin DM**, Miller CCJ. 2008. The Fe65 proteins and Alzheimer's disease. *Journal of Neuroscience Research* **86**:744–754. DOI: <https://doi.org/10.1002/jnr.21532>, PMID: 17828772
- Meeuse MW**, Hauser YP, Morales Moya LJ, Hendriks G-J, Eglinger J, Bogaarts G, Tsiarris C, Großhans H. 2020. Developmental function and state transitions of a gene expression oscillator in *Caenorhabditis elegans*. *Molecular Systems Biology* **16**:e9498. DOI: <https://doi.org/10.15252/msb.209975>, PMID: 33438821
- Mejias J**, Truong NM, Abad P, Favery B, Quentin M. 2019. Plant proteins and processes targeted by parasitic nematode effectors. *Frontiers in Plant Science* **10**:970. DOI: <https://doi.org/10.3389/fpls.2019.00970>, PMID: 31417587
- Mello C**, Fire A. 1995. Dna transformation. *Methods in Cell Biology* **48**:451–482 PMID: 8531738.
- Michelitsch MD**, Weissman JS. 2000. A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions. *PNAS* **97**:11910–11915. DOI: <https://doi.org/10.1073/pnas.97.22.11910>, PMID: 11050225
- Mittag T**, Parker R. 2018. Multiple modes of protein-protein interactions promote RNP granule assembly. *Journal of Molecular Biology* **430**:4636–4649. DOI: <https://doi.org/10.1016/j.jmb.2018.08.005>, PMID: 30099026
- Monsalve GC**, Van Buskirk C, Frand AR. 2011. LIN-42/PERIOD controls cyclical and developmental progression of *C. elegans* molts. *Current Biology* **21**:2033–2045. DOI: <https://doi.org/10.1016/j.cub.2011.10.054>, PMID: 22137474
- Motohashi T**, Tabara H, Kohara Y. 2006. WormBook. The *C. Elegans* Research Community.
- Mountain GA**, Keating CD. 2020. Formation of multiphase complex coacervates and partitioning of biomolecules within them. *Biomacromolecules* **21**:630–640. DOI: <https://doi.org/10.1021/acs.biomac.9b01354>, PMID: 31743027

- Moussian B.** 2010. Recent advances in understanding mechanisms of insect cuticle differentiation. *Insect Biochemistry and Molecular Biology* **40**:363–375. DOI: <https://doi.org/10.1016/j.ibmb.2010.03.003>, PMID: 20347980
- Muiznieks LD,** Sharpe S, Pomès R, Keeley FW. 2018. Role of liquid-liquid phase separation in assembly of elastin and other extracellular matrix proteins. *Journal of Molecular Biology* **430**:4741–4753. DOI: <https://doi.org/10.1016/j.jmb.2018.06.010>, PMID: 29886015
- Murray DT,** Kato M, Lin Y, Thurber KR, Hung I, McKnight SL, Tycko R. 2017. Structure of FUS protein fibrils and its relevance to self-assembly and phase separation of low-complexity domains. *Cell* **171**:615–627. DOI: <https://doi.org/10.1016/j.cell.2017.08.048>, PMID: 28942918
- Muthukrishnan S,** Merzendorfer H, Arakane Y, Yang Q. 2019. Chitin organizing and modifying enzymes and proteins involved in remodeling of the insect cuticle. *Advances in Experimental Medicine and Biology* **1142**:83–114. DOI: https://doi.org/10.1007/978-981-13-7318-3_5, PMID: 31102243
- Packer JS,** Zhu Q, Huynh C, Sivaramakrishnan P, Preston E, Dueck H, Stefanik D, Tan K, Trapnell C, Kim J, Waterston RH, Murray JI. 2019. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* **365**:eaax1971. DOI: <https://doi.org/10.1126/science.aax1971>, PMID: 31488706
- Page AP,** Johnstone IL. 2007. The cuticle. *WormBook* **1–15**:1–15. DOI: <https://doi.org/10.1895/wormbook.1.138.1>, PMID: 18050497
- Park J-O,** Pan J, Möhrlen F, Schupp M-O, Johnsen R, Baillie DL, Zapf R, Moerman DG, Hutter H. 2010. Characterization of the astacin family of metalloproteases in *C. elegans*. *BMC Developmental Biology* **10**:14. DOI: <https://doi.org/10.1186/1471-213X-10-14>, PMID: 20109220
- Patel A,** Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, Stoynev S, Mahamid J, Saha S, Franzmann TM, Pozniakovski A, Poser I, Maghelli N, Royer LA, Weigert M, Myers EW, Grill S, Drechsel D, Hyman AA, Alberti S. 2015. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* **162**:1066–1077. DOI: <https://doi.org/10.1016/j.cell.2015.07.047>, PMID: 26317470
- Reed EH,** Hammer DA. 2018. Redox sensitive protein droplets from recombinant oleosin. *Soft Matter* **14**:6506–6513. DOI: <https://doi.org/10.1039/c8sm01047a>, PMID: 30043819
- Roncero C,** Valdivieso MH, Ribas JC, Durán A. 1988. Isolation and characterization of *Saccharomyces cerevisiae* mutants resistant to calcofluor white. *Journal of Bacteriology* **170**:1950–1954. DOI: <https://doi.org/10.1128/jb.170.4.1950-1954.1988>, PMID: 3280554
- Roy PJ.** 2022. Temporal regulation of gene expression in post-mitotic cells is revealed from a synchronized population of *c. elegans* larvae. *MicroPublication Biology* **1–5**. DOI: <https://doi.org/10.17912/micropub.biology.000587>, PMID: 35783576
- Sali A,** Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* **234**:779–815. DOI: <https://doi.org/10.1006/jmbi.1993.1626>, PMID: 8254673
- Schneider TD,** Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* **18**:6097–6100. DOI: <https://doi.org/10.1093/nar/18.20.6097>, PMID: 2172928
- Schrimpf SP,** Weiss M, Reiter L, Ahrens CH, Jovanovic M, Malmström J, Brunner E, Mohanty S, Lercher MJ, Hunziker PE, Aebersold R, von Mering C, Hengartner MO. 2009. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biology* **7**:e48. DOI: <https://doi.org/10.1371/journal.pbio.1000048>, PMID: 19260763
- Sieriebriennikov B,** Sommer RJ. 2018. Developmental plasticity and robustness of a nematode mouth-form polyphenism. *Frontiers in Genetics* **9**:382. DOI: <https://doi.org/10.3389/fgene.2018.00382>, PMID: 30254664
- Sievers F,** Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology* **7**:539. DOI: <https://doi.org/10.1038/msb.2011.75>, PMID: 21988835
- Smit RB,** Schnabel R, Gaudet J. 2008. The HLH-6 transcription factor regulates *C. elegans* pharyngeal gland development and function. *PLoS Genetics* **4**:e1000222. DOI: <https://doi.org/10.1371/journal.pgen.1000222>, PMID: 18927627
- Sparacio AP,** Trojanowski NF, Snetselaar K, Nelson MD, Raizen DM. 2020. Teething during sleep: ultrastructural analysis of pharyngeal muscle and cuticular grinder during the Molt in *Caenorhabditis elegans*. *PLoS ONE* **15**:e0233059. DOI: <https://doi.org/10.1371/journal.pone.0233059>, PMID: 32433687
- Stenvall J,** Fierro-González JC, Swoboda P, Saamarthy K, Cheng Q, Cacho-Valadez B, Arnér ESJ, Persson OP, Miranda-Vizuete A, Tuck S. 2011. Selenoprotein TRXR-1 and GSR-1 are essential for removal of old cuticle during molting in *Caenorhabditis elegans*. *PNAS* **108**:1064–1069. DOI: <https://doi.org/10.1073/pnas.1006328108>, PMID: 21199936
- Sviben S,** Spaeker O, Bennet M, Albéric M, Dirks J-H, Moussian B, Fratzi P, Bertinetti L, Politi Y. 2020. Epidermal cell surface structure and chitin-protein co-assembly determine fiber architecture in the locust cuticle. *ACS Applied Materials & Interfaces* **12**:25581–25590. DOI: <https://doi.org/10.1021/acsami.0c04572>, PMID: 32343541
- Tan Y,** Hoon S, Guerette PA, Wei W, Ghadban A, Hao C, Miserez A, Waite JH. 2015. Infiltration of chitin by protein coacervates defines the squid beak mechanical gradient. *Nature Chemical Biology* **11**:488–495. DOI: <https://doi.org/10.1038/nchembio.1833>, PMID: 26053298
- Telford MJ,** Bourlat SJ, Economou A, Papillon D, Rota-Stabelli O. 2008. The evolution of the ecdysozoa. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **363**:1529–1537. DOI: <https://doi.org/10.1098/rstb.2007.2243>, PMID: 18192181

- Thompson JD**, Higgins DG, Gibson TJ. 1994. 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**:4673–4680. DOI: <https://doi.org/10.1093/nar/22.22.4673>, PMID: 7984417
- van der Lee R**, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu RV, Tompa P, Uversky VN, Wright PE, Babu MM. 2014. Classification of intrinsically disordered regions and proteins. *Chemical Reviews* **114**:6589–6631. DOI: <https://doi.org/10.1021/cr400525m>, PMID: 24773235
- Vassar PS**, Culling CF. 1959. Fluorescent stains, with special reference to amyloid and connective tissues. *Archives of Pathology* **68**:487–498 PMID: 13841452.
- Vernon RM**, Forman-Kay JD. 2019. First-Generation predictors of biological protein phase separation. *Current Opinion in Structural Biology* **58**:88–96. DOI: <https://doi.org/10.1016/j.sbi.2019.05.016>, PMID: 31252218
- White JG**, Southgate E, Thomson JN, Brenner S. 1986. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **314**:1–340. DOI: <https://doi.org/10.1098/rstb.1986.0056>, PMID: 22462104
- Willis JH**. 2010. Structural cuticular proteins from arthropods: annotation, nomenclature, and sequence characteristics in the genomics era. *Insect Biochemistry and Molecular Biology* **40**:189–204. DOI: <https://doi.org/10.1016/j.ibmb.2010.02.001>, PMID: 20171281
- Wojciechowski JW**, Kotulska M. 2020. Path-prediction of amyloidogenicity by threading and machine learning. *Scientific Reports* **10**:7721. DOI: <https://doi.org/10.1038/s41598-020-64270-3>, PMID: 32382058
- Wootton JC**, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry* **17**:149–163. DOI: [https://doi.org/10.1016/0097-8485\(93\)85006-X](https://doi.org/10.1016/0097-8485(93)85006-X)
- Wozniak PP**, Kotulska M. 2015. AmyLoad: website dedicated to amyloidogenic protein fragments. *Bioinformatics* **31**:3395–3397. DOI: <https://doi.org/10.1093/bioinformatics/btv375>, PMID: 26088800
- Wright KA**, Thomson JN. 1981. The buccal capsule of *Caenorhabditis elegans* (Nematoda: rhabditoidea): an ultrastructural study. *Canadian Journal of Zoology* **59**:1952–1961. DOI: <https://doi.org/10.1139/z81-266>
- Wu Y**, Wu Z, Butko P, Christen Y, Lambert MP, Klein WL, Link CD, Luo Y. 2006. Amyloid-beta-induced pathological behaviors are suppressed by Ginkgo biloba extract EGb 761 and ginkgolides in transgenic *Caenorhabditis elegans*. *The Journal of Neuroscience* **26**:13102–13113. DOI: <https://doi.org/10.1523/JNEUROSCI.3448-06.2006>, PMID: 17167099
- Wu C**, Scott J, Shea JE. 2012. Binding of Congo red to amyloid protofibrils of the Alzheimer A β (9-40) peptide probed by molecular dynamics simulations. *Biophysical Journal* **103**:550–557. DOI: <https://doi.org/10.1016/j.bpj.2012.07.008>, PMID: 22947871
- Zhang Y**, Foster JM, Nelson LS, Ma D, Carlow CKS. 2005. The chitin synthase genes *chs-1* and *chs-2* are essential for *C. elegans* development and responsible for chitin deposition in the eggshell and pharynx, respectively. *Developmental Biology* **285**:330–339. DOI: <https://doi.org/10.1016/j.ydbio.2005.06.037>, PMID: 16098962
- Zhang W**, Watanabe R, Konishi HA, Fujiwara T, Yoshimura SH, Kumeta M. 2020. Redox-sensitive cysteines confer proximal control of the molecular crowding barrier in the nuclear pore. *Cell Reports* **33**:108484. DOI: <https://doi.org/10.1016/j.celrep.2020.108484>, PMID: 33326779

RESEARCH ARTICLE

Exploring a diverse world of effector domains and amyloid signaling motifs in fungal NLR proteins

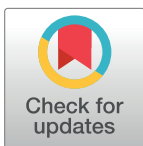
Jakub W. Wojciechowski¹✉, Emirhan Tekoglu^{2,3}✉, Marlena Gąsior-Głogowska¹, Virginie Coustou⁴, Natalia Szulc¹, Monika Szefczyk⁵, Marta Kopaczyńska¹, Sven J. Saupe^{4*}, Witold Dyrka^{1*}

1 Katedra Inżynierii Biomedycznej, Wydział Podstawowych Problemów Techniki, Politechnika Wrocławska, Wrocław, Poland, **2** Biyomühendislik Bölümü, Y İd z Teknik Üniversitesi, İstanbul, Turkey, **3** Wydział Chemiczny, Politechnika Wrocławska, Poland, **4** Institut de Biochimie et de Genetique Cellulaire, UMR 5095 CNRS, Université de Bordeaux, Bordeaux, France, **5** Katedra Chemii Bioorganicznej, Wydział Chemiczny, Politechnika Wrocławska, Wrocław, Poland

✉ These authors contributed equally to this work.

✉ Current address: Koc University, School of Medicine, İstanbul, Turkey

* sven.saupe@ibgc.cnrs.fr (SJS); witold.dyrka@pwr.edu.pl (WD)



OPEN ACCESS

Citation: Wojciechowski JW, Tekoglu E, Gąsior-Głogowska M, Coustou V, Szulc N, Szefczyk M, et al. (2022) Exploring a diverse world of effector domains and amyloid signaling motifs in fungal NLR proteins. *PLoS Comput Biol* 18(12): e1010787. <https://doi.org/10.1371/journal.pcbi.1010787>

Editor: William Stafford Noble, University of Washington, UNITED STATES

Received: February 10, 2022

Accepted: December 2, 2022

Published: December 21, 2022

Copyright: © 2022 Wojciechowski et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript, its [Supporting information files](#), and at Zenodo (doi: [10.5281/zenodo.7352382](https://doi.org/10.5281/zenodo.7352382)).

Funding: WD, MG-G were supported by the Narodowe Centrum Nauki (ncn.gov.pl) grant no. 2019/35/B/NZ2/03997. WD was also supported by the Wrocławskie Centrum Sieciowo-Komputerowe, Politechnika Wrocławska (wcss.pl), grant no. 98. NS was supported by the Narodowe Centrum

Abstract

NLR proteins are intracellular receptors constituting a conserved component of the innate immune system of cellular organisms. In fungi, NLRs are characterized by high diversity of architectures and presence of amyloid signaling. Here, we explore the diverse world of effector and signaling domains of fungal NLRs using state-of-the-art bioinformatic methods including MMseqs2 for fast clustering, probabilistic context-free grammars for sequence analysis, and AlphaFold2 deep neural networks for structure prediction. In addition to substantially improving the overall annotation, especially in basidiomycetes, the study identifies novel domains and reveals the structural similarity of MLKL-related HeLo- and Goodbye-like domains forming the most abundant superfamily of fungal NLR effectors. Moreover, compared to previous studies, we found several times more amyloid motif instances, including novel families, and validated aggregating and prion-forming properties of the most abundant of them *in vitro* and *in vivo*. Also, through an extensive *in silico* search, the NLR-associated amyloid signaling was identified in basidiomycetes. The emerging picture highlights similarities and differences in the NLR architectures and amyloid signaling in ascomycetes, basidiomycetes and other branches of life.

Author summary

All living organisms possess an immune system allowing them to cope with pathogens and, more broadly, to manage interactions with other organisms. One of its conserved components are the so-called NLR proteins, which are found in bacteria, plants, animals and fungi. NLRs are intracellular sensors that trigger a host response upon the detection of non-self markers, which is typically performed by effector domains of NLRs. We

Badań i Rozwoju (ncbr.gov.pl) project no. POWR.03.02.00-00-1003/16. VC, SJS were supported by the Agence Nationale de la Recherche (anr.fr) grant no. SFAS R-17-CE11-0035. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

investigate the repertoire of these domains in almost 500 fungal strains. We identify several major effector classes, most of which are involved in regulated cell death. Some NLRs do not have built-in effector domains but instead activate separate effector proteins via prion-like signal propagation. This activation is triggered by passing the amyloid fold from a short signaling domain on the NLR to its counterpart on the effector. Using innovative computational approaches, we identify new amyloid signaling motifs and find them overall several times more common in fungal NLRs than previously reported, including the evidence of amyloid signaling in basidiomycetes. Our results describe the global ensemble of NLRs effector domains in fungi and thus enhance our comparative view of this nearly-universally conserved immune receptor family.

Background

NLR proteins

All living organisms possess an immune system allowing them to cope with viral or cellular pathogens. Among the central and conserved components of the innate immune system in animals and plants are the NLR proteins. NLRs are intracellular immune receptors that induces various host responses including regulated cell death upon the detection of non-self cues [1–3]. A typical NLR protein functions following a ligand-induced oligomerization and activation process. Its tripartite domain architecture displays 1) a central Nucleotide-binding and Oligomerization Domain (NOD), 2) a C-terminal domain composed of superstructure forming repeats that is typically involved in detection of non-self cues in the form of DAMPs or MAMPs (Damage- or Microbe-Associated Molecular Patterns) and 3) a N-terminal effector domain whose activation induces various downstream host responses including regulation of the infected cell death [4–8]. While historically, NLRs were mostly studied within the animal and plant kingdoms (as Nod-Like Receptors and NBS-LRR Receptors respectively) [9, 10], their homologs were identified in bacteria and fungi [4, 11–13].

In fungi, homologs of NLR proteins were initially identified in the context of the study of a non-self recognition process termed heterokaryon incompatibility [14]. This reaction occurs in filamentous fungi in the event of the fusion (anastomosis) of the hyphae of genetically incompatible individuals, resulting in the death of mixed fusion cells [15, 16]. Incompatibility prevents in particular the transmission of mycoviruses between isolates during the anastomosis events. In *Podospora anserina*, HET-E, one of the proteins controlling heterokaryon incompatibility is a homolog of NLR proteins (although its N- and C-terminal domains differ from those known in animals and plants, a situation typical for NLR architecture proteins outside of the plant and animal kingdom [4, 11, 17]). Its central NOD domain is one of the original founding members used to define the NACHT domain (Pfam PF05729) common in animal NLRs (the H in the NACHT acronym stands for HET-E) [10, 18]. The C-terminal domain of HET-E protein, built of hypervariable WD40 repeats recognizes a non-self cue, here polymorphic variants of a host protein termed HET-C, a glycolipid transfer protein universally conserved in eukaryotes that could represent a pathogen effector target [19]. In such event, the N-terminal HET domain of the HET-E protein is activated which ultimately leads to regulated cell death [19]. The HET domain (PF06985) [18] is a cell death inducing domain with a remote homology to TIR domains [20, 21], including conservation of a functionally relevant glutamate [11, 22]. Several other fungal cell death inducing incompatibility pathways in *Podospora* and other species are controlled by NLR proteins [5, 23]. Yet, apparently only a small fraction of the existing fungal NLRs are involved in heterokaryon incompatibility and it is proposed that

these proteins have more general functions in immune defense and establishment of symbiotic interactions in fungi [5, 24]. Indeed, NLR proteins are abundant in multicellular filamentous fungi (no NLR protein was found in unicellular yeasts). In a recent study, a total of about 36 000 NLR proteins have been found in around 880 strains of over 560 species of fungi with on average 57 NLRs per genome and numerous species displaying hundreds of NLR genes [5, 11].

In terms of domain annotation fungal NLRs differ from their typical animal and plant counterparts. Unlike more homogenous NLR proteins in animals and plants, the central domain of fungal NLRs can be either of the NACHT [10] or the NB-ARC type (PF00931) [9]. Then fungal NLRs display ankyrin repeats (ANK, Pfam CL0465), tetratricopeptide repeats (TPR, CL0020) and beta-propellers of the WD40 meta-family (CL0186) in place of the LRR repeats found in most animal and plant NLRs. The NBS-TPR architecture was proposed to correspond to the ancestral architecture whilst NLR proteins in multicellular bacteria also typically display TPR, ANK or WD repeats [4, 11, 12, 17]. Consistent with a role in immune defense C-terminal repeated domains of fungal NLRs display marks of positive selection and are highly variable [11, 23, 25]. In addition, the C-terminal domains show original modes of functional diversification. First, about 1/6 of these C-terminal repeat domains consist of highly similar repeats with only a few highly variable positions under positive selection [11, 26]. These repeats arrays with high internal similarity are hypervariable loci in which individual repeats are exchanged and reshuffled resulting in functional diversification [25, 26]. High internal similarity of repeats is both a cause and a result of an unequal crossing over mechanism, a process which is 5–6 orders of magnitude faster than the point mutation [27]. Then, in the truffle *Tuber melanosporum* a superfamily of NACHT-ANK NLR encoding genes displays dozens of 3 bp mini-exons whose alternative splicing can considerably diversify the repertoire of potential C-terminal recognition domain [28]. These striking modes of recognition domain diversification are consistent with the proposed role of NLR proteins in the immune response, as capability of quickly adapting to evolving pathogens is a condition of success in the constant arms race against them [25].

For about 50% of fungal NLR proteins, N-terminal domain annotations could be determined with the Pfam [29] and similar HMM profiles [11], which make up for 12–13 major meta-families [5, 11]. Functionally, the characterized N-terminal domains belong to three basic types: enzymatic, signaling, and regulated cell death induction [30]. The four largest families of fungal NLR effectors are the Alpha/Beta hydrolases [31], the purine and uridine phosphorylases [32, 33], both associated with enzymatic functions, pore-forming domains homologous to HeLo [34–38], and functionally and structurally uncharted Goodbye homologs [11, 37]. The first three families are widespread in various branches of life. For example, the HeLo domain is a fungal homolog of human MLKL, plant RPW8 and bacterial Bell domains [12, 37, 38]. It is understood that upon oligomerization, these domains, whose central part is a four-helix bundle, expel a N-terminal alpha-helix to form a pore targeting the membrane and thus induce cell death [39, 40]. Out of 72 theoretically possible NLR architectures made with the most common domain families (12 types of N-terminal domains, 2 types of central domains and 3 clans of C-terminal domains), as many as 32 were identified in fungal proteomes [11]. Interestingly, in about 20 cases, the closest orthologs of the central domain sequences were bound to different N-terminal domains (including in two different strains of the same species). Moreover, the maximum-likelihood phylogenetic trees generated separately for the N-terminal and central domains were mutually incompatible, and distribution of the N-terminal domains over the branches of central domains trees generated for selected species was scattered. Together with a relatively high number of NLRs without ortholog in other strains of the same species, these findings indicate high plasticity of the architecture of NLR proteins and the occurrence of the *death-and-birth evolution* process [5, 11].

Amyloid signaling motifs

Another notable feature of fungal NLRs is the occurrence of amyloid-forming motifs at their N-termini [30]. A series of studies derived from the characterization of the *Podospora anserina* [Het-s] prion protein, which controls regulated cell death in the context of heterokaryon incompatibility, has revealed that a fraction of the fungal NLRs employ amyloid signaling to activate downstream cell death effector domains [30, 41]. The paradigmatic example of such amyloid NLR signalosomes is the HET-S/NWD2 two-component system of *P. anserina*. HET-S encodes a cell death execution protein with a globular N-terminal HeLo domain (PF14479) and a C-terminal amyloid forming prion domain composed of two elementary repeats r1 and r2 which are able to adopt a specific β -solenoid amyloid fold [36, 42–44]. Amyloid transconformation of the C-terminal domain induces activation of the HeLo domain, which turns into a pore-forming toxin. NWD2 is a NLR, encoded by the gene immediately adjacent to *het-S*, and displays at its N-terminus a motif termed r0 which is homologous to the elementary r1 and r2 repeats [37, 41]. When activated by their cognate ligand, engineered variants of NWD2 are capable of triggering transconformation of HET-S and to induce its toxicity. In this system, activation of the NLR leads to amyloid folding of its N-terminus which then serves as template to activate a cognate cell death execution protein [30]. Throughout this paper the term amyloid signaling refers to passing information from one protein to another by transmitting the amyloid fold due to the compatibility of amyloid motifs [30].

The r0, r1, and r2 motifs, collectively referred to as the HET-s *motif*, represent one of the best studied examples of an amyloid signaling motif (ASM). Homologs of the HET-s motif can be grouped in 5 subclasses (collectively denoted as HET-s Related Amyloid Motifs or HRAM) [45], which co-occur in N-termini of fungal NLR proteins and in C-termini of HeLo [34–36] and HeLo-like (PF17111) proteins [11, 37] encoded by genes adjacent to NLR-encoding genes in the genome. In some organisms, two or three subclasses of HRAMs exist simultaneously, which allows for maintaining distinct signaling pathways [45, 46].

There are two other families of fungal ASMs with similar functionality in the NLR protein system, namely σ (named after the σ prion, which contains this motif [47]) and PP (pseudopalindromic due to the amino acid pattern N×G×Q×G×N at its core) [37]. The PP motif bears significant resemblance to the mammalian RHIM motif [38, 48, 49] with remote homologs also in multicellular bacteria [12].

Still, this repertoire of already described fungal ASMs is significantly smaller in comparison to bacterial amyloid signaling motifs. A recent *in silico* analysis of over 100,000 available bacterial genomes in search of sequence motifs repeated in adjacent genes encoding the Bell (bacterial homolog of fungal HeLo) and NLR proteins revealed ten families of Bacterial Amyloid Signal Sequences (BASS) widespread in multicellular Actinomycetes, Cyanobacteria and in Archaea [12]. Despite their sequence-level diversity, at least some if not all known bacterial and fungal ASMs are believed to share the beta-arch fold [50–52].

While it is not fully understood why the NLR/effector pairs involving amyloid signaling are generally encoded by clustered genes, the same situation has been recently reported in regulated cell death pathways involving protease/gasdermine clustered gene pairs [53, 54]. The most likely explanation for this genomic clustering relates to genetic inheritability of such clusters. Genetic association of the genes encoding the receptor and effector moiety of the cell death pathway favors both its vertical (meiotic) and horizontal (transposition driven) inheritance of the pathway as a whole. There is evidence that NLRs in fungi can be preferentially associated with and carried by transposons [55].

When compared to the NLR proteins in plant and animal kingdoms, the fungal NLR proteins display larger diversity of architectures. In addition, NLR-associated amyloid signaling

appears specific to fungal and bacterial kingdoms although amyloid motifs also occur in immune pathways in animals [56, 57]. The dominant view, until recently, was that the architecture and immunological function of NLR proteins in plants and animals resulted from the convergent evolution [17]. However, higher diversity of NLRs in fungi than in animals and plants, as well as presence of NLRs in prokaryotes [4, 12, 13] suggest the early evolutionary origins of the architecture and the immune function of NLR proteins [5, 30]. Exploration of the diversity of fungal NLRs is an important asset for deciphering of the potential roles of these immune receptors in fungal biology in addition to their documented role in cell death related to incompatibility. In addition, comparative studies of NLRs in the different kingdoms can provide a more global view of the long term evolution of these central components of immunity in both microbes and macro-organisms. The aim of the current study is to improve the annotation and characterization of the vast ensemble of N-terminal domain of fungal NLRs with particular emphasis on short domains (shorter than 150 amino acids) and amyloid-like motifs.

Results

Overview of N-terminal domains of fungal NLRs

In roughly 36 000 fungal NLRs identified in a previous study [12], over 90% proteins had N-terminal extension to the NOD domain at least 20 amino-acids long and therefore capable to accommodating a functional domain (Fig 1a). Only 57% of them was previously annotated using the Pfam [29] or inhouse profiles [11]. To improve the Pfam annotation coverage, we clustered the set of N-termini with MMseqs2 [58] and then, for each cluster with at least 20 members, searched for homologs in UniRef30 [59, 60] and subsequently in Pfam using HHblits [61] (see Computational methods for details). The procedure resulted in assigning the Pfam-based annotations to 3003 additional N-termini, thus increasing the annotation coverage to 66%.

N-terminal annotations of fungal NLRs are not evenly distributed. The length distribution of N-termini varied significantly with regard to the fungal phylum (Fig 1b and Fig A in S1 Text): while Basidiomycota were over-represented among short N-termini (below 100 amino acids), Ascomycota made up for 85% of termini longer than 200 amino acids. The Pfam annotation coverage was also not evenly distributed. While almost 90% of longer N-terminal domains (200 aa or more) were at least partially annotated, the figure was below 40% for the middle range, and—not surprisingly—a few percent for domains shorter than 100 amino acids, which constituted 1/4 of all NLR N-termini (Fig 1c). The Pfam annotation coverage also strongly depended on taxonomic scope: N-termini from Ascomycota were more completely annotated (72%) than N-termini from Basidiomycota (23%), even though our new clustering-based annotation scheme increased coverage of the latter phylum roughly twice (Fig 1d). This inequality held as well when N-termini in the same length ranges (above 100 aa) were compared in both branches. In the clustering-based approach, Pfam annotations were found for more than 80% N-termini with the UniRef homologs outside the Fungi kingdom, but only for around 20% sequences with fungal-only homologs (Fig 1e). While better coverage of more universally spread domains is not surprising, taken together, our results highlight the fact that the NLRs of fungi, and especially Basidiomycota, are still not sufficiently represented in Pfam.

Novel annotations include the ubiquitin, TIR, and purine nucleoside phosphorylase domains. The updated annotations of fungal NLR N-termini were summarized in Fig 1f and in Fig B in S1 Text. Vast majority of newly added annotations belonged to domain families already described as fungal NLR effectors (Table A in S1 Text). The exceptions were the Crinkler domain of the Ubiquitin clan only recently included in Pfam [62–66], and the Sterile Alpha

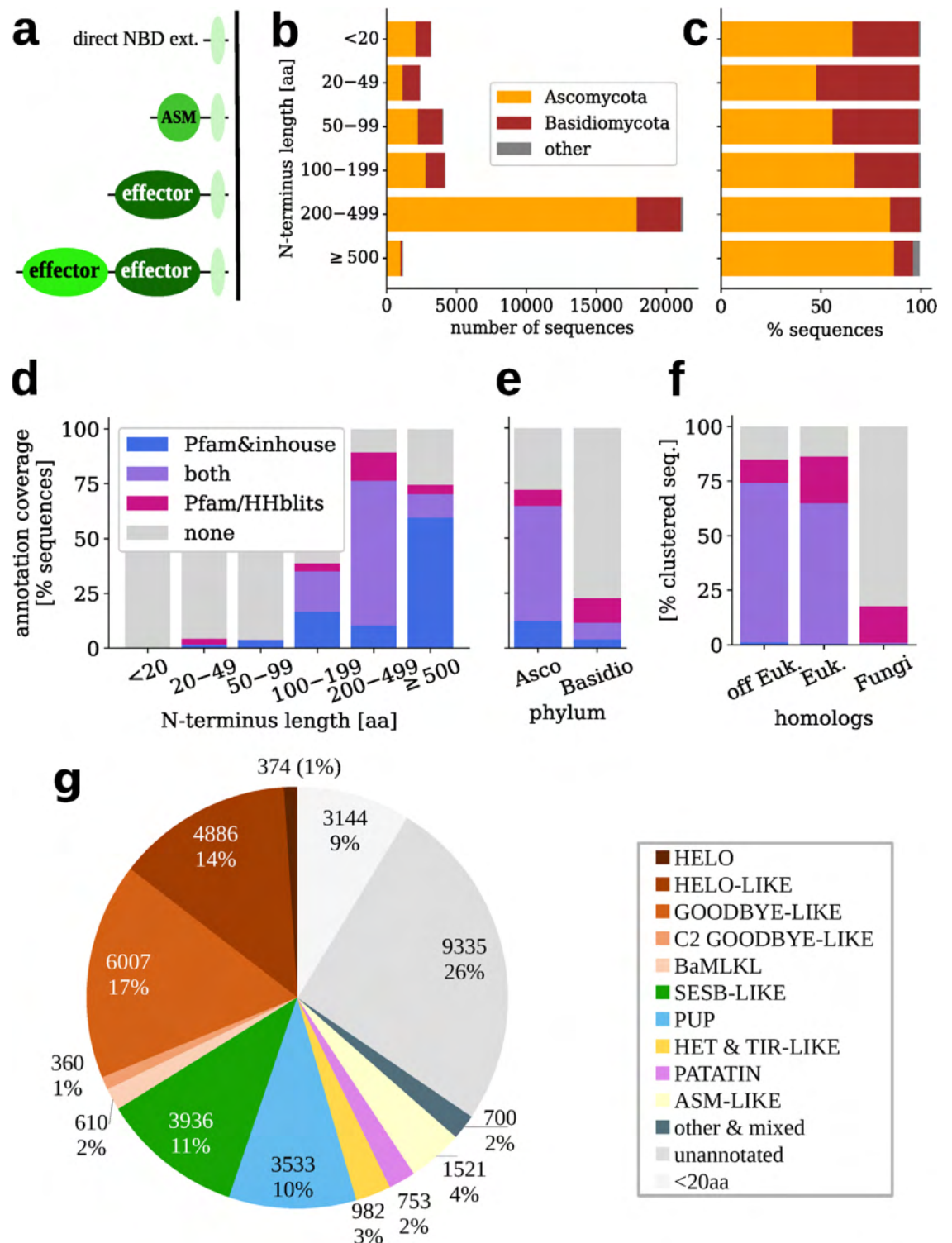


Fig 1. Fungal NLR N-termini. a) Major general architectures; b) N-termini length distribution with taxonomic division, c) Same data scaled to 100% for each length range; Annotation coverage with regard to d) N-terminus length, e) taxonomic division; f) Annotation coverage of the MMSeqs clustered N-termini with regard to presence of taxonomically distant homologs in UniRef top hits (see Results). Euk. denotes *Eukaryota*, off Euk. category includes *Bacteria*, *Archaea* and *Viruses*. Colored bars indicate fraction of Pfam & inhouse annotated sequences (blue: only direct Pfam hits, violet: direct and with clustering & HHblits, rose: only with clustering & HHblits).

Inhouse profiles were used only for direct Pfam searches. g) Distribution of domain families. Additional non-Pfam annotations included, see [Results](#) and [Methods](#). N-termini shorter than 20 amino acids are distinguished, as unlikely to contain functional domains.

<https://doi.org/10.1371/journal.pcbi.1010787.g001>

Motif family SAM_Ste50p [67]. SAMs are involved in homologous and heterologous protein-protein interactions [68], notably they are present in SARM1 protein of the Toll-Interleukin-1 Receptor (TIR) family [69–71]. Moreover, the new scheme increased the number of the Purine and Uridine Phosphorylase (PUP) superfamily annotations, mostly due to the matches to the purine Nucleoside Permease (NUP) profile [72]. In addition, dozens of Pezizomycotina species contained NLR N-termini comprising of C-terminal part of the PNP_UDP_1 fold (cf. pdb:6po4B, residues 176–234). A large number of agaricomycetal N-termini displayed the double domain C2 Goodbye-like architecture [11, 37], the architecture which was specific to Agaricomycetes. The Goodbye-like domain was found also in other double domain architectures of NLR N-termini (Table A in [S1 Text](#)). Please refer to [S1 Text](#) for additional notes on the updated annotations.

Some effector domains are absent in basidiomycetal NLRs. Overall, several most abundant domain classes including the Goodbye-, HeLo-, SesB-like and PUP families, accounted for majority of fungal NLR N-termini ([Fig 1f](#)). The two latter superfamilies were common in ascomycetal NLRs (13–14% each) but were almost (SesB-like) or completely (PUP) missing from basidiomycetal NLRs ([Fig B](#) in [S1 Text](#)). The complete lack of PUP (and HET) domains in basidiomycetal NLRs contrasted with the presence of these domains in other (non-NLR) domain architectures in this division.

Relation between HeLo-, Goodbye- and basidiomycetal MLKL-likes

HeLo- and Goodbye-like annotations overlap in basidiomycetal homologs of human MLKL. Notably, we found clusters with apparently overlapping HeLo/HeLo-like and HeLo-like/Goodbye-like domain annotations. The latter situation was found in Basidiomycota and mostly involved sequences annotated as MLKL_NTD according to Conserved Domain Database (CDD) [73]. Moreover, there were additional basidiomycetal clusters with CDD MLKL_NTD annotation and/or with Pfam HeLo- or Goodbye-like annotations just below the assignment threshold, surmounting to a total of 600 basidiomycetal MLKL-like (BaMLKL) sequences. This made the superfamily of Goodbye/HeLo/MLKL_NTD-like domains the most frequent in Basidiomycota (nearly 2000 sequences, 23% of all), similarly to Ascomycota (10 000 sequences or 38%, [Fig B](#) in [S1 Text](#)).

We analyzed the largest cluster with the overlapping Goodbye-like and HeLo-like annotations assigned through the HHblits-based procedure (OBZ65626, 106 sequences). Several sequences in the cluster received also hits from various MLKL-related Pfam profiles when sequences were searched individually (sequence and domain E-values of $1e-3$, [Fig 2a](#)). Not surprisingly, the multiple sequence alignment of the cluster closely matched (HHpred [74, 75] probability above 98%) the sequence of human MLKL executioner domain with an experimentally solved three-dimensional structure (pdb:6vzo [76], [Fig 2b](#)). In fact the MLKL domain was almost perfectly aligned with the HeLo_like_N profile match, while the related SesA profile match was slightly shorter. At the same time, the matches to the two Goodbye-like profiles, Goodbye and NACHT_N [11], were both shifted N-terminally with regard to the MLKL-like domain resulting in a partial overlap, significantly longer for NACHT_N. Importantly, the multiple sequence alignment was well conserved for the combined stretch of Goodbye- and HeLo-like matches regardless of Pfam annotations of individual sequences ([Fig 2a](#)).

HeLo-, Goodbye- and basidiomycetal MLKL-like proteins share a core structural fold. Then, we attempted structure prediction for the largest MLKL-like clusters using AlphaFold2

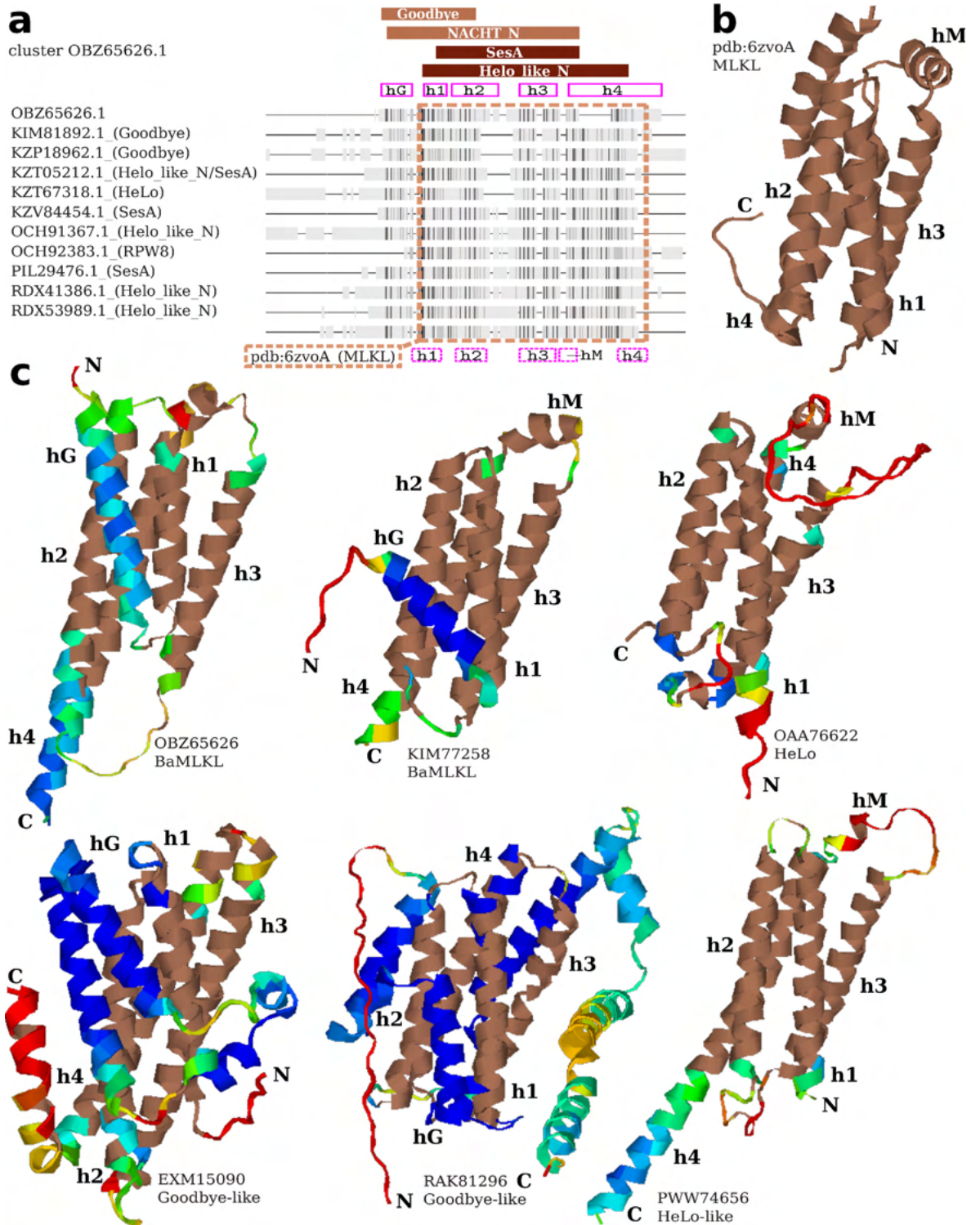


Fig 2. MLKL-like N-termini. a) Fingerprint alignment of the doubly (Goodbye-like & Helo-like) annotated OBZ65626 cluster including non-redundant sequences with direct Pfam annotations. The alignment was truncated C-terminally. Darker shade implies higher conservation, while gaps are represented as lines. Columns matched with Pfam profiles of MLKL-like domains are indicated with brown bars. Columns corresponding to helices in a predicted OBZ65626 model are indicated with solid magenta boxes. Columns alignable to the human MLKL structure are framed with a brown dashed line. Columns corresponding to helices in the aligned MLKL structure are indicated with dashed

magenta boxes. **b**) The human MLKL structure (pdb:6zvoA). **c**) Structural models of various MLKL-like domains predicted with AlphaFold2 (see [Methods](#)). Regions aligned to the human MLKL structure with TM-align are shown in brown. Rainbow colors indicate model quality in terms of pLDDT (below or 50: red, 60: yellow, 70: green, 80: cyan, above 90: blue).

<https://doi.org/10.1371/journal.pcbi.1010787.g002>

[77] through the ColabFold advanced notebook [78]. The predictions were carried out solely using multiple sequence alignments of each cluster. Except for the largest HeLo-like cluster, all other predictions resulted in very good quality models (pLDDT around 0.80) sharing a four-helix core (Fig 2c), which is characteristic to the solved MLKL structure. When aligned to the latter using TM-align [79], the predicted models achieved TM-scores between 0.51 to 0.64. The four-helix bundle configuration was supported with alignment conservation scores, calculated with ConSurf [80, 81], which were consistently high for residues facing the interior of the bundle (Fig C in [S1 Text](#)). The most notable difference between structural models obtained for various clusters was an additional N-terminal helix in basidiomycotal MLKL_NTD homologs and Goodbye-like (hG in Fig 2), not found in MLKL and HeLo-like. However, Goodbye-like models presented longer and more complex N-terminal extension than BaMLKLs. Noteworthy was the relatively high conservation of hG residues facing the bundle and h1 residues facing the exterior of the bundle (Fig C in [S1 Text](#)). Also, Goodbye-like lacked a short perpendicular helix (hM) between helices h3 and h4, which seemed to be a common feature of human and basidiomycotal MLKLs and HeLo-like (Fig 2c).

Taken together, these analyses indicate that although Goodbye-like profiles share a core region with the MLKL bundle and HeLo and HeLo-like profiles, they also differ by the presence of an N-terminal extension ahead of the region corresponding to the first helix in MLKL/RPW8/HeLo proteins. Considering the critical role of this region in the oligomerization, membrane targeting and ion specificity of these animal, plant and fungal proteins, further experimental investigation are needed before a potential cell death inducing activity can be firmly attributed to Goodbye-like profiles [39, 40, 44, 82, 83].

Unannotated longer N-termini

A novel helical effector domain is shared between Pezizomycotina and Mortierellomycetes. In addition, largest unannotated clusters were carefully examined and subjected to structural modeling using AlphaFold2 [77, 78] (see Computational methods). The identified domains were listed Table B in [S1 Text](#) and briefly characterized in [S1 Text](#). Notably two clusters, mutually homologous, consisted of relatively long domains (N-terminal length above 500 aa) from Pezizomycotina and Mortierellomycetes predicted to be made of multiple alpha-helices forming two stretches of the alpha solenoid-like structure (NLR_Helical in Table B and Fig Dab in [S1 Text](#)). Interestingly, homologous domains were also found in bacteria, mainly in *Mycoavidus cysteinexigens*. As this betaproteobacteria is an endosymbiont of *Linnemannia (Mortierella) elongata* AG-77 (a fungus with the largest number of these proteins [2]), this may suggest possibility of the horizontal gene transfer.

TIR-like effectors are present in Pezizomycotina. Another unannotated cluster consisted of moderately long NLR N-termini (median length of 389 aa), from various Pezizomycotina species, which partially resembled the SEFIR family [84, 85] of TIR clan. A good quality structural model predicted with AlphaFold2 supported homology to TIR and HET domains (Fig Dc in [S1 Text](#)). Importantly, the TIR domain was reported in NLRs from plants, bacteria and Chytridiomycota [5, 12, 21, 86]. Interestingly, homologous domains were also present as separate proteins in Mucormycota *Rhizophagus irregularis*, a species related to *Mortierella*, and in *Mycoavidus cysteinexigens*, in accordance with the possibility of horizontal gene transfer [87].

Specialized effector domains are abundant in fungal NLRs. Importantly, all other longer domains were represented by less than 100 sequences. With the limitation that in some cases larger families may have been superficially partitioned into small clusters, this indicates that the current Pfam annotations (plus MLKL_NTD and a few inhouse profiles) cover all widely spread abundant domains. At the same time, there seems to exist a substantially large corpus of thousands of specialized N-termini, sometimes confined to narrow taxonomic branches. While some of them may be formed with a tuple of known domains, other could represent novel families (likely being difficult targets for structure prediction due to small alignments). With regard to our previous analyses [5, 11], the current study suggests less diversity in major effector classes (5–7 rather than 12–13), but highlights a likely abundance of specialized domains.

Amyloid-like motifs in short N-termini

A novel *in silico* approach finds amyloid-like motifs in 1/6 of all short NLR N-termini. The largest deficiency in the annotation coverage concerned short N-terminal sequences (length below 150 amino acids). Only less than six percent of them (645 out of 11 634) received any Pfam-based annotation, while less than two percent (214) was annotated as so called prion-forming domains (PFD) [11, 37], consisting of the three known families of fungal ASMs. As more than 3 000 short N-terminal domains were assigned to clusters made with at least 20 sequences, this suggested presence of conserved sequential features. Therefore, we searched for potential additional fungal amyloid signaling motifs using an approach that combined filtering with a probabilistic grammatical model inferred from ten families of bacterial ASMs (BASS1–10 [12]), shown to be sensitive to fungal amyloid signaling sequences [52], with the MEME motif extraction [88] (see Computational methods for details). The procedure resulted in identifying 16 grammar-compatible motifs (Fig 3a). Then, we used profile HMMs of these motifs to scan all NLR N-termini at least 10 amino-acids long, and found hits in 1537 sequences (Table 1), which represented 17% (36%) of all (clustered) short N-termini. The number included 204 out of all 242 sequences already annotated as PFD-LIKEs (84% sensitivity).

Amyloid-like motifs in fungal NLRs cluster to nine classes likely assuming the beta-arch fold. Not surprisingly, some of the 16 motifs clearly corresponded to the three fungal ASM families: HRAM (NLR13, found in 131 sequences), PP (NLR07, 296), and σ (NLR28, 71). The overall recall of 498 hits was twice higher in comparison to the combined Pfam-based approaches (242). Several hits of another two motifs, NLR12 and NLR40, overlapped with the NLR13 (HRAM) matches (Fig 3b). Moreover, the HMM scan with a generalized HRAM profile based on HRAM dataset from [45] recognized 27/51 NLR12 and 14/22 NLR40 motifs, thus indicating that these two classes were related to HRAM. Indeed, the NLR12 motif (Fig 3a) is apparently similar to HRAM3 [45]. In addition, the G-hydrophobic-Q-hydrophobic-G pattern of NLR39 motif resembled NLR07 (PP). Five other motifs (NLR17, NLR19, NLR20, NLR32 and NLR34, in 138 sequences altogether) were difficult to assign to the known families. The final and the largest subgroup (689 sequences) consisted of five motifs (NLR05/08/22/29/44) with hits substantially overlapping NLR22 hits. This large group was specific to basidiomycetes except of a dozen of NLR22 hits overlapping ascomycotal NLR28 (σ) (Fig 3b). While most motifs were distributed in larger taxonomic branches, two motifs were more restricted: NLR17 was specific to *Amanita muscaria* (strain Koide) and NLR19 to genus *Tuber*. A combined NLR19 + NLR34 configuration was found in five highly homologous sequences from *Tuber melanosporum* (Fig 3b). All 16 motifs are likely to assume the beta-arch fold typical to known fungal and bacterial ASMs as from 45 to 95% motif instances passed the fold prediction

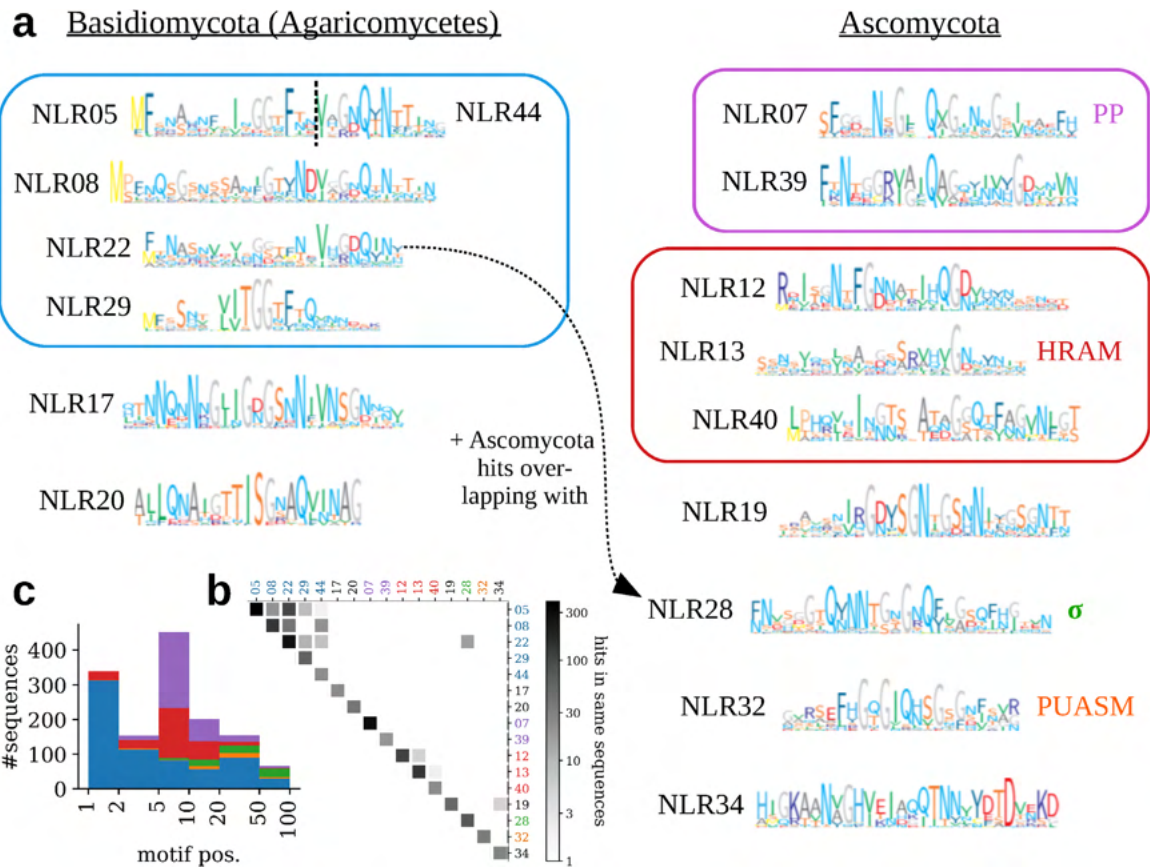


Fig 3. Amyloid-like motifs in short N-termini of NLRs. Clusters of N-termini containing sequences resembling bacterial amyloid signaling motifs were identified using a probabilistic grammatical model [52]. Motifs were extracted with MEME [88] and iteratively refined with profile HMMs. (a) Profile HMM-based motif logos—grouped according to overlapping hits in NLR N-termini, as shown in panel (b) Overlapping hits in NLR N-termini. See Results and Methods for details. (c) Stacked histogram of motif hits positions in NLR N-termini for the five largest motif families, color-coded as in panel (a).

<https://doi.org/10.1371/journal.pcbi.1010787.g003>

threshold of ArchCandy (column AC+ in Table 1). The only exceptions were two shortest motifs, NLR05 (28%) and NLR44 (none), probably because they comprise only parts of the actual amyloid-like motif (Fig 3a).

For four motifs, the amyloid signaling is supported by genomic co-localization of effectors. Significant numbers of similar sequence stretches in C-termini (100aa) of genomically neighboring (20kbp) proteins were found only for motifs representing the three fungal ASM families (NLR07 in 37 sequences, NLR12 in 4, NLR13 in 16, and NLR28 in 42) and for NLR32 (in 11 sequences). This suggests that NLR32 defines a new family of amyloid signaling motifs. (For further computational and experimental verification, see below).

Amyloid-like motifs differ in their position in NLR N-termini. While instances of the NLR05/22 group were usually situated in the very terminus, most HRAMs (NLR12/13/40) and PPs (NLR07/39) were located at positions 5–9. Moreover, NLR32 and σ motifs (NLR28) were shifted further C-terminally with relative majority at positions 20–49 and 50–99, respectively (Fig 3c). In addition, a couple of dozens of amyloid-like sequences of various families (including 17 NLR05 and 7 NLR07) were found located centrally or C-terminally in longer N-termini. Some of them formed combined architectures with annotated domains, most notably with

Table 1. Amyloid-like motifs in short N-termini of NLRs. Motif id indicates ranks in the MEME output. Motifs are grouped based on overlapping hits in NLRs and similar sequence patterns. Established and proposed motif annotation labels are given where applicable. L is the motif length. #NLR and #nei. indicate number of sequences with a given motif in short N-termini of NLRs and C-termini of their genomic neighbors, respectively. AC+ indicates a proportion of motif instances for which ArchCandy score is 0.56 or above. Major taxonomic branch including vast majority of NLRs with the motif is given. #eff. indicates total number of effector proteins (with established association to NLRs) [5] with a given motif in C-termini. #cooc. indicates number of sequences with a given motif in short N-termini of NLRs / C-termini of effector proteins cooccurring in the same strains (genome assemblies). #str. is a number of such strains with cooccurrence. Exp. indicates selected studies reporting experimental validation of some properties typical to ASM for a motif instance in a given family.

| Id | Annot. | L | #NLR | AC+ | Major tax. | #nei. | #eff. | #cooc. | #str. | Exp. |
|----|----------|----|------|------|--------------------|-------|-------|--------|-------|------------|
| 05 | — | 17 | 387 | 0.28 | Agaricomycetes | 0 | 0 | — | — | — |
| 08 | — | 30 | 138 | 0.55 | Agaricomycetes | 0 | 0 | — | — | — |
| 22 | — | 24 | 263 | 0.53 | Agaricomycetes | 0 | 0 | — | — | — |
| 29 | — | 20 | 60 | 0.53 | Agaricales | 0 | 0 | — | — | — |
| 44 | — | 12 | 24 | 0.00 | Agaricomycetes | 0 | 0 | — | — | — |
| 07 | PP | 23 | 296 | 0.64 | Ascomycota | 37 | 106 | 157/76 | 56 | [106] |
| 39 | PP | 24 | 20 | 0.80 | Ascomycota | 0 | 2 | 1/1 | 1 | — |
| 12 | HRAM3 | 26 | 110 | 0.48 | Sordariomycetes | 4 | 17 | 80/14 | 13 | — |
| 13 | HRAM | 24 | 131 | 0.57 | Ascomycota | 16 | 43 | 41/24 | 20 | [43, 46] |
| 40 | HRAM | 26 | 24 | 0.92 | Ascomycota | 0 | 0 | — | — | — |
| 17 | — | 26 | 24 | 0.88 | <i>A. muscaria</i> | 0 | 0 | — | — | — |
| 19 | — | 27 | 53 | 0.60 | <i>Tuber</i> | 0 | 0 | — | — | — |
| 20 | — | 21 | 37 | 0.68 | Agaricales | 0 | 0 | — | — | — |
| 28 | σ | 28 | 71 | 0.80 | Ascomycota | 42 | 62 | 45/43 | 40 | [52] |
| 32 | PUASM | 22 | 33 | 0.45 | Sordariomycetes | 11 | 29 | 25/21 | 18 | this study |
| 34 | — | 28 | 29 | 0.52 | Ascomycota | 0 | 0 | — | — | — |

<https://doi.org/10.1371/journal.pcbi.1010787.t001>

NLR_PRDR (NLR05 in 10 sequences from *A. bisporus*) and MLKL-likes (5 BaMLKL + NLR05 in *Laccaria bicolor*, 4 HeLo-like + NLR28 and 1 HeLo-like + NLR07 in various Ascomycota).

A reverse approach: Amyloid-like motifs in C-termini of effector proteins

Two novel amyloid-like motifs are uniquely associated with the PNP_UDP effector domain. In order to complement the search for amyloid signaling motifs in NLRs and verify discovery of the fourth NLR-related fungal ASM family, we adapted the approach recently used for identification of 10 families of bacterial ASMs in NLR-related proteins in bacteria [12]. The procedure, which also used MEME for motif extraction, started from known effector proteins [5], and relied on genomic proximity of their genes and genes encoding NLRs (see Computational methods for details). Consequently, we identified 22 motifs, and clustered them on the basis of their co-occurrence in 190 pairs of genomically neighboring proteins (Fig E in S1 Text). Three clusters clearly corresponded to the already known families PP, σ , and HRAM (Fig E in S1 Text). Two additional motifs with few pairs apparently resembled HRAM2 and HRAM4 [45], respectively. The fourth largest family of motifs exhibited a distinctive conserved pattern FxGxGxQxxGxGxF, which clearly corresponded to the NLR32 motif in Fig 3. Since in both searches the motif was found associated uniquely with the PNP_UDP domain, we termed it PUASM, or the Pnp_Udp-associated Amyloid Signaling Motif. The NLRs with the PUASM motif proteins were annotated either as NACHT or NACHT WD40. All matched instances of the PUASM motif came from various Pezizomycotina species. Finally, we found one more distinct motif related to PNP_UDP, however only present in four pairs (PF01048_015 in Fig E in S1 Text).

Amyloid-like motifs differ in the effector domain association. Overall, the ASM differed in type of associated effector domain, either pore-forming (HeLo and HeLo-like for

HRAM/NLR13), enzymatic (PNP_UDP for HRAM/NLR12, NLR32 and PF01048_015), or both (PP/NLR07 and σ /NLR28). Interestingly, while the NLR13 motif was typically found as a double in C-termini of HeLo and HeLo-like domains, for the second HRAM-like, NLR12, only single instances were found in C-termini of PNP_UDP_1 effector proteins. This may suggest a different mode of operation despite their similar sequence profiles. Notably, the occurrence of ASMs as single instances or two (or three) fold repeats was also reported for bacterial ASMs [12].

Amyloid signaling suspected between NLRs and effectors encoded by non-adjacent genes. To check the possibility that proteins cooperating through the amyloid signaling are encoded by non-adjacent genes, we analyzed co-occurrence of particular amyloid-like motifs in N-termini of NLRs and C-termini of established effector domains [5] in entire genomes. Non-singular C-terminal hits and genomic co-occurrences were found only for the three established fungal ASM families and PUASM (Table 1). Such cases were relatively most frequent for HRAM/NLR12, in parallel with the high ratio between the NLR-side and the effector-side motifs in some genomes (mean ratio 5.7:1, Table 1).

Amyloid-like motifs in Basidiomycota

Genome-wide motif searches suggest the NLR-related amyloid signaling in Agaricomycetes. With the NLR-related amyloid signaling previously described in multicellular bacteria and Ascomycota, apparent is the lack of evidence of this mechanism in Basidiomycota. On the other hand, we found numerous homologs of the pore-forming HeLo and HeLo-like domains in Basidiomycotal NLRs. Thus, we used them for searching the entire Basidiomycota genomes for homologs separate from NLR domains. We identified hundreds of such putative singular pore-forming domains, which—because of their potential to cause the cell death—can be expected to be under control of other proteins. As in Ascomycota such control is exerted by NLRs through the amyloid signaling sequences, we scanned the identified BaMLKL homologs against ASM profiles and grammars. However, fragments resembling ASMs were identified only in a few out of 500 sequences and in no case similar fragments were found in the neighboring NLRs. Yet in two cases pairs of amyloid-like motif instances occurred when entire genomes were considered (Fig 4a). In *Moniliophthora roreri* (strain MCA 2997) there was a 18 amino-acid long motif apparently shared between two BaMLKL C-termini and 26 short NACHT N-termini (Fig F in S1 Text). In addition, in *Fibularhizoctonia* sp. CBS 109695 there was a conserved pattern shared between two BaMLKL C-termini, eight short NLR N-termini (including KZP25847 with NLR20 instance), and additional five NLR proteins with the pattern situated between BaMLKL and NACHT domains (including KZP30127 and KZP3012 with NLR22 instances)—see alignment in Fig G in S1 Text. It would suggest a possibility that in *Fibularhizoctonia* proteins with the N-terminal and C-terminal amyloid-like sequences were pseudogenes, especially that three NLRs in this group were atypically short (less than 200 amino acids). However, NLRs with N-terminal and mid-sequence ASMs differed in domain configuration with the former belonging to NACHT, NACHT ANK and NACHT VHS architectures, while the latter were all of the NACHT TPR type (Fig 4a). (In *M. roreri*, we found only one protein with the BaMLKL + NOD architecture (ESK90106.1) and the linker sequence between the domains did not resemble an amyloid-like motif).

Amyloid-like motifs in agaricomycetal NLRs share features with the HET-s motif homologs. In addition, we investigated two Agaricomycetes species with proteins comprising of a singular HeLo domain and a C-terminal double HET-s motif. In the genome of *Sphaerobolus stellatus* (strain SS14), which included four such C-termini, we found at least eight NACHT NLRs with N-termini comprising of single HRAM-like sequences (Fig H in

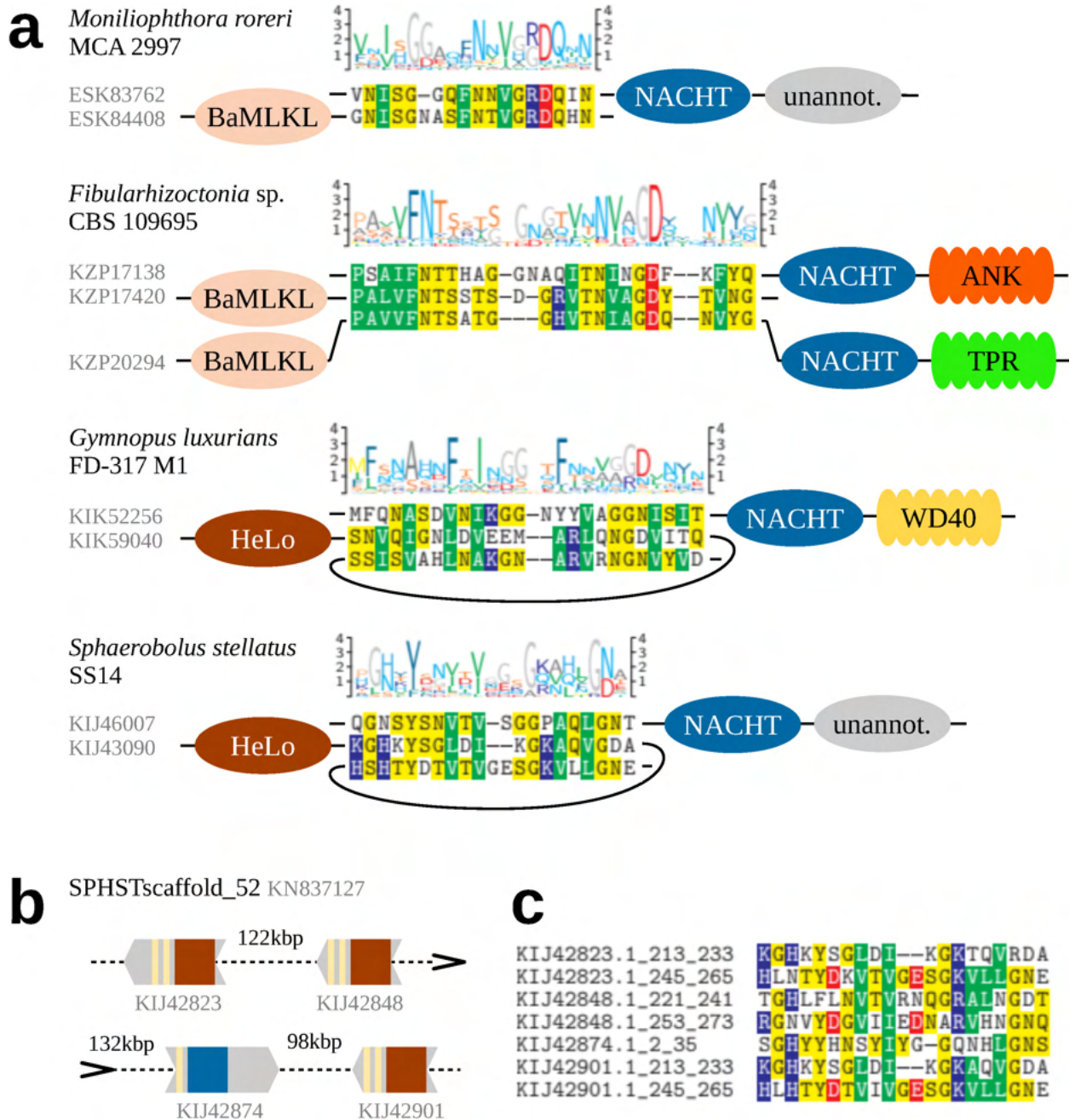


Fig 4. Potentially interacting amyloid-like motifs in Agaricomycetes. (a) Motif logos, sequence alignments and domain architectures of selected motif instances. (b) Schematic representation of a cluster of amyloid-like motifs in contig SPHSTscaffold_52 from genome assembly GCA_000827215.1 of *Sphaerobolus stellatus* SS14. (c) Multiple sequence alignment of motif instances in (b).

<https://doi.org/10.1371/journal.pcbi.1010787.g004>

[S1 Text](#)). Two instances (in KIJ28522 and KIJ30800) resembled the NLR13 HRAM motif. This strain was the only case where an NLR and three HeLo proteins were situated on a single contig in genome assembly (Fig 4b and 4c). The shortest distance between genes encoding NLR and HeLo was relatively large 95 kbp. The second species, *Gymnopus luxurians* (strain FD-317 M1), included one protein with HeLo + double HET-s motif architecture. While we did not

find any typical HRAMs in N-termini of 200 NLRs, several dozens included an instance of the NLR05/08/22/44 motif meta-family. When fragments of N-terminal sequences best-fitting the PCFG model were aligned, it revealed a 25-residue long core pattern. Interestingly, the alignment exhibited features characteristic to HRAMs: the N-terminal pattern of three hydrophobic residues and the C-terminal G [DN] bigram (Fig I in [S1 Text](#)). In total 32 amyloid-like motif instances were associated with NB-ARC, NACHT, NACHT WD and NACHT TPR domain architectures. Taken together these analyses strongly suggest that the NLR-associated amyloid signaling process also occurs in Basidiomycota.

N-terminal amyloid motifs often found in dozens of NLRs per basidiomycetal strain. Taken together, presented results support the presence of amyloid signaling in Basidiomycota, or more specifically in Agaricomycetes, in the context of NLR-based regulation of HeLo-/MLKL_NTD-likes. Moreover, they suggest that NLR05/08/22/29/44 meta-family of motifs is a basidiomycotal variety of the HRAM motif or its homolog. However, there were significant differences with regard to Ascomycota. First, while NLR-side amyloid signaling motifs were present in roughly half of Ascomycota strains, they were only found in 1/4 (30%) of Basidiomycota (Agaricomycetes) strains. Second, while there were typically only few amyloid signaling sequences per ascomycotal strain, there were usually dozens per basidiomycotal strain. At the same time, basidiomycotal effector-side C-terminal ASM sequences were seemingly less frequent than NLR-side N-terminal ASM sequences (Fig F-I in [S1 Text](#)). Indeed, the high number of NLR-side ASM sequences corresponded to enrichment of basidiomycotal sequences among shorter N-terminal domains (Fig 1b and Fig A in [S1 Text](#)).

Experimental validation of a novel amyloid signaling motif

PUASM displays sequence patterns typical to amyloid-like motifs. The alignment of PUASM instances (Fig 5a) revealed high similarity of PNP_UDP- and NLR-side sequences in the core region covered with the NANBNtm_035 pattern. Some divergence was present C-terminally, with pattern GND prevailing in PNP_UDP-side motifs and pattern ARD in NLR-side motifs. Interestingly, these 3-mers can be found in C-termini of already known amyloid signaling motifs HRAM1 [45] and BASS2 [12], respectively. Further four residues of the C-terminal extension of the motif exhibited a hydrophobic pattern well-conserved in pairwise alignments (Fig 5a). On the other side, N-terminal extensions of the PUASM profile matches often included histidine on the PNP_UDP side and glutamic acid on the NLR side. This, together with the overall composition of the N-terminal extensions, suggests some role of the charge complementarity.

Aggregation of synthetic PUASM peptides examined with ATR-FTIR, AFM and ThT assay. To check if biochemical properties of PUASM are consistent with its presumed role as the amyloid signaling motif, we experimentally analyzed a representative pair of motifs of this family, namely, PNP_UDP-side C-terminal EQB50682.1_332_355 and NLR-side N-terminal EQB50683.1_9_31 from a plant pathogenic fungus *Colletotrichum gloeosporioides* Cg-14 [89] (Table C and Fig J in [S1 Text](#)). The selected fragments entirely covered the matches of PUASM profiles and the pairwise conserved C-terminal extensions. The aggregation propensities of the PUASM peptides were determined experimentally using the Attenuated Total Reflectance—Fourier Transform Infrared spectroscopy (ATR-FTIR), Atomic Force Microscopy (AFM), and the Thioflavin T fluorescence assay (ThT). The ATR-FTIR spectroscopy allows determination of secondary structure and monitoring structural changes of peptides upon aggregation processes [90–92], while AFM is useful for detection and visualization of aggregates [93]. In turn, ThT is considered to be the “gold standard” for identifying amyloid fibrils [94, 95]. It is

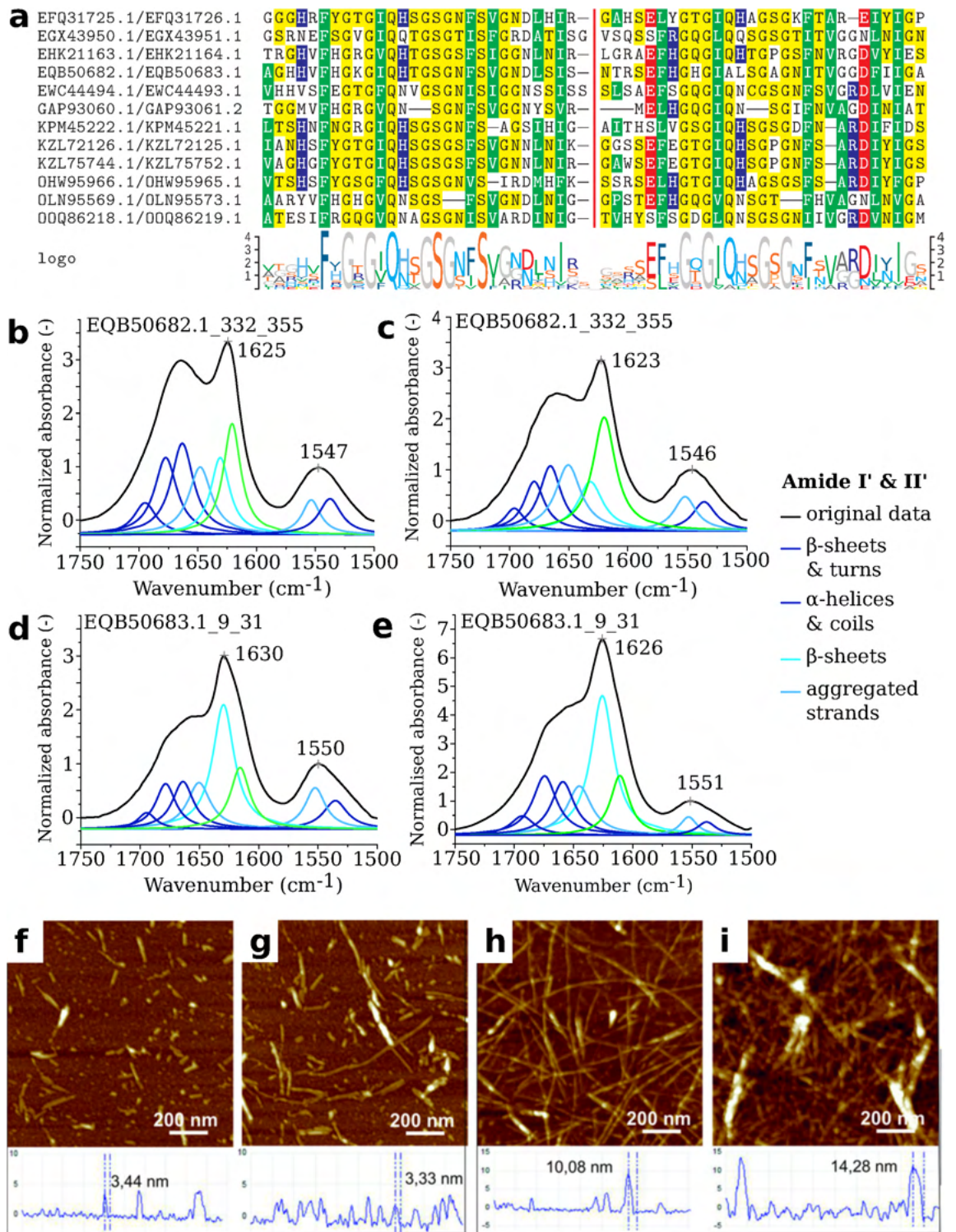


Fig 5. The PUASM motif. Alignment of the PUASM sequence pairs (a) from effector C-terminus (left) and NLR N-terminus (right). Colors indicate residue hydrophobicity, curly brackets—the motif ranges. Deconvolution of ATR-FTIR spectra of air-dried peptide films of EQB50682.1_332_355 (bc) and EQB50683.1_9_31 (de) in the amide bands region (1750–1500 cm⁻¹). Spectra registered at 20°C (68°F) after dissolving (bd) or after 40 days of incubation at 37°C (98.6°F) (ce). AFM images with cross-section profiles of peptides EQB50682.1_332_355 (f) and EQB50683.1_9_31 (g). Samples imaged after dissolving (fg) or after 40 days of incubation at 37°C (98.6°F) (hi).

<https://doi.org/10.1371/journal.pcbi.1010787.g005>

widely accepted that such a combination of experimental techniques is necessary to ascertain whether a particular peptide or protein is able to form the amyloid assemblies [96–98].

PUASM peptides display intramolecular β -structures and intermolecular β -sheets.

Analysis of the ATR-FTIR spectra in the range of 1750–1500 cm^{-1} (Fig 5b, 5c, 5d and 5e, Fig K, Table D and Table E in S1 Text) confirmed aggregation properties of studied peptides. The position of the Amide I band maximum was observed at 1625 cm^{-1} and 1630 cm^{-1} for EQB50682.1_332_355 and EQB50683.1_9_31, respectively. This signature is considered to be a spectroscopic marker of the cross- β amyloid architecture [91, 99]. High absorbances in the region of 1670–1660 cm^{-1} were observed in both spectra. The assignment of this band is still discussed in the literature [100–102]. The overall spectral line in Amide I was similar to the spectra observed for β -solenooidal proteins, including HET-s [103] and PrP^{Sc} [104]. While for both studied peptides the aggregation process was observed immediately after dissolving, N-terminal EQB50683.1_9_31 aggregated quicker and formed more well-ordered structures [105] (Fig 5b and 5d). A band curve-fitting method allowed to resolve individual Amide I band components and obtain a more detailed information about secondary structure of studied peptides. In the wavenumber range of 1640–1610 cm^{-1} two components were visible. The subband at about 1635 cm^{-1} corresponds to intramolecular β -structures. The percentage area of this component was 31% and 14% for peptide EQB50683.1_9_31 and EQB50682.1_332_355, respectively. In turn, the second component at about 1620 cm^{-1} corresponds to intermolecular β -sheets. Peptide EQB50683.1_9_31 displayed this subband at 1616 cm^{-1} , while EQB50682.1_332_355 at 1620 cm^{-1} , indicating a looser fibrillar structure of the latter.

PUASM peptides form amyloid-like aggregates that elongate during incubation.

Atomic Force Microscopy images of both PUASM peptides were acquired for two conditions related to the spectroscopy studies: after dissolving, and after 40 days of incubation at 37°C. The aggregation process of the peptides was present already in the sample after dissolving as the fibers with height of 3.44 ± 0.3 nm and 3.33 ± 0.3 nm, respectively for EQB50682.1_332_355 and EQB50683.1_9_31, were observed (Fig 5f and 5g). The height of the object observable in AFM is comparable with the size of the HET-s peptides obtained by the solid-state NMR technique (pdb:2kj3) [43]. Peptide aggregation was further enhanced in the samples imaged after 40 days of incubation at 37°C (Fig 5h and 5i), when the height of the aggregates reached 10.08 ± 0.9 nm and 14.28 ± 1.3 nm, respectively for EQB50682.1_332_355 and EQB50683.1_9_31. This clearly visible increasing aggregation process was in line with the ATR-FTIR measurements (Fig 5c and 5e).

PUASM peptides show an increase in ThT fluorescence in the assembly process. Thioflavin T (ThT) fluorescence assay is the most common assay to follow amyloid formation. We thus determined whether the PUASM peptides bind ThT. We observed an increase in ThT fluorescence over time with a sigmoidal curve for PNP_UDP-side C-terminal peptide EQB50682.1_332_355, starting with a lag phase of 2 hours (Fig L in S1 Text), followed by a rapid growth phase from 2–2.20 h, and ending at a stable plateau with the maximum ThT intensity. A significant increase in the fluorescence emission was observed for NLR-side N-terminal peptide EQB50683.1_9_31 (about 5 times higher than for EQB50682.1_332_355). The lag phase was not observed (Fig L in S1 Text). The steeper ThT curve with quicker attainment of plateau might indicate faster aggregation process of peptide EQB50683.1_9_31 in comparison to peptide EQB50682.1_332_355. While it is clear that both peptides showed an increase in ThT fluorescence during the assembly process, the presence of short fibrils in the AFM study complicate the comparative study of the aggregation kinetics of the two peptides.

GFP-PUASM spontaneously forms cytoplasmic foci *in vivo* alike other amyloid-like motifs. It was previously reported that fungal, bacterial and mammalian amyloid motifs

could form prions *in vivo* in the *Podospora anserina* model [12, 41, 46, 106]. To determine if PUASMs could also form prions *in vivo*, we expressed the PNP_UDP-side C-terminal EQB50682.1_332_355 from a plant pathogenic fungus *Colletotrichum gloeosporioides* Cg-14 [89] in *P. anserina* as GFP or RFP fusions. Three different constructs were generated: a N-terminal GFP fusion (GFP-PUASM) and C-terminal RFP and GFP fusions (respectively PUASM-RFP and PUASM-GFP). The three constructs were expressed from a strong constitutive promoter. In GFP-PUASM, the motif thus occurs C-terminally to the GFP domain, an organization that is analogous to that of the native full length EQB50682.1, in which the motif occurs C-terminally to the phosphorylase domain (PF01048). Prion formation was monitored using fluorescence microscopy by following the formation of cytoplasmic fluorescent foci as previously described for other amyloid signaling motifs expressed in *P. anserina* [12, 41, 46, 106]. A GFP fusion with an instance of previously characterized ASM (BASS3 of *Streptomyces atratus*) was used as positive control and its two proline mutants (BASS3 Q113P and Q120P) were used as negative controls [12]. GFP-BASS3 led to foci formation while the proline mutants did not. The GFP-PUASM fusion led initially to a diffused fluorescence signal (Fig 6, Table F in S1 Text). Upon subculturing, the number of transformants showing cytoplasmic foci gradually increased over time as typically observed for other prion amyloid motifs [106]

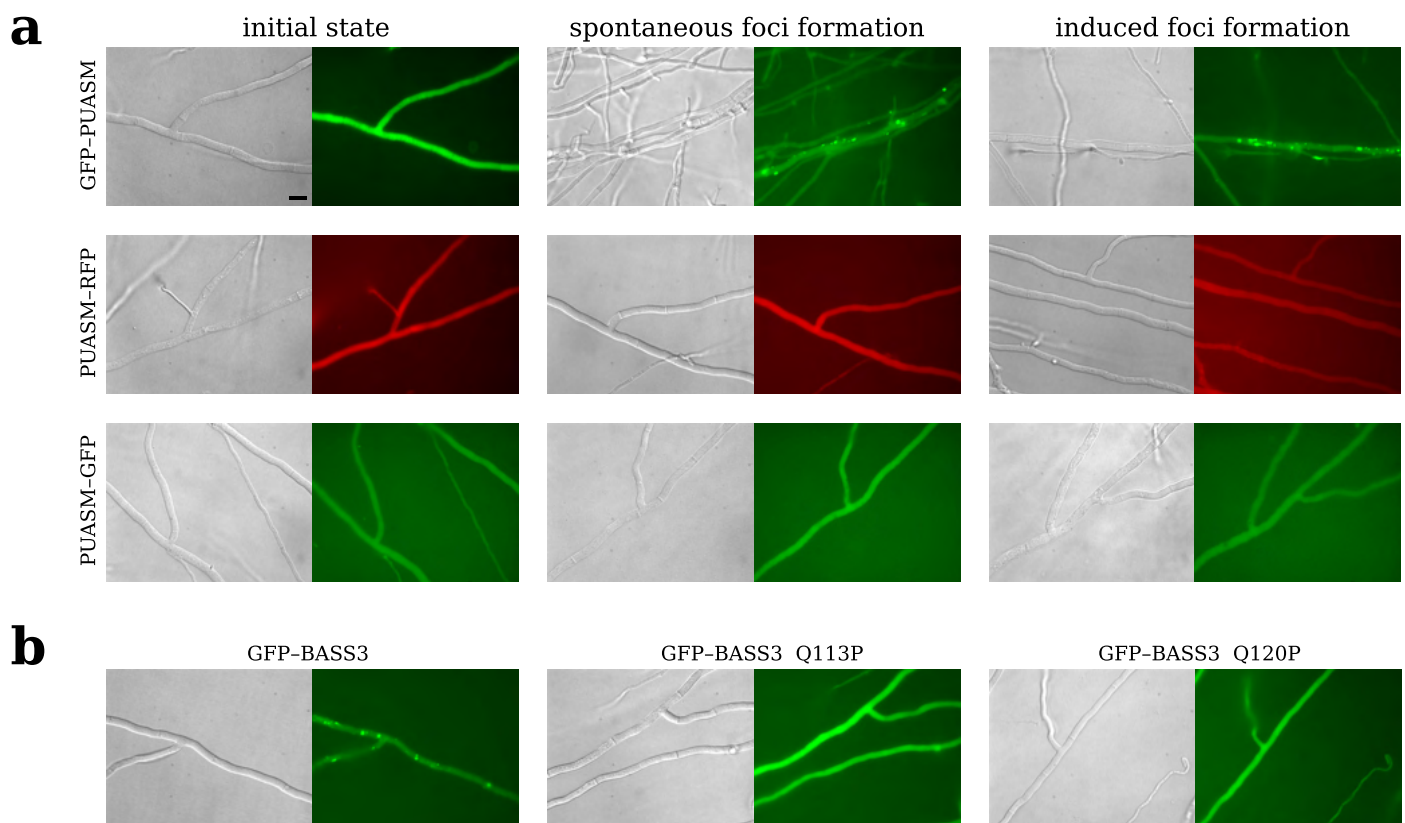


Fig 6. Expression of GFP/RFP-fused PUASM motifs in *Podospora anserina*. a) Micrographs of *P. anserina* strains expressing molecular fusions of PUASM with GFP or RFP, as indicated on the left; scale bar: 5 μ m. Strains were analyzed in their initial state after transfection (left panels marked) and either after several days of subculturing (middle panels, spontaneous foci formation) or after cytoplasmic contact with a strain expressing GFP-PUASM in the foci state (right panels, induced foci formation). Note that the GFP-PUASM construct (but not PUASM-RFP and PUASM-GFP) leads to spontaneous and induced foci-formation. Quantification of the rate of foci formation is given in Table F in S1 Text. b) Micrographs of strains expressing GFP-BASS3—positive control for the spontaneous foci formation (BASS3 motif of WP_037701008.1 of *Streptomyces atratus*, positions 70 to 124, left panel), and two GFP-BASS3 mutants—negative controls (Q113P, middle, and Q120P, right panel).

<https://doi.org/10.1371/journal.pcbi.1010787.g006>

(Table F in [S1 Text](#)). In contrast to the GFP–PUASM construct, fusion constructs displaying the motif N-terminally (PUASM–GFP and PUASM–RFP) remained diffused and did not form foci even upon prolonged subculturing. A similar situation was observed previously for the HELLF and RHIM motifs for which N-terminal position of the GFP/RFP inhibited foci formation [106]. We conclude from these experiments, that GFP–PUASM (but not RFP–PUASM and GFP–PUASM) spontaneously forms cytoplasmic foci as previously reported for other amyloid signaling motifs.

GFP–PUASM behaves as a prion *in vivo* in the *Podospora* model. To determine whether the foci state is infectious, strains expressing GFP–PUASM, PUASM–GFP or PUASM–RFP in the diffuse state were confronted to strains expressing GFP–PUASM in the foci state to induce cytoplasmic contact. At several time points after cytoplasmic contact, the recipient strains were subcultured and monitored to presence of foci (Table F in [S1 Text](#)). In this induced prion formation assay, GFP–PUASM strains were efficiently converted to the foci state after cytoplasmic contact with a GFP–PUASM strain in the foci state (Fig 6, Table F in [S1 Text](#)). In 96 hours after contact with the inducing strain, all tested strains displayed dots. In comparison, spontaneous dot formation was only detected in about 3% of the strains after 5 days of subculturing (Table F in [S1 Text](#)). Thus contact with a strain expressing GFP–PUASM dots induced dot formation in the recipient strain. Again, for the PUASM–GFP and PUASM–RFP proteins prion conversion was not observed. After confrontation with a strain expressing GFP–PUASM foci, when strains were subcultured no foci formation was detected. We conclude from these experiments that the GFP–PUASM fusion protein behaved as a prion *in vivo* in the *Podospora* model. Apparently, as in the case of other amyloid signaling motifs, the C-terminal position of the GFP/RFP inhibited foci formation [106]. In addition, the spontaneous and induced prion conversion of GFP–PUASM was somewhat less efficient than for other amyloid motifs that have been previously tested in the same way [12, 46, 106].

Discussion

In previous studies we computationally screened N-terminal domains of fungal NLRs using profile Hidden Markov Models (HMM) from the Pfam database directly and complemented the search with several Pfam-like inhouse models [5, 11]. Here we expanded the most recent analysis with a more sensitive search using the state-of-the-art clustering offered by MMseqs2 and HMM–HMM searches with HHblits. The study increased the overall Pfam annotation coverage of N-terminal domains by about 16% (or 19% when MLKL_NTD from CDD is counted), but also highlighted remarkable deficiencies in availability of annotations. Our results highlight the fact that the NLRs of fungi, and especially Basidiomycota, are still not sufficiently described.

Goodbye resembles Helo but with an additional N-terminal extension

The identification of a common structural core of Helo-like and Goodbye-like domains, the four-helix bundle, raises the question of their functional similarity. Both the distribution of associated nucleotide-binding domain and C-terminal domains, and the paralog-to-ortholog ratio for Goodbye-like and HeLo-like domains are similar [11], which may suggest similarities in their mode of operation. However, Goodbye-likes in NLR N-termini are often associated with another annotated effector domains, which is untypical for HeLo-likes. Moreover, the opposite is true for association with the amyloid signaling motifs, which is common to HeLos, HeLo-likes and basidiomycotal MLKL_NTDs but not to Goodbye-likes. In a plant homolog of HeLo, the N-terminal helix of the bundle (and entire protein) is known to play a significant role in triggering the cell death process [82]. However, in BaMLKL and Goodbye-likes, the

bundle is extended N-terminally by one or more helices, respectively. Thus, while the common evolutionary ancestry of HeLo-like and Goodbye-like is rather evident, the question of their functional similarity remains open. In particular, the functional role of the N-terminal extension of Goodbye-likes remains to be explored. For example, it can be speculated that the helices of the N-terminal extension in Goodbye-likes are displaced to enable the oligomerization process occur, possibly under the control of a non-amyloid mechanism involving domains associated with Goodbye-likes.

A large fraction of the effector domains are involved in regulated cell death

With the limitation that the evolutionary relation of Goodbye-like domains does not necessarily imply functional similarity, it appears that a substantial fraction of the effector domains in both ascomycetes and basidiomycetes is predicted to control regulated cell death. Involvement in regulated cell death has been reported not only for the HeLo/MLKL group but also for the HET domain [107], the Patatin [23] domain and more indirectly for the SesB-like domain [38]. One needs to add to this list the amyloid signaling motifs that control separate downstream cell-death effector domains. Globally, it would appear of at least one-third to half of the fungal NLRs could be involved in some kind of regulated cell death process. This high proportion raises the question of whether some of the other domains (whether annotated or not) could also play a role in regulated cell death. For example, it was recently reported that genes encoding fungal NLRs with N-terminal CHAT and S8 protease effector domains reside adjacent to Gasdermin-encoding genes [53].

Annotation of very short domains requires more complex methods than profile HMMs

While the vast majority of longer domains is at least partially annotated, this is true only for a definite minority of shorter domains. The shortage of annotations cannot be easily explained by the lack of conserved sequential features. Instead, one of the reasons is the profile HMM model itself, which by assessing each alignment position independently (except for indels) is not statistically powerful enough when dealing with short sequences. In other words, profile HMM models of more diverse families of short sequence fragments (e.g. 20–40 amino-acids long) cannot be sensitive and specific at the same time [52]. Currently, the problem can be at least partially addressed by using more complex and computationally demanding protein sequence models, such as probabilistic context-free grammars (PCFG) [12, 52, 108] and co-evolutionary Potts models [109–111]. Another viable option are the recurrent and attention-based neural networks, which have enough computational power to describe relevant dependencies in protein sequences [112–114]. However, while modern neural networks have been successfully applied to annotation of protein families [115, 116], their performance in modeling short protein sequence fragments is yet too to be evaluated.

NLR-associated amyloid-like motifs are less diverse in fungi compared to bacteria

In Ascomycota, we discovered two new amyloid signaling motif family, which are uniquely associated with the PNP_UDP domain. The amyloid properties of the more abundant PUASM motif were confirmed experimentally using a representative pair of N- and C-terminal sequences. Both of them generated amyloid-like fibers in the *in vitro* condition. (In depth study of the co-aggregation process is left for a separate study.) The effector-side PUASM sequence was shown to be capable of forming prions *in vivo* in the *Podospora anserina* model.

Despite the extensive search, the expand diversity of ASM remains lower in fungi than in bacteria using similar identification procedure, which is not inconsistent of the larger phylogenetic breath of the scanned bacterial genome as compared to the fungal ensemble.

Two strategies emerge for facilitating inheritance of amyloid signaling

Similarly to other ascomycotal amyloid signaling motifs, HRAM, PP and σ , effector proteins with C-terminal motif are often coded by direct genomic neighbors of the motif–NLR genes. Such genomic co-localization may facilitate co-inheritance of the two genes of the functional unit in the event of a recombination process. This may be of special importance for the NLR signaling pathway, which is polymorphic in population given the death-and-birth evolution. A notable exception is a PNP_UDP-associated HRAM motif variant (NLR12), for which only in a few cases the effector–motif and motif–NLR pairs present in the genome were co-localized (that is encoded by adjacent genes), while its NLR-side instances were relatively more frequent than the effector-side instances in some genomes. In Basidiomycota, virtually none of the hundreds of instances of amyloid-like motifs found in our survey in N-termini of NLRs was genomically co-localized with amyloid-like motif instances in effector C-termini. Again, in genomes where ASM co-occurred in both types of proteins, the NLR-side N-terminal instances were more frequent than the effector-side C-terminal instances. We speculate that the presence of many NLRs controlling the same effector could potentially relieve the need for genomic co-localization of NLRs and their effectors linked by amyloid signaling sequences.

Internal ASM instances may serve as scaffolds to stabilize the NLR oligomers

One interesting finding is presence of NLRs with intra-proteins amyloid-like motifs in *Fibularhizoctonia* sp. CBS 109695. Different central and C-terminal domain association in comparison to NLRs with N-terminal ASM-like motifs suggest also different functions of the motifs in both cases. Therefore, we hypothesize that these internal ASM instances may serve as scaffolds to stabilize the NLR oligomers, similar to cRHIM in the RIP1K/RIP3K complex [57]. In these lines, it is possible that also some other amyloid-like sequences identified in the current study but with no matching effector-side counterparts participate in the assembly of the NLR signalosome or are involved in interactions with motifs located outside the C-terminus of the associated protein.

Materials and methods

Computational methods

Annotation of NLR N-termini. A set of 36,141 NLR proteins from 487 fungal strains was identified in a previous study through the PSI-BLAST [117] search among completely sequenced fungal genomes in the NCBI nr database [11, 12] (the full list of accessions with their corresponding NOD domain boundaries is included in S1 Table). 32,962 N-termini at least 20 amino-acids long (91%), delimited according to the NACHT or NB-ARC query matches, were further considered, of which 18,674 (57%) were annotated using direct matches to Pfam [29] or inhouse HMM profiles (S1 Data) [11, 12]. The set of N-termini at least 20 amino-acid long was clustered with MMseqs2 [58] in mode 1 (21,758 N-termini in 127 clusters, 15,105 already annotated). Then, sequences in each cluster with at least 20 members were aligned using Clustal-Omega [118] (S2 Data) and searched for homologs in UniRef30 [59, 60] using HHblits [61] (parameters: `-e 0.001 -n 2 -E 0.01 -Z 1000000 -M 50`). Subsequently, the resulting alignments were used to search Pfam (HHblits parameters:

-e 0.001 -n 1 -E 1 -Z 1000000). The clustering required mutual coverage of at least 80% of sequence length, and the annotations were only assigned to sequences which covered at least 50% of the match to the Pfam profile. The resulting cluster-level annotations were retained only if the alignment match to the Pfam profile covered at least 50% of the profile length, and assigned only to individual sequences which covered at least 50% of the match. After completing the main processing, the set of N-termini was re-scanned for the Crinkler domain (PF20147) added recently to the Pfam database.

The tabularized results of the annotation are provided in [S1 Table](#). The overlapping Pfam annotations were resolved as in [11, 12]. The double HeLo/HeLo-like annotations were kept in [S1 Table](#) and in Table A in [S1 Text](#) but were represented as HeLo in [Fig 1f](#) and [Fig B](#) in [S1 Text](#). In addition, basidiomycotal sequences from clusters doubly annotated as Goodbye-like/Helo-like, as well as from clusters with CDD [73] MLKL_NTD annotations, were denoted as BaMLKL (see [Results](#)).

Comparative analysis of Goodbye-, HeLo- and MLKL-likes. For the largest clusters annotated as HeLo, HeLo-like, Goodbye-like and BaMLKL, their representative sequences were submitted to AlphaFold2 structure prediction [77] through the ColabFold advanced notebook [78]. Standard parameters of the notebook were applied except of (1) using the cluster alignments instead of searching genetic databases, (2) trimming off fragments just upstream the NACHT domain were applicable. Successful models—with the mean predicted pLDDT score [77] above 70 overall, and around 80 or more for the core helix bundle—and respective ColabFold outputs are provided in [S3 Data](#). For each cluster, the highest rank model was selected and structurally aligned to the experimentally solved MLKL domain (pdb:6zvo) using TM-align with default parameters [79]. Alignment conservation scores were calculated using the ConSurf webserver with default parameters [80, 81] based on the cluster alignments and AlphaFold2 structural models.

Characterization of unannotated longer N-termini. In addition, the largest MMSeqs-produced clusters, which did not get any Pfam annotation through the HHblits procedure, were carefully examined. For five unannotated clusters with at least ten members at the identity threshold of 70% and the median length above 100 amino acids, homologs were searched in UniProt [119] through the web-based hmmsearch with standard parameters [120], and predictions of the three dimensional structure for their representative sequences were attempted using AlphaFold2 [77] through the ColabFold advanced notebook [78]. Standard parameters were used except of adding the MMSeqs2 alignments to input (sequences just upstream the NACHT domain was trimmed off). Good quality structures (the predicted pLDDT score above 70) were obtained for three clusters, KEY84097, KFH66451 and PQE30996 ([S4 Data](#)). The proposed annotations for the five clusters (Table B in [S1 Text](#)) are assigned to member sequences in [S1 Table](#) and included in the TIR-like and “other” groups in [Fig 1f](#) and [Fig B](#) in [S1 Text](#).

Extraction of amyloid-like motifs in short N-termini. A subset of 54 NLR N-termini clusters with mean/median sequence length of at most 160/161 amino acids was selected. It consisted of N-termini of 3441 sequences, which were scanned using the PCFG-CM software [52, 121] probabilistic grammatical model inferred from ten families of bacterial ASMs (BASS1–10) [12, 52] ([S5 Data](#)) with scanning window of 20 to 40 amino acids and the smoothing factor of 10 PAM [52]. Very high scoring fragments (maximum log₁₀ score at least 3.5, mean log₁₀ score above 1.67) were found in 18 clusters with 1456 sequences ([S6 Data](#)). This included all 8 clusters (592 sequences) with at least one PFD-like annotation. The N-terminal sequences were made non-redundant at the identity level of 90% using CD-HIT 4.7 [122, 123] and submitted to motif extraction with MEME 5.0.5 [88, 124] with the following parameters: -nmotifs 100, -minsites 10, -maxsites 500, -minw 10, -maxw 30,

-allw, -evt 1. For each of 51 motifs found at the E-value threshold of 1, HMM profiles were built with HMMER 3.2.1 [125] and used for searching against the full set of grammar-fitting N-terminals (at the sequence and domain E-values of $1e - 2$). Then, obtained hits were extended by 5 amino acids in each direction and realigned using Clustal-Omega with the `auto` parameter. For each motif, the extended sequences were re-examined for consistency with the grammatical model (maximum log₁₀ score at least 3, mean log₁₀ score above 1). For 16 motifs which passed the grammatical filter, the alignments were used to build final HMM profiles (S7 Data).

Analysis of N-terminal amyloid-like motifs. The HMM profiles of the 16 motifs were used for scanning all N-termini longer than 10 amino acids (domain (independent) E-value threshold of $1e - 2$), comprising also sequences not included in the 127 clusters with 20 or more members. The resulting hits in 1538 sequences are included in S1 Table with coordinates (outermost in rare cases of double ASM hits). For further analysis only hits in N-termini shorter than 200 amino acids not located beyond position 150 were considered. Motif sequences in envelopes of 5 amino acids were tested for the beta-arch structure with ArchCandy 2.0 [51] using the recommended threshold of 0.56. Constituent sequences of the motifs were scanned using a generalized HRAM profile (S8 Data) at the domain (independent) E-value of $1e - 2$. The profile was built from HRAM motif sequences in Supplementary File 2 from [45], realigned using Mafft [126] (in the `auto` mode) and pruned of columns with more than 50% gaps using trimAl [127].

For each motif-containing NLR sequence, proteins coded by genes within the ± 20 kbp neighborhood of the genes encoding these NLRs were fetched from NCBI GenBank [128] or EMBL ENA [129] using an in-house Python (version 3.7.3) script aided by packages `requests` [130] and `xmldict` [131] (S2 Table). The set was then confined to proteins in the length range of 200–400 amino acids (S9 Data), which is typical for proteins with single domain architectures known to be associated to NLRs via amyloid signaling [12, 45]. Next, C-termini (100 amino acids) of the found neighboring proteins were scanned for the presence of the motifs using HMMER (domain (independent) E-value threshold of $1e - 2$, all heuristic filters off). Pairwise hits of the same motifs in N-termini of NLRs and C-termini of genomically neighboring proteins are collected in S3 Table.

Note that common occurrence of amyloid motifs at the N-termini of NLRs and at the C-termini of effector domains encoded by adjacent genes was repeatedly used for the identification of such motifs both in fungi and bacterial genomes [12, 37]. This criterion adds sensitivity and specificity to the identification of amyloid motifs.

Homology search of effector domains. Remote homologs of effector domains related to NLR proteins were iteratively searched for, starting from 19 Pfam profiles of N-terminal domains of NLRs reported in [5]: Pkinase (PF00069), Peptidase_S8 (PF00082), C2 (PF00168), PNP_UDP_1 (PF01048), TIR (PF01582), Patatin (PF01734), RelA_SpoT (PF04607), DUF676 (PF05057), HET (PF06985), PK_Tyr_Ser-Thr (PF07714), PGAP1 (PF07819), Abhydrolase_6 (PF12697), CHAT (PF12770), TIR_2 (PF13676), HeLo (PF14479), NACHT_N (PF17100), SesA (PF17107), Goodbye (PF17109) and Helo_like_N (PF17111). First, Pfam HMM profiles for each of the domains were used for searching against a local copy of the non-redundant protein sequences database (NCBI's "nr", downloaded in November 2019) [128] using HMMER 3.2.1 with the sequence inclusion E-value of $1e - 2$. Found proteins were then used to build the new HMM profiles and the search was repeated (this time with the more stringent sequence inclusion E-value of $1e - 3$) until the number of hits did not change by more than 7%. Final profiles were used to delimit domain boundaries through yet another `hmmsearch` run with the same E-value parameter but all heuristic filters turned off and the initial search space set to 12 155 478 (S10 Data). In addition, all fungal NACHT (PF05729) and NB-ARC (PF00931)

proteins were retrieved from the Pfam database (as of January 2020). C-termini of effector domains and N-termini of NACHT/NB-ARC NLRs were extracted and—in both cases—only fragments between 10 and 150 aa were selected for further analysis. (This effectively excluded nearly all proteins with effector + NOD architectures.) The final set included around 235k (nr: 187k) of effector C-termini and 6.8k (nr: 5.1k) NLR N-termini (S11 and S12 Data, respectively).

Identification of paired amyloid motifs. The sets of N- and C-termini were clustered using CD-HIT to reduce redundancy at the 70% similarity threshold (separately for each effector domain, together for NACHT and NB-ARC). Then, motif search was performed using MEME with the following parameters: `-nmotifs 100` for effectors or `-nmotifs 50` for NLRs, `-minsites 1%` of sequences but no less than 5 and no more than 10, `-maxsites 500`, `-minw 10`, `-maxw 30`, `-mod anr`. For each of 818 motifs identified at the E-value threshold of 1, including 769 motifs in effector C-termini and 49 motifs in NLR N-termini, HMM profiles were built in the two-stage procedure, as described above (see S13 Data). Next, the N- and C-termini were scanned with the combined set of effector- and NLR-side motif profiles. The same-motif hits in effector proteins and in NLRs (at domain (independent) E-value of $1e-2$) were matched based on genomic proximity (up to 20kbp) of genes encoding the proteins (see S4 Table for the genomic neighborhoods of genes encoding NACHT and NB-ARC proteins with short N-termini). At least 3 non-redundant pairs of motif instances were found for 22 motifs (S13 Data), which were then clustered on the basis of their co-occurrence in 190 pairs of genomically neighboring proteins (S5 Table and Fig E in S1 Text).

Finally, hits of the 16 ASM motif profiles in short N-termini of NLRs (previously analyzed) and hits in short C-termini of effector domains (from the homology search, included at domain (independent) E-value of $1e-2$ over the entire set) were matched on the strain level (through the BioSample and BioProject identifiers; entries with incomplete pairs of identifiers were rejected) in order to identify potentially correlated pairs, which are not co-localized in genomes (S6 Table).

Specialized searches for amyloid motifs in Basidiomycota. BaMLKL homologs were searched in UniProt [119] through the web-based hmmsearch [120] with standard parameters starting from the alignment of the largest BaMLKL cluster in Basidiomycota (representative protein: KIM77258), trimmed to the NACHT_N match. Hits were further restricted to GenBank sequences with length up to 400 amino acids and no Pfam P-loop_NTPase clan (CL0023) annotation at E-value of 1. C-termini (100aa) of resulting 241 BaMLKL homologs (S14 Data) were scanned with the PCFG BASS model (S5 Data) with the same parameters as above (except the minimum scanning window length of 15). For proteomes with the most promising hits in BaMLKL homologs (log10 score above 3, eight sequences from six species), N-termini (150 aa) of all NLR proteins were again scanned with the grammars. Promising N-terminal hits were obtained for *Moniliophthora roreri* (strains 2995 and 2997), *Laccaria amethystina* (strain LaAM-08-1), and *Fibularhizoctonia* sp. CBS 109695. The matched fragments were aligned with their C-terminal counterparts on the per genome basis with Mafft [126] in an accurate mode (`-maxiterate 1000 -localpair`). The NLR N-terminal and BaMLKL C-terminal ASM-like sequences aligned satisfactorily for *M. roreri* (we only analyzed strain 2997 due to high similarity between the strains) and *Fibularhizoctonia* sp. CBS 109695. The alignments were then extended and trimmed manually (Fig F and Fig G in S1 Text). In addition, the sequences were scanned with the 16 HMM profiles of amyloid-like motifs (domain (independent) E-value of $1e-2$).

Next, fungal proteomes in UniProt were scanned using web-based jackhmmmer [120] with standard parameters starting from the double HET-s motif from Q03689 (AAB94631) (residues 218–289) of *Podospora anserina*, which resulted in finding five complete HeLo-

HARAM-HRAM proteins in two Agaricomycetes: four from *Sphaerobolus stellatus* SS14 and one from *Gymnopus luxurians* FD-317 M1 (see Fig 4). NLRs in these genomes were then scanned with the PCFG model and the hits exceeding the log10 score threshold of 2.33 were aligned with their C-terminal counterparts on the per genome basis with Mafft [126] in the accurate mode. Finally, the alignments were curated manually (poorly aligned sequences were excluded, sequences were extended or trimmed if necessary, Fig H and Fig I in S1 Text).

Visualization. Basic data processing and visualization was conducted in Python using pandas [132, 133], matplotlib [134] and seaborn [135] packages, as well as in LibreOffice, GIMP and Inkscape. Multiple sequence alignments and logos were generated using TeXshade [136]. The graph of logos in Fig E in S1 Text was generated with graphviz 2.40.1 [137]. Visualizations of structural models were generated with RasMol [138] (Fig 2) or taken directly from the ColabFold notebook [78] (Fig D in S1 Text).

Experimental methods

In vitro analysis. Peptide synthesis. All commercially available reagents and solvents were purchased from Merck, Sigma-Aldrich and Lipopharm.pl, and used without further purification. Peptides EQB50682.1_332_355 (VFHGKGIQHTGSGNFSVGNLDSIS) and EQB50683.1_9_31 (FHGHGIALSGAGNITVGGDFIIG) were synthesized with an automated solid-phase peptide synthesizer (Liberty Blue, CEM) using rink amide AM resin (loading: 0.59 mmol/g). Fmoc deprotection was achieved using 20% piperidine in DMF for 1 min at 90°C. A double-coupling procedure was performed with 0.5 M solution of DIC and 0.25 M solution of OXYMA (1:1) in DMF for 4 min at 90°C. Cleavage of the peptides from the resin was accomplished with the mixture of TFA/TIS/H₂O (95:2.5:2.5) after 3 h of shaking. The crude peptide was precipitated with ice-cold Et₂O and centrifuged (8000 rpm, 15 min, 2°C). Peptides were purified using preparative HPLC (Knauer Prep) with a C18 column (Thermo Scientific, Hypersil Gold 12 μl, 250 × 20 mm) with water/acetonitrile (0.05% TFA) eluent system.

Peptide analytics. Analytical high-performance liquid chromatography (HPLC) was performed using Kinetex 5μ EVO C18 100A 150 × 4.6 mm column. Program (eluent A: 0.05% TFA in H₂O, eluent B: 0.05% TFA in acetonitrile, flow 0.5 mL/min): A: t = 0 min, 90% A; t = 45 min (25 min in case of EQB50682.1_332_355). Peptides were studied by WATERS LCT Premier XE System consisting of high resolution mass spectrometer (MS) with a time of flight (TOF).

Attenuated Total Reflectance—Fourier Transform Infrared Spectroscopy (ATR-FTIR). Lyophilized peptides were dissolved in D₂O (deuterium oxide, 99.8% D, Carl Roth, GmbH, Germany) to final concentration of ca. 814 μM. The spectroscopic measurements were performed directly after dissolving peptides in a solvent, after 7 and 40 days of incubation process at 37°C (98.6°F). In addition peptides were measured after 40 days of incubation at 4°C (39.2°F, Fig K in S1 Text). Each time, 10 μl of peptide solution was dropped directly on the diamond surface and was allowed to dry out. ATR-FTIR spectra were recorded using a Nicolet 6700 FTIR Spectrometer (Thermo Scientific, USA) with Golden Gate Mk II ATR Accessory with Heated Diamond Top-plate (PIKE Technologies). The spectrometer was continuously purged with dry air. Directly before sampling, the background spectrum of diamond/air was collected as a reference. For each spectrum 512 scans with a resolution of 4 cm⁻¹ were co-added. All spectra were obtained in the range of 4000–450 cm⁻¹ at 20°C (68.0°F).

Spectroscopy data treatment. ATR-FTIR spectra were initially preprocessed using OMNIC software (version 8, Thermo Fisher Scientific, USA): atmospheric and ATR correction. All spectra were analyzed using the OriginPro (version 2019, OriginLab Corporation, USA). The analysis included: baseline correction, smoothing using the Savitzky-Golay

polynomial filter (polynomial order 2, a window size of 9 points) [139] and normalization to 1 for the Amide II' band. Spectra in the amide bands region (1750–1500 cm^{-1}) were deconvoluted into subcomponents using the Lorentz function based on second and fourth derivative spectra (R-Square 0.997).

Atomic Force Microscopy. AFM images were acquired in tapping mode using a Nanoscope IIIa scanning probe microscope with Extender Module (Bruker) in the dynamic modus. An active vibration isolation platform was applied. Olympus etched silicon cantilevers were used with a typical resonance frequency in the range of 100–200 kHz and a spring constant of 40 N/m. The set-point amplitude of the cantilever was maintained by the feedback circuitry at 80% of the free oscillation amplitude of the cantilever. The volume of 10 μL of 0.814 μM peptide was applied to freshly cleaved ultra-clean mica (Nano and More) and incubated at room temperature for 30 s. The mica discs were then rinsed with ultra-clean purified 18.2 M Ω deionized water and dried using gentle nitrogen gas flow. All samples were measured at room temperature in air. Structural analysis and height measurements of acquired images were performed with Nanoscope v.6.13 software.

Thioflavin T fluorescence assay. ThT powder was dissolved in MilliQ to final concentration 2 mM and filtered through 0.22 μm syringe. ThT solution was dissolved in 50 mM Tris-HCl (pH = 7.4) to final concentration 10 μM and filtered. The 90 μL of ThT buffer was mixed with 10 μL of peptide solution (concentration 400 μM) in the 96-wells plate. Samples were measured on the SpectraMax Gemini XPS Microplate (Molecular Devices LLC). The measurements were conducted in room temperature. The excitation wavelength was set at 450 nm and the emission was recorded in the range from 470 to 500 nm. Each group of experiment contained three parallel samples and the data were averaged after measurements.

In vivo analysis. Strains and plasmids. The *Podospora anserina* Δhellp (ΔPa_5_8070) $\Delta\text{het-s}$ (ΔPa_3_620) Δhellf (ΔPa_3_9900) strain [106] was used as recipient strain for the expression of molecular fusions of PUASM (PNP_UDP-side C-terminal EQB50682.1_332_355 VFHGKGIQHTGSGNFSVGNLDSIS) from the plant pathogenic fungus *Colletotrichum gloeosporioides* Cg-14 [89] and the GFP (green fluorescent protein) or RFP (red fluorescent protein). These fusions were expressed from plasmids based on the pGEM-T backbone (Promega) named pOP [38] and containing either the GFP or RFP encoding gene, or in a derivative of the pAN52.1 GFP vector [140], named pGB6-GFP and containing the GFP encoding gene. In both cases, the molecular fusions were under the control of the strong constitutive *P. anserina* *gpd* (glyceraldehyde-3-phosphate dehydrogenase) promoter. The Δhellp $\Delta\text{het-s}$ Δhellf strain was transformed as described [141] with a fusion construct along with a second vector carrying a ble phleomycin-resistance gene, pPaBle (using a 10:1 molar ratio). Phleomycin-resistant transformants were selected, grown for 30 h at 26°C and screened for the expression of the transgenes using fluorescence microscopy. PUASM was amplified with specific primers either 5' ggcttaattaaATGGTCTTTCATGGCAAGGGCATCC 3' and 5' ggcagatcttgcctccGGA GATGCTGAGATCG 3' for cloning in pOP plasmids, or 5' ggcgc ggcgcGTCTTTCATGG CAAGGGCATC 3' and 5' ggcGGATC-CTTAGGAGATGCTGAGATCGTTGCC 3' for cloning in the pGB6 plasmid (capital letters correspond to the PUASM sequence). The PCR products were cloned upstream of the GFP or RFP coding sequence in the pOP plasmids using PacI/BglII restriction enzymes to generate the pOPPUASM-GFP and pOPPUASM-RFP vectors in which in addition to the BglII site, a two amino acid linker (GA) was introduced between the sequences encoding PUASM and GFP or RFP and cloned downstream of the GFP using NotI/BamHI restriction enzymes to generate the pGB6-GFP-PUASM plasmid.

Microscopy. *P. anserina* hyphae were inoculated on solid medium and cultivated for 24 to 48 h at 26°C. The medium was then cut out, placed on a glass slide and examined with a Leica

DMRXA microscope equipped with a Micromax CCD (Princeton Instruments) controlled by the Metamorph 5.06 software (Roper Scientific). The microscope was fitted with a Leica PL APO 63X immersion lens.

Prion propagation. Methods for determination of prion formation and propagation were previously described [12, 142]. Prion formation and propagation can be observed using microscopy by monitoring the formation of fluorescent dots. Spontaneous prion formation is first monitored as the rate of spontaneously acquired prion phenotype (dot formation) in the initially prion-free subculture after 5, 11, 18, 32, 49 and 75 days of growth at 26°C on corn-meal agar using microscopy as described. Prion formation can also be measured as the ability to propagate prions from a donor strain (containing prion) to a prion-free strain (induced strain). In practice, prion-free strains are confronted on solid corn-meal agar medium for 2 to 5 days (contact between strains was observed after 24 to 36 hours of culture) before being subcultured and observed by fluorescence microscopy for the presence of dots (this test is referred to as induced prion formation). At least 18 different transformants were used and the tests were realized in triplicates. It is to note that transformants were randomly tested for prion formation allowing various expression levels of the transgene (high levels of expression are usually associated with rapid spontaneous prion formation) except for the induced conversion test where transformants expressing moderate level of transgene were preferred to limit the rate of spontaneous transition within the timing of the experiment that could mask the prion induction.

As a control, we also imaged anew GFP fusion proteins with the wild-type and mutant form of a previously characterized amyloid signaling motif the BASS3 motif found in WP_037701008.1 from *Streptomyces atratus* described in [12]. Two proline mutants substituting conserved glutamine residues that were found previously to abolish in vivo dot formation were used (Q113P and Q120P).

Supporting information

S1 Text. Supplementary online materials. The document includes supplementary text, tables (Table A–F in S1 Text) and figures (Fig A–L in S1 Text).
(PDF)

S1 Table. Tabularized results of N-termini annotation. The table aggregates results presented in the manuscript.
(CSV)

S2 Table. Genomic neighbors of candidate short N-termini NLRs with ASMs. The list includes accessions of proteins encoded by genes within the neighborhood of 20kbp of genes encoding the query proteins (S6 Data).
(CSV)

S3 Table. Pairwise hits of the same ASMs in N-termini of NLRs and C-termini of genomically neighboring proteins. The table is based on S6 and S7 Data, S2 Table and S9 Data. See Computational methods for details.
(CSV)

S4 Table. Genomic neighbors of candidate short N-termini Pfam NACHT and NB-ARC proteins. The list includes accessions of proteins encoded by genes within the neighborhood of 20kbp of genes encoding the query proteins (S12 Data).
(CSV)

S5 Table. Pairwise hits of the same ASMs in N-termini of NACHT/NB-ARC NLRs and C-termini of genomically neighboring effector proteins. The table is based on [S11](#), [S12](#) and [S13](#) Data and [S4 Table](#). See Computational methods for details.

(CSV)

S6 Table. Pairwise hits of the same ASMs in N-termini of NLRs and C-termini of genomically co-occurring effector proteins. The table is based on [S6](#) and [S7](#) Data and [S11 Data](#). See Computational methods for details.

(CSV)

S1 Data. Profile HMMs of NLR effector domains. The file includes previously unpublished models used in [\[5, 11\]](#).

(HMM)

S2 Data. Multiple sequence alignments of N-termini clusters. The alignments were calculated using ClustalOmega for 127 MMseqs2 clusters with at least 20 member sequences.

(GZ)

S3 Data. Structure prediction of HeLo-/Goodbye-/MLKL-like domains. Full AlphaFold2/ColabFold outputs.

(GZ)

S4 Data. Structure prediction of previously unannotated domains. Full AlphaFold2/ColabFold outputs.

(GZ)

S5 Data. PCFGs for BASS. The file includes previously unpublished grammars used in [\[52\]](#) and a sample scanning configuration.

(GZ)

S6 Data. Candidate short NLR N-termini with ASMs. The FASTA file includes sequences from clusters with high content of ASM-like sequences, according to the BASS PCFGs ([S5 Data](#)).

(FA)

S7 Data. Profile HMMs of ASMs found in short NLR N-termini. Please refer to Computational methods for the profile generation process.

(HMM)

S8 Data. Profile HMM of HeLo-related HRAMs. The profile is based on the motifs identified in [\[45\]](#).

(HMM)

S9 Data. Short C-termini of 200–400 aa long proteins genomically neighboring candidate short NLR N-termini with ASMs. The FASTA file concerns target proteins listed in [S2 Table](#).

(FA)

S10 Data. Lists of HMMER domain hits of effector domain profiles. The lists were obtained through iterative searches in NCBI “nr” starting from Pfam profiles of known NLR effector domains.

(GZ)

S11 Data. Short C-termini of effector proteins. The FASTA file concerns target proteins listed in [S10 Data](#).

(FA)

S12 Data. Short N-termini of Pfam NACHT and NB-ARC proteins. The FASTA file concerns proteins from NCBI “nr” associated with the two families in the Pfam database.
(FA)

S13 Data. Profile HMMs of ASMs found both in effector C-termini and NLR N-termini of genomically neighboring proteins. Please refer to Computational methods for the profile generation process.
(HMM)

S14 Data. BaMLKL homologs identified with hmmsearch in Basidiomycota. A FASTA file.
(FA)

Acknowledgments

The authors acknowledge the use of the E-SCIENCE.PL infrastructure.

Author Contributions

Conceptualization: Sven J. Saupe, Witold Dyrka.

Data curation: Jakub W. Wojciechowski, Emirhan Tekoglu, Witold Dyrka.

Formal analysis: Jakub W. Wojciechowski, Emirhan Tekoglu, Witold Dyrka.

Funding acquisition: Sven J. Saupe.

Investigation: Jakub W. Wojciechowski, Emirhan Tekoglu, Marlena Gąsior-Głogowska, Virginie Coustou, Natalia Szulc, Monika Szefczyk, Marta Kopaczyńska, Sven J. Saupe, Witold Dyrka.

Methodology: Jakub W. Wojciechowski, Emirhan Tekoglu, Marlena Gąsior-Głogowska, Virginie Coustou, Sven J. Saupe, Witold Dyrka.

Project administration: Witold Dyrka.

Resources: Marlena Gąsior-Głogowska, Virginie Coustou, Monika Szefczyk, Marta Kopaczyńska, Sven J. Saupe.

Software: Jakub W. Wojciechowski, Emirhan Tekoglu, Witold Dyrka.

Supervision: Marta Kopaczyńska, Witold Dyrka.

Validation: Emirhan Tekoglu, Marlena Gąsior-Głogowska, Virginie Coustou, Natalia Szulc, Monika Szefczyk, Marta Kopaczyńska, Sven J. Saupe, Witold Dyrka.

Visualization: Marlena Gąsior-Głogowska, Virginie Coustou, Natalia Szulc, Marta Kopaczyńska, Witold Dyrka.

Writing – original draft: Jakub W. Wojciechowski, Emirhan Tekoglu, Marlena Gąsior-Głogowska, Natalia Szulc, Monika Szefczyk, Marta Kopaczyńska, Sven J. Saupe, Witold Dyrka.

Writing – review & editing: Jakub W. Wojciechowski, Marlena Gąsior-Głogowska, Natalia Szulc, Sven J. Saupe, Witold Dyrka.

References

1. Jones JDG, Vance RE, Dangl JL. Intracellular innate immune surveillance devices in plants and animals. *Science*. 2016; 354(6316):aaf6395. <https://doi.org/10.1126/science.aaf6395> PMID: 27934708

2. Uehling J, Deveau A, Paoletti M. Do fungi have an innate immune response? An NLR-based comparison to plant and animal immune systems. *PLoS Pathogens*. 2017; 13(10):e1006578. <https://doi.org/10.1371/journal.ppat.1006578> PMID: 29073287
3. Duxbury Z, Wu Ch, Ding P. A Comparative Overview of the Intracellular Guardians of Plants and Animals: NLRs in Innate Immunity and Beyond. *Annual Review of Plant Biology*. 2021; 72(1):155–184. <https://doi.org/10.1146/annurev-arplant-080620-104948> PMID: 33689400
4. Koonin EV, Aravind L. Origin and evolution of eukaryotic apoptosis: the bacterial connection. *Cell Death & Differentiation*. 2002; 9:394–404. <https://doi.org/10.1038/sj.cdd.4400991> PMID: 11965492
5. Daskalov A, Dyrka W, Saupe SJ. NLR function in fungi as revealed by the study of self/non-self recognition systems. In: Benz JP, Schipper K, editors. *NLR Function in Fungi as Revealed by the Study of Self/Non-self Recognition Systems*. Cham: Springer International Publishing; 2020. p. 123–141.
6. Xiong Y, Han Z, Chai J. Resistosome and inflammasome: platforms mediating innate immunity. *Current opinion in plant biology*. 2020; 56:47–55. <https://doi.org/10.1016/j.pbi.2020.03.010> PMID: 32554225
7. Bi G, Zhou JM. Regulation of Cell Death and Signaling by Pore-Forming Resistosomes. *Annual review of phytopathology*. 2021; 59:239–263. <https://doi.org/10.1146/annurev-phyto-020620-095952> PMID: 33957051
8. Saur IML, Panstruga R, Schulze-Lefert P. NOD-like receptor-mediated plant immunity: from structure to cell death. *Nat Rev Immunol*. 2021; 21(5):305–318. <https://doi.org/10.1038/s41577-020-00473-z> PMID: 33293618
9. van der Biezen EA, Jones JDG. The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. *Current Biology*. 1998; 8(7):R226–R228. [https://doi.org/10.1016/S0960-9822\(98\)70145-9](https://doi.org/10.1016/S0960-9822(98)70145-9) PMID: 9545207
10. Koonin E, Aravind L. The NACHT family—a new group of predicted NTPases implicated in apoptosis and MHC transcription activation. *Trends in Biochemical Sciences*. 2000; 25:223–224. [https://doi.org/10.1016/S0968-0004\(00\)01577-2](https://doi.org/10.1016/S0968-0004(00)01577-2) PMID: 10782090
11. Dyrka W, Lamacchia M, Durrrens P, Kobe B, Daskalov A, Paoletti M, et al. Diversity and Variability of NOD-Like Receptors in Fungi. *Genome Biology and Evolution*. 2014; 6:3137–3158. <https://doi.org/10.1093/gbe/evu251> PMID: 25398782
12. Dyrka W, Coustou V, Daskalov A, Lends A, Bardin T, Berbon M, et al. Identification of NLR-associated amyloid signaling motifs in bacterial genomes. *Journal of Molecular Biology*. 2020; 432(23):6005–6027. <https://doi.org/10.1016/j.jmb.2020.10.004> PMID: 33058872
13. Gao LA, Wilkinson ME, Strecker J, Makarova KS, Macrae RK, Koonin EV, et al. Prokaryotic innate immunity through pattern recognition of conserved viral proteins. *Science*. 2022; 377(6607): eabm4096. <https://doi.org/10.1126/science.abm4096> PMID: 35951700
14. Saupe S, Turcq B, Begueret J. A gene responsible for vegetative incompatibility in the fungus *Podospora anserina* encodes a protein with a GTP-binding motif and G beta homologous domain. *Gene*. 1995; 162:135–139. [https://doi.org/10.1016/0378-1119\(95\)00272-8](https://doi.org/10.1016/0378-1119(95)00272-8) PMID: 7557402
15. Paoletti M. Vegetative incompatibility in fungi: From recognition to cell death, whatever does the trick. *Fungal Biol Rev*. 2016; 30(4):152–162. <https://doi.org/10.1016/j.fbr.2016.08.002>
16. Goncalves AP, Heller J, Rico-Ram rez AM, Daskalov A, Rosenfield G, Glass NL. Conflict, Competition, and Cooperation Regulate Social Interactions in Filamentous Fungi. *Annual Review of Microbiology*. 2020; 74(1):693–712. <https://doi.org/10.1146/annurev-micro-012420-080905> PMID: 32689913
17. Urbach JM, Ausubel FM. The NBS-LRR architectures of plant R-proteins and metazoan NLRs evolved in independent events. *Proceedings of the National Academy of Sciences*. 2017; 114:1063–1068. <https://doi.org/10.1073/pnas.1619730114> PMID: 28096345
18. Espagne E, Balhadere P, Penin ML, Barreau C, Turcq B. HET-E and HET-D belong to a new subfamily of WD40 proteins involved in vegetative incompatibility specificity in the fungus *Podospora anserina*. *Genetics*. 2002; 161:71–81. <https://doi.org/10.1093/genetics/161.1.71> PMID: 12019224
19. Bastiaans E, Debets AJ, Aanen DK, van Diepeningen AD, Saupe SJ, Paoletti M. Natural variation of heterokaryon incompatibility gene *het-c* in *Podospora anserina* reveals diversifying selection. *Molecular Biology and Evolution*. 2014; 31:962–974. <https://doi.org/10.1093/molbev/msu047> PMID: 24448643
20. Armant MA, Fenton MJ. Toll-like receptors: a family of pattern-recognition receptors in mammals. *Genome biology*. 2002; 3(8):REVIEWS3011. <https://doi.org/10.1186/gb-2002-3-8-reviews3011> PMID: 12186654
21. Toshchakov VY, Neuwald AF. A survey of TIR domain sequence and structure divergence. *Immunogenetics*. 2020; 72(3):181–203. <https://doi.org/10.1007/s00251-020-01157-7> PMID: 32002590

22. Lapin D, Johandrees O, Wu Z, Li X, Parker JE. Molecular innovations in plant TIR-based immunity signaling. *The Plant Cell*. 2022; 34(5):1479–1496. <https://doi.org/10.1093/plcell/koac035> PMID: 35143666
23. Heller J, Clave C, Gladieux P, Saube SJ, Glass NL. NLR surveillance of essential SEC-9 SNARE proteins induces programmed cell death upon allorecognition in filamentous fungi. *Proc Natl Acad Sci U S A*. 2018; 115(10):E2292–E2301. <https://doi.org/10.1073/pnas.1719705115> PMID: 29463729
24. Paoletti M, Saube SJ. Fungal incompatibility: Evolutionary origin in pathogen defense? *BioEssays*. 2009; 31(11):1201–1210. PMID: 19795412
25. Paoletti M, Saube SJ, Clave C. Genesis of a Fungal Non-Self Recognition Repertoire. *PLoS ONE*. 2007; 2(3):e283. <https://doi.org/10.1371/journal.pone.0000283> PMID: 17356694
26. Chevanne D, Saube S, Clave C, Paoletti M. WD-repeat instability and diversification of the *Podospora anserina* hwd non-self recognition gene family. *BMC Evolutionary Biology*. 2010; 10(1):134. <https://doi.org/10.1186/1471-2148-10-134> PMID: 20459612
27. Rando OJ, Verstrepen KJ. Timescales of Genetic and Epigenetic Inheritance. *Cell*. 2007; 128:655–668. <https://doi.org/10.1016/j.cell.2007.01.023> PMID: 17320504
28. Iotti M, Rubini A, Tisserant E, Kholer A, Paolucci F, Zambonelli A. Self/nonself recognition in *Tuber melanosporum* is not mediated by a heterokaryon incompatibility system. *Fungal Biology*. 2012; 116(2):261–275. <https://doi.org/10.1016/j.funbio.2011.11.009> PMID: 22289772
29. Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*. 2016. <https://doi.org/10.1093/nar/gkv1344> PMID: 26673716
30. Saube SJ. Amyloid Signaling in Filamentous Fungi and Bacteria. *Annual Review of Microbiology*. 2020; 74(1):673–691. <https://doi.org/10.1146/annurev-micro-011320-013555> PMID: 32689912
31. Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, et al. The alpha/beta hydrolase fold. *Protein Engineering, Design and Selection*. 1992; 5(3):197–211. <https://doi.org/10.1093/protein/5.3.197> PMID: 1409539
32. Mushegian A, Koonin E. Unexpected sequence similarity between nucleosidases and phosphoribosyltransferases of different specificity. *Protein science: a publication of the Protein Society*. 1994; 3(7):1081–1088. <https://doi.org/10.1002/pro.5560030711> PMID: 7920254
33. Mao C, Cook W, Zhou M, Koszalka G, Krenitsky T, Ealick S. The crystal structure of *Escherichia coli* purine nucleoside phosphorylase: a comparison with the human enzyme reveals a conserved topology. *Structure (London, England: 1993)*. 1997; 5(10):1373–1383. [https://doi.org/10.1016/S0969-2126\(97\)00287-6](https://doi.org/10.1016/S0969-2126(97)00287-6) PMID: 9351810
34. Balguerie A, Dos Reis S, Ritter C, Chaignepain S, Couлары-Salin B, Forge V, et al. Domain organization and structure-function relationship of the HET-s prion protein of *Podospora anserina*. *The EMBO Journal*. 2003; 22(9):2071–2081. <https://doi.org/10.1093/emboj/cdg213> PMID: 12727874
35. Fedorova ND, Badger JH, Robson GD, Wortman JR, Nierman WC. Comparative analysis of programmed cell death pathways in filamentous fungi. *BMC Genomics*. 2005; 6:177. <https://doi.org/10.1186/1471-2164-6-177> PMID: 16336669
36. Greenwald J, Buhtz C, Ritter C, Kwiatkowski W, Choe S, Maddelein ML, et al. The mechanism of prion inhibition by HET-S. *Molecular Cell*. 2010; 38:889–899. <https://doi.org/10.1016/j.molcel.2010.05.019> PMID: 20620958
37. Daskalov A, Paoletti M, Ness F, Saube SJ. Genomic Clustering and Homology between HET-S and the NWD2 STAND Protein in Various Fungal Genomes. *PLoS ONE*. 2012; 7(4):e34854. <https://doi.org/10.1371/journal.pone.0034854> PMID: 22493719
38. Daskalov A, Habenstein B, Sabate R, Berbon M, Martinez D, Chaignepain S, et al. Identification of a novel cell death-inducing domain reveals that fungal amyloid-controlled programmed cell death is related to necroptosis. *Proceedings of the National Academy of Sciences of the United States of America*. 2016; 113(10):2720–2725. <https://doi.org/10.1073/pnas.1522361113> PMID: 26903619
39. Wang J, Hu M, Wang J, Qi J, Han Z, Wang G, et al. Reconstitution and structure of a plant NLR resistosome conferring immunity. *Science*. 2019; 364(6435):eaav5870. <https://doi.org/10.1126/science.aav5870> PMID: 30948527
40. Bi G, Su M, Li N, Liang Y, Dang S, Xu J, et al. The ZAR1 resistosome is a calcium-permeable channel triggering plant immune signaling. *Cell*. 2021; 184(13):3528–3541.e12. <https://doi.org/10.1016/j.cell.2021.05.003> PMID: 33984278
41. Daskalov A, Habenstein B, Martinez D, Debets AJ, Sabate R, Loquet A, et al. Signal transduction by a fungal NOD-like receptor based on propagation of a prion amyloid fold. *PLoS Biology*. 2015; 13(2):e1002059. <https://doi.org/10.1371/journal.pbio.1002059> PMID: 25671553

42. Wasmer C, Lange A, Van Melckebeke H, Siemer AB, Riek R, Meier BH. Amyloid Fibrils of the HET-s (218–289) Prion Form a Beta Solenoid with a Triangular Hydrophobic Core. *Science*. 2008; 319(5869):1523–1526. <https://doi.org/10.1126/science.1151839> PMID: 18339938
43. van Melckebeke H, Wasmer C, Lange A, AB E, Loquet A, Bockmann A, et al. Atomic-Resolution Three-Dimensional Structure of HET-s(218–289) Amyloid Fibrils by Solid-State NMR Spectroscopy. *Journal of the American Chemical Society*. 2010; 132(39):13765–13775. <https://doi.org/10.1021/ja104213j> PMID: 20828131
44. Seuring C, Greenwald J, Wasmer C, Wepf R, Saupe SJ, Meier BH, et al. The mechanism of toxicity in HET-S/HET-s prion incompatibility. *PLoS Biology*. 2012; 10(12):e1001451. <https://doi.org/10.1371/journal.pbio.1001451> PMID: 23300377
45. Daskalov A, Dyrka W, Saupe SJ. Theme and variations: evolutionary diversification of the HET-s functional amyloid motif. *Scientific Reports*. 2015; 5:12494. <https://doi.org/10.1038/srep12494> PMID: 26219477
46. Daskalov A, Martinez D, Coustou V, El Mammeri N, Berbon M, Andreas LB, et al. Structural and molecular basis of cross-seeding barriers in amyloids. *Proceedings of the National Academy of Sciences*. 2021; 118(1). <https://doi.org/10.1073/pnas.2014085118> PMID: 33443172
47. Graziani S, Silar P, Daboussi M. Bistability and hysteresis of the 'Secteur' differentiation are controlled by a two-gene locus in *Nectria haematococca*. *BMC Biology*. 2004; 2:18. <https://doi.org/10.1186/1741-7007-2-18> PMID: 15312233
48. Sun X, Yin J, Starovasnik MA, Fairbrother WJ, Dixit VM. Identification of a novel homotypic interaction motif required for the phosphorylation of receptor-interacting protein (RIP) by RIP3. *The Journal of Biological Chemistry*. 2002; 277(11):9505–9511. <https://doi.org/10.1074/jbc.M109488200> PMID: 11734559
49. Rebsamen M, Heinz LX, Meylan E, Michallet MC, Schroder K, Hofmann K, et al. DAI/ZBP1 recruits RIP1 and RIP3 through RIP homotypic interaction motifs to activate NF- κ B. *EMBO reports*. 2009; 10(8):916–922. <https://doi.org/10.1038/embor.2009.109> PMID: 19590578
50. Kajava AV, Klopffleisch K, Chen S, Hofmann K. Evolutionary link between metazoan RHIM motif and prion-forming domain of fungal heterokaryon incompatibility factor HET-s/HET-s. *Scientific Reports*. 2014; 4(1):1–6. <https://doi.org/10.1038/srep07436> PMID: 25500536
51. Ahmed AB, Znassi N, Chateau MT, Kajava AV. A structure-based approach to predict predisposition to amyloidosis. *Alzheimer's & Dementia*. 2015; 11(6):681–690. <https://doi.org/10.1016/j.jalz.2014.06.007> PMID: 25150734
52. Dyrka W, Gašior-Gogowska M, Szefczyk M. Searching for universal model of amyloid signaling motifs using probabilistic context-free grammars. *BMC Bioinformatics*. 2021; 22:222. <https://doi.org/10.1186/s12859-021-04139-y> PMID: 33926372
53. Clave C, Dyrka W, Turcotte EA, Granger-Farbos A, Ibarlosa L, Pinson B, et al. Fungal gasdermin-like proteins are controlled by proteolytic cleavage. *Proceedings of the National Academy of Sciences*. 2022; 119(7):e2109418119. Personal statement: WD considers the use of the HEK293T embryonic cell line should be avoided due to its origin from an aborted human foetus. <https://doi.org/10.1073/pnas.2109418119> PMID: 35135876
54. Johnson AG, Wein T, Mayer ML, Duncan-Lowey B, Yirmiya E, Oppenheimer-Shaanan Y, et al. Bacterial gasdermins reveal an ancient mechanism of cell death. *Science*. 2022; 375(6577):221–225. <https://doi.org/10.1126/science.abj8432> PMID: 35025633
55. Gluck-Thaler E, Ralston T, Konkel Z, Ocampos CG, Ganeshan VD, Dorrance AE, et al. Giant Starship Elements Mobilize Accessory Genes in Fungal Genomes. *Molecular Biology and Evolution*. 2022; 39(5). <https://doi.org/10.1093/molbev/msac109> PMID: 35588244
56. Li J, McQuade T, Siemer AB, Napetschnig J, Moriwaki K, Hsiao YS, et al. The RIP1/RIP3 necrosome forms a functional amyloid signaling complex required for programmed necrosis. *Cell*. 2012; 150(2):339–350. <https://doi.org/10.1016/j.cell.2012.06.019> PMID: 22817896
57. Kleino A, Ramia NF, Bozkurt G, Shen Y, Nailwal H, Huang J, et al. Peptidoglycan-Sensing Receptors Trigger the Formation of Functional Amyloids of the Adaptor Protein Imd to Initiate *Drosophila* NF- κ B Signaling. *Immunity*. 2017; 47(4):635–647.e6. <https://doi.org/10.1016/j.immuni.2017.09.011> PMID: 29045898
58. Steinegger M, Soeding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*. 2017; 35:1026–1028. <https://doi.org/10.1038/nbt.3988> PMID: 29035372
59. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*. 2014; 31(6):926–932. <https://doi.org/10.1093/bioinformatics/btu739> PMID: 25398609

60. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Soding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*. 2016; 45(D1):D170–D176. <https://doi.org/10.1093/nar/gkw1081> PMID: 27899574
61. Remmert M, Biegert A, Hauser A, Soeding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*. 2012; 9(2):173–175. <https://doi.org/10.1038/nmeth.1818>
62. Oliva R, Win J, Raffaele S, Boutemy L, Bozkurt TO, Chaparro-Garcia A, et al. Recent developments in effector biology of filamentous plant pathogens. *Cellular Microbiology*. 2010; 12(6):705–715. <https://doi.org/10.1111/j.1462-5822.2010.01471.x> PMID: 20374248
63. Liu T, Ye W, Ru Y, Yang X, Gu B, Tao K, et al. Two host cytoplasmic effectors are required for pathogenesis of *Phytophthora sojae* by suppression of host defenses. *Plant physiology*. 2011; 155(1):490–501. <https://doi.org/10.1104/pp.110.166470> PMID: 21071601
64. Zhang D, Burroughs AM, Vidal ND, Iyer LM, Aravind L. Transposons to toxins: the provenance, architecture and diversification of a widespread class of eukaryotic effectors. *Nucleic Acids Research*. 2016; 44(8):3513–3533. <https://doi.org/10.1093/nar/gkw221> PMID: 27060143
65. Voss S, Betz R, Heidt S, Corradi N, Requena N, RiCRN1, a Crinkler Effector From the Arbuscular Mycorrhizal Fungus *Rhizophagus irregularis*, Functions in Arbuscule Development. *Frontiers in microbiology*. 2018; 9:2068. <https://doi.org/10.3389/fmicb.2018.02068> PMID: 30233541
66. Hetmann A, Kowalczyk S. Supresja odporności podstawowej typu PTI przez syntetyzowane w fitopatogenach białka efektorowe wprowadzane do wnętrza komorek infekowanej rośliny. *Postępy Biochemii*. 2019; 65(1):58–71. https://doi.org/10.18388/pb.2019_257
67. Grimshaw SJ, Mott HR, Stott KM, Nielsen PR, Everts KA, Hopkins LJ, et al. Structure of the sterile alpha motif (SAM) domain of the *Saccharomyces cerevisiae* mitogen-activated protein kinase pathway-modulating protein STE50 and analysis of its interaction with the STE11 SAM. *The Journal of Biological Chemistry*. 2004; 279(3):2192–2201. <https://doi.org/10.1074/jbc.M305605200> PMID: 14573615
68. Ray S, Chee L, Matson DR, Palermo NY, Bresnick EH, Hewitt KJ. Sterile α -motif domain requirement for cellular signaling and survival. *Journal of Biological Chemistry*. 2020; 295(20):7113–7125. <https://doi.org/10.1074/jbc.RA119.011895> PMID: 32241909
69. O'Neill LAJ, Bowie AG. The family of five: TIR-domain-containing adaptors in Toll-like receptor signaling. *Nature reviews Immunology*. 2007; 7(5):353–364. <https://doi.org/10.1038/nri2079> PMID: 17457343
70. Zhang Q, Zmasek CM, Cai X, Godzik A. TIR domain-containing adaptor SARM is a late addition to the ongoing microbe-host dialog. *Developmental and comparative immunology*. 2011; 35(4):461–468. <https://doi.org/10.1016/j.dci.2010.11.013> PMID: 21110998
71. Peterson ND, Icsó JD, Salisbury JE, Rodriguez T, Thompson PR, Pukkila-Worley R. Pathogen infection and cholesterol deficiency activate the *C. elegans* p38 immune pathway through a TIR-1/SARM1 phase transition. *eLife*. 2022; 11:e74206. <https://doi.org/10.7554/eLife.74206> PMID: 35098926
72. Detke S. Cloning of the *Candida albicans* nucleoside transporter by complementation of nucleoside transport-deficient *Saccharomyces*. *Yeast*. 1998; 14(14):1257–1265. [https://doi.org/10.1002/\(SICI\)1097-0061\(199810\)14:14%3C1257::AID-YEA326%3E3.0.CO;2-6](https://doi.org/10.1002/(SICI)1097-0061(199810)14:14%3C1257::AID-YEA326%3E3.0.CO;2-6) PMID: 9802205
73. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Research*. 2019; 48(D1):D265–D268. <https://doi.org/10.1093/nar/gkz991>
74. Soeding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics*. 2005; 21(7):951–960. <https://doi.org/10.1093/bioinformatics/bti125>
75. Zimmermann L, Stephens A, Nam SZ, Rau D, Kubler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology*. 2018; 430(15):2237–2243. <https://doi.org/10.1016/j.jmb.2017.12.007> PMID: 29258817
76. Rubbelke M, Fiegen D, Bauer M, Binder F, Hamilton J, King J, et al. Locking mixed-lineage kinase domain-like protein in its auto-inhibited state prevents necroptosis. *Proceedings of the National Academy of Sciences*. 2020; 117(52):33272–33281. <https://doi.org/10.1073/pnas.2017406117>
77. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021; 596(7873):583–589. <https://doi.org/10.1038/s41586-021-03819-2> PMID: 34265844
78. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nature Methods*. 2022; 19(6):679–682. <https://doi.org/10.1038/s41592-022-01488-1> PMID: 35637307

79. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*. 2005; 33(7):2302–2309. <https://doi.org/10.1093/nar/gki524> PMID: 15849316
80. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, et al. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Research*. 2005; 33(suppl_2):W299–W302. <https://doi.org/10.1093/nar/gki370> PMID: 15980475
81. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research*. 2016; 44(W1):W344–W350. <https://doi.org/10.1093/nar/gkw408> PMID: 27166375
82. Adachi H, Contreras MP, Harant A, Wu Ch, Derevnina L, Sakai T, et al. An N-terminal motif in NLR immune receptors is functionally conserved across distantly related plant species. *eLife*. 2019; 8:e49956. <https://doi.org/10.7554/eLife.49956> PMID: 31774397
83. McNamara DE, Dovey CM, Hale AT, Quarato G, Grace CR, Guibao CD, et al. Direct Activation of Human MLKL by a Select Repertoire of Inositol Phosphate Metabolites. *Cell Chemical Biology*. 2019; 26(6):863–877.e7. <https://doi.org/10.1016/j.chembiol.2019.03.010> PMID: 31031142
84. Novatchkova M, Leibbrandt A, Werzowa J, Neubuser A, Eisenhaber F. The STIR-domain superfamily in signal transduction, development and immunity. *Trends in biochemical sciences*. 2003; 28(5):226–229. [https://doi.org/10.1016/S0968-0004\(03\)00067-7](https://doi.org/10.1016/S0968-0004(03)00067-7) PMID: 12765832
85. Wu B, Gong J, Liu L, Li T, Wei T, Bai Z. Evolution of prokaryotic homologues of the eukaryotic SEFIR protein domain. *Gene*. 2012; 492(1):160–166. <https://doi.org/10.1016/j.gene.2011.10.033> PMID: 22037611
86. Jacob F, Vernaldi S, Maekawa T. Evolution and Conservation of Plant NLR Functions. *Frontiers in Immunology*. 2013; 4:297. <https://doi.org/10.3389/fimmu.2013.00297> PMID: 24093022
87. Guo Y, Narisawa K. Fungus-Bacterium Symbionts Promote Plant Health and Performance. *Microbes and Environments*. 2018; 33(3):239–241. <https://doi.org/10.1264/jsme2.ME3303rh> PMID: 30270261
88. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*. 2009; 37(suppl_2):W202–W208. <https://doi.org/10.1093/nar/gkp335> PMID: 19458158
89. Alkan N, Meng X, Friedlander G, Reuveni E, Sukno S, Sherman A, et al. Global Aspects of pacC Regulation of Pathogenicity Genes in *Colletotrichum gloeosporioides* as Revealed by Transcriptome Analysis. *Molecular Plant-Microbe Interactions*. 2013; 26(11):1345–1358. <https://doi.org/10.1094/MPMI-03-13-0080-R> PMID: 23902260
90. Sarroukh R, Goormaghtigh E, Ruyschaert JM, Raussens V. ATR-FTIR: a “rejuvenated” tool to investigate amyloid proteins. *Biochimica et biophysica acta*. 2013; 1828(10):2328–2338. <https://doi.org/10.1016/j.bbamem.2013.04.012> PMID: 23746423
91. Shivu B, Seshadri S, Li J, Oberg KA, Uversky VN, Fink AL. Distinct β -Sheet Structure in Protein Aggregates Determined by ATR-FTIR Spectroscopy. *Biochemistry*. 2013; 52(31):5176–5183. <https://doi.org/10.1021/bi400625v> PMID: 23837615
92. Ruyschaert JM, Raussens V. ATR-FTIR analysis of amyloid proteins. *Methods Mol Biol*. 2018; 1777:69–81. https://doi.org/10.1007/978-1-4939-7811-3_3 PMID: 29744828
93. Ruggeri FS, Sneideris T, Vendruscolo M, Knowles TPJ. Atomic force microscopy for single molecule characterisation of protein aggregation. *Archives of Biochemistry and Biophysics*. 2019; 664:134–148. <https://doi.org/10.1016/j.abb.2019.02.001> PMID: 30742801
94. Biancalana M, Koide S. Molecular mechanism of Thioflavin-T binding to amyloid fibrils. *Biochimica et Biophysica Acta (BBA) Proteins and Proteomics*. 2010; 1804(7):1405–1412. <https://doi.org/10.1016/j.bbapap.2010.04.001> PMID: 20399286
95. Xue C, Lin TY, Chang D, Guo Z. Thioflavin T as an amyloid dye: fibril quantification, optimal concentration and effect on aggregation. *Royal Society Open Science*. 2017; 4(1):160696. <https://doi.org/10.1098/rsos.160696> PMID: 28280572
96. Nilsson MR. Techniques to study amyloid fibril formation in vitro. *Methods*. 2004; 34(1):151–160. <https://doi.org/10.1016/j.ymeth.2004.03.012> PMID: 15283924
97. Li H, Rahimi F, Sinha S, Maiti P, Bitan G, Murakami K. Amyloids and Protein Aggregation Analytical Methods. In: Meyers RA, editor. *Encyclopedia of Analytical Chemistry*; 2009.
98. Martins PM, Navarro S, Silva A, Pinto MF, Sarkany Z, Figueiredo F, et al. MIRRAGGE Minimum Information Required for Reproducible AGGregation Experiments. *Frontiers in Molecular Neuroscience*. 2020; 13:222. <https://doi.org/10.3389/fnmol.2020.582488> PMID: 33328883
99. Wilkosz N, Czaja M, Seweryn S, Skirlińska-Nosek K, Szymonski M, Lipiec E, et al. Molecular Spectroscopic Markers of Abnormal Protein Aggregation. *Molecules*. 2020; 25(11). <https://doi.org/10.3390/molecules25112498> PMID: 32471300

100. Khurana R, Fink AL. Do Parallel β -Helix Proteins Have a Unique Fourier Transform Infrared Spectrum? *Biophysical Journal*. 2000; 78(2):994–1000. [https://doi.org/10.1016/S0006-3495\(00\)76657-4](https://doi.org/10.1016/S0006-3495(00)76657-4) PMID: 10653812
101. Zou Y, Li Y, Hao W, Hu X, Ma G. Parallel β -Sheet Fibril and Antiparallel β -Sheet Oligomer: New Insights into Amyloid Formation of Hen Egg White Lysozyme under Heat and Acidic Condition from FTIR Spectroscopy. *The Journal of Physical Chemistry B*. 2013; 117(15):4003–4013. <https://doi.org/10.1021/jp4003559> PMID: 23537140
102. Grelich-Mucha M, Garcia AM, Torbeev V, Özga K, Berlicki , Olesiak-Bañska J. Autofluorescence of Amyloids Determined by Enantiomeric Composition of Peptides. *The Journal of Physical Chemistry B*. 2021; 125(21):5502–5510. <https://doi.org/10.1021/acs.jpcc.1c00808> PMID: 34008978
103. Berthelot K, Ta HP, Gean J, Lecomte S, Cullin C. In Vivo and In Vitro Analyses of Toxic Mutants of HET-s: FTIR Antiparallel Signature Correlates with Amyloid Toxicity. *Journal of Molecular Biology*. 2011; 412(1):137–152. <https://doi.org/10.1016/j.jmb.2011.07.009> PMID: 21782829
104. Requena JR, Wille H. The structure of the infectious prion protein. *Prion*. 2014; 8(1):60–66. <https://doi.org/10.4161/pri.28368> PMID: 24583975
105. Moran SD, Zanni MT. How to Get Insight into Amyloid Structure and Formation from Infrared Spectroscopy. *The journal of physical chemistry letters*. 2014; 5(11):1984–1993. <https://doi.org/10.1021/jz500794d> PMID: 24932380
106. Bardin T, Daskalov A, Barrouilhet S, Granger-Farbos A, Salin B, Blancard C, et al. Partial Prion Cross-Seeding between Fungal and Mammalian Amyloid Signaling Motifs. *mBio*. 2021; 12(1):e02782–20. <https://doi.org/10.1128/mBio.02782-20> PMID: 33563842
107. Paoletti M, Clave C. The Fungus-Specific HET Domain Mediates Programmed Cell Death in *Podospora anserina*. *Eukaryotic Cell*. 2007; 6(11):2001–2008. <https://doi.org/10.1128/EC.00129-07> PMID: 17873080
108. Dyrka W, Nebel JC. A Stochastic Context Free Grammar based Framework for Analysis of Protein Sequences. *BMC Bioinformatics*. 2009; 10:323. <https://doi.org/10.1186/1471-2105-10-323> PMID: 19814800
109. Wilburn GW, Eddy SR. Remote homology search with hidden Potts models. *PLOS Computational Biology*. 2020; 16(11):1–22. <https://doi.org/10.1371/journal.pcbi.1008085> PMID: 33253143
110. Muntoni AP, Pagnani A, Weigt M, Zamponi F. Aligning biological sequences by exploiting residue conservation and coevolution. *Phys Rev E*. 2020; 102:062409. <https://doi.org/10.1103/PhysRevE.102.062409> PMID: 33465950
111. Talibart H, Coste F. PPAalign: optimal alignment of Potts models representing proteins with direct coupling information. *BMC Bioinformatics*. 2021; 22(1):317. <https://doi.org/10.1186/s12859-021-04222-4> PMID: 34112081
112. Merrill W, Weiss G, Goldberg Y, Schwartz R, Smith NA, Yahav E. A Formal Hierarchy of RNN Architectures. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics; 2020. p. 443–459.
113. Bhattamishra S, Patel A, Goyal N. On the Computational Power of Transformers and Its Implications in Sequence Modeling. In: *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics; 2020. p. 455–475.
114. Merrill W. Formal Language Theory Meets Modern NLP. *CoRR*. 2021;abs/2102.10094.
115. Nambiar A, Heflin M, Liu S, Maslov S, Hopkins M, Ritz A. Transforming the Language of Life: Transformer Neural Networks for Protein Prediction Tasks. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. BCB'20. New York, NY, USA: Association for Computing Machinery; 2020.
116. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics*. 2022; 38(8):2102–2110. <https://doi.org/10.1093/bioinformatics/btac020> PMID: 35020807
117. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 1997; 25(17):3389–402. <https://doi.org/10.1093/nar/25.17.3389> PMID: 9254694
118. Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*. 2011; 7:539. <https://doi.org/10.1038/msb.2011.75> PMID: 21988835
119. Consortium TU. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*. 2020; 49(D1):D480–D489. <https://doi.org/10.1093/nar/gkaa1100>
120. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. *Nucleic Acids Research*. 2018; 46(W1):W200–W204. <https://doi.org/10.1093/nar/gky448> PMID: 29905871

121. Dyrka W, Pyzik M, Coste F, Talibert H. Estimating probabilistic context-free grammars for proteins using contact map constraints. *PeerJ*. 2019; 7:e6559. <https://doi.org/10.7717/peerj.6559> PMID: 30918754
122. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006; 22:1658–1659. <https://doi.org/10.1093/bioinformatics/btl158> PMID: 16731699
123. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012; 28(23):3150–3152. <https://doi.org/10.1093/bioinformatics/bts565> PMID: 23060610
124. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, California; 1994. p. 28–36.
125. Eddy SR. Accelerated Profile HMM Searches. *PLoS Computational Biology*. 2011; 7(10):e1002195. <https://doi.org/10.1371/journal.pcbi.1002195> PMID: 22039361
126. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*. 2013; 30:772–80. <https://doi.org/10.1093/molbev/mst010> PMID: 23329690
127. Capella-Gutierrez S, Silla-Martinez JM, T G. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009; 25:1972–3. <https://doi.org/10.1093/bioinformatics/btp348> PMID: 19505945
128. Coordinators NR. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. 2017; 46(D1):D8–D13. <https://doi.org/10.1093/nar/gkx1095>
129. Harrison PW, Ahamed A, Aslam R, Alako BTF, Burgin J, Buso N, et al. The European Nucleotide Archive in 2020. *Nucleic Acids Research*. 2020; 49(D1):D82–D85. <https://doi.org/10.1093/nar/gkaa1028>
130. Reitz K. Requests: HTTP for Humans; 2022. Available from <https://requests.readthedocs.io/>
131. Blech M. xmltodict: Python module that makes working with XML feel like you are working with JSON; 2019. Available from <https://github.com/martinblech/xmltodict>
132. McKinney W. Data Structures for Statistical Computing in Python. In: St fan van der Walt, Jarrod Millman, editors. *Proceedings of the 9th Python in Science Conference*; 2010. p. 56–61.
133. McKinney W. pandas—Python Data Analysis Library; 2021. Available from: <https://pandas.pydata.org>.
134. Hunter JD. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*. 2007; 9(3):90–95. <https://doi.org/10.1109/MCSE.2007.55>
135. Waskom ML. seaborn: statistical data visualization. *Journal of Open Source Software*. 2021; 6(60):3021. <https://doi.org/10.21105/joss.03021>
136. Beitz E. TeXshade: shading and labeling of multiple sequence alignments using LaTeX2e. *Bioinformatics*. 2000; 16(2):135–139. <https://doi.org/10.1093/bioinformatics/16.2.135> PMID: 10842735
137. Gansner ER, North SC. An open graph visualization system and its applications to software engineering. *Software Practice and Experience*. 2000; 30(11):1203–33. [https://doi.org/10.1002/1097-024X\(200009\)30:11%3C1203::AID-SPE338%3E3.0.CO;2-N](https://doi.org/10.1002/1097-024X(200009)30:11%3C1203::AID-SPE338%3E3.0.CO;2-N)
138. Sayle RA, Milner-White EJ. RASMOL: biomolecular graphics for all. *Trends in Biochemical Sciences*. 1995; 20(9):374–376. [https://doi.org/10.1016/S0968-0004\(00\)89080-5](https://doi.org/10.1016/S0968-0004(00)89080-5) PMID: 7482707
139. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*. 1964; 36:1627–1639. <https://doi.org/10.1021/ac60214a047>
140. Balguerie A, Dos Reis S, Couлары-Salin B, Chaignepain S, Sabourin M, Schmitter JM, et al. The sequences appended to the amyloid core region of the HET-s prion protein determine higher-order aggregate organization in vivo. *Journal of Cell Science*. 2004; 117(12):2599–2610. <https://doi.org/10.1242/jcs.011116> PMID: 15159455
141. Bergès T, Barreau C. Heat Shock at an Elevated Temperature Improves Transformation Efficiency of Protoplasts from *Podospora anserina*. *Microbiology*. 1989; 135(3):601–604. <https://doi.org/10.1099/00221287-135-3-601> PMID: 2621442
142. Benkemoun L, Sabate R, Malato L, Reis SD, Dalstra H, Saupe SJ, et al. Methods for the in vivo and in vitro analysis of [Het-s] prion infectivity. *Methods*. 2006; 39(1):61–67. <https://doi.org/10.1016/j.ymeth.2006.04.006> PMID: 16750391

AmyloGraph: a comprehensive database of amyloid–amyloid interactions

Michał Burdukiewicz^{1,2,*}, Dominik Rafacz³, Agnieszka Barbach⁴, Katarzyna Hubicka⁴, Laura Bąkała³, Anna Lassota⁵, Jakub Stecko⁶, Natalia Szymańska⁶, Jakub W. Wojciechowski⁴, Dominika Kozakiewicz⁷, Natalia Szulc⁴, Jarosław Chilimoniuk⁸, Izabela Jęskowiak⁹, Marlena Gąsior-Głogowska⁴ and Małgorzata Kotulska⁴

¹Institute of Biotechnology and Biomedicine, Autonomous University of Barcelona, Campus Universitat Autònoma de Barcelona Plaça Cívica Bellaterra, s/n, 08193 Cerdanyola del Vallès, Barcelona, Spain, ²Clinical Research Centre, Medical University of Białystok, Kilińskiego 1, 15-369 Białystok, Poland, ³Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland, ⁴Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland, ⁵School of Biosciences, College of Life and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom, ⁶Faculty of Medicine, Wrocław Medical University, Ludwika Pasteura 1, 50-367 Wrocław, Poland, ⁷Laboratory of Microbiome Immunobiology, Hirszfeld Institute of Immunology and Experimental Therapy, Polish Academy of Sciences, Weigla 12, 53-114 Wrocław, Poland, ⁸Department of Genomics, Faculty of Biotechnology, University of Wrocław, Fryderyka Joliot-Curie 14a, 50-383 Wrocław, Poland and ⁹Department of Pharmacology, Wrocław Medical University, Mikulicza-Radeckiego 2, 50-345 Wrocław, Poland

Received August 14, 2022; Revised September 22, 2022; Editorial Decision September 27, 2022; Accepted September 30, 2022

ABSTRACT

Information about the impact of interactions between amyloid proteins on their fibrillization propensity is scattered among many experimental articles and presented in unstructured form. We manually curated information located in almost 200 publications (selected out of 562 initially considered), obtaining details of 883 experimentally studied interactions between 46 amyloid proteins or peptides. We also proposed a novel standardized terminology for the description of amyloid–amyloid interactions, which is included in our database, covering all currently known types of such a cross-talk, including inhibition of fibrillization, cross-seeding and other phenomena. The new approach allows for more specific studies on amyloids and their interactions, by providing very well-defined data. AmyloGraph, an online database presenting information on amyloid–amyloid interactions, is available at (<http://AmyloGraph.com/>). Its functionalities are also accessible as the R package (<https://github.com/KotulskaLab/AmyloGraph>). AmyloGraph is the

only publicly available repository for experimentally determined amyloid–amyloid interactions.

INTRODUCTION

Amyloids are proteins able to self-assembly into insoluble β -sheet supra-molecular fibrils characterized by very regular beta-cross structures. Some of them interact with each other during fibrillization, which may accelerate or slow down development of fibrils or even lead to the formation of heterogeneous fibrils (1). Interactions between amyloid proteins raise a growing interest since they may contribute to amyloid-related diseases. The aggregation of amyloid fibrils can be associated with pathologies observed in a wide range of diseases known as amyloidoses. For example, amyloid fibrils which aggregate in the brain and central nervous system, are related to Alzheimer's and Parkinson's diseases (2). Another example is a prion conversion, where ingested misfolded proteins can seed the aggregation of their homologous polypeptide sequence (3). Similar mechanisms can trigger other amyloidoses (4), but experimental data considering such phenomena are dispersed and often very incompatible.

The importance of interactions between amyloid proteins makes them a subject of numerous experimental studies.

*To whom correspondence should be addressed. Email: michalburdukiewicz@gmail.com

However, reviews on interactions between amyloid proteins show that available experimental results are often contradictory (5). Although the information on amyloid proteins is collected in several databases (6–10), until now there has been no database consolidating the results of numerous experiments studying interactions of amyloids. Importantly, existing efforts to present the overview of amyloid–amyloid interactions do not allow for a more in-depth inspection of data (11).

Other problems arise from the lack of clear definitions of field. Although there have been attempts to standardize the vocabulary (12,13) or a list of requirements necessary in reporting amyloid studies (14), the practices are still not being fully implemented. Therefore, comparing different studies, especially those regarding amyloid–amyloid interactions, is problematic and even fundamental concepts may be understood incompatibly.

Therefore, we designed a structured vocabulary to describe amyloid–amyloid interactions more rigorously. It covers descriptors that fully define the exact nature of the influence of an interactor on an interactee. Using the proposed methodology, we manually curated a majority of reported interactions between amyloids and presented this information in the form of an interactive graph and a tabular database.

MATERIALS AND METHODS

Standardized terminology

To describe interactions between amyloid proteins, we created a precisely controlled vocabulary. First, we defined six possible scenarios of amyloid–amyloid interactions (Figure 1A). All scenarios assume that there are only two participants in each interaction, and an interactor modulates self-assembly of an interactee. We are aware that in reality the distinction between the interactee and interactor may not always be clear or, depending on other factors, a specific interaction can fall under more than one scenario at the same time. However, this simplification allowed us to better design the standardized terminology and led to a more structured description of interactions between amyloid proteins.

Next, we developed three descriptors to more rigorously describe details of the scenarios, based on one of the following: the fibrillization speed, presence of physical binding between both interacting proteins and appearance of heterogeneous fibrils (Figure 1B).

Each descriptor provides specific dictionary terms of possible states. For example, descriptor I, ‘The impact on the speed of fibrillization’, enables the choice of one of the following states: ‘faster fibrillization’, ‘slower fibrillization’, ‘no fibrillization’, ‘no effect’ and ‘no information’. The designed states are mutually exclusive and provide in-depth description to relate them to relevant experimental results (Supplementary Information, section *Descriptors* and Supplementary Figure S1).

It is essential to stress that most publications provided information only on the interactee’s (a protein whose self-assembly is modulated by interactee) ability to create amyloid-like fibrils. Therefore, the descriptors focus only on the behavior of interactee. However, some manuscripts reported on the self-assembly of both interactee and interac-

tor. In this case, such interactions were reported bidirectionally, where protein A acts as an interactor and protein B as the interactee, and *vice versa*.

Our descriptors do not replace existing terminology, but rather standardize it. For example, a combination of answers to descriptor I ‘The impact on the speed of fibrillization’: ‘Faster aggregation’ and descriptor II ‘Physical binding between interactor and interactee’: ‘Yes, direct evidence’ or ‘Yes, implied by kinetics’ can be related to either cross-seeding or co-incubation (12). These two experiments are drastically different design-wise: co-incubation requires both interactee and interactor to be in monomeric form, while in cross-seeding experiment interactee is monomeric and interactor in the form of small aggregates. However, studies frequently do not mention the exact form of interactee and interactor making general descriptors easier to use and more accurate.

Database scope

The scope of the current AmyloGraph version is limited to interactions between two proteins, each of them able to form an amyloid-like aggregate by its wild type. Additionally, we allowed for non-aggregating homologs of a well-known amyloid protein, such as rat amylin (15).

While selecting the source publications, we focused on *in vitro* studies published after 2000. The complete list of eligibility criteria is available in the Supplementary Information, section *Manuscript collection*.

Data acquisition and curation

To ensure the highest possible quality of the collected data, the data acquisition and curation were executed in a three-stage pipeline, including: pre-screen of manuscripts, manual curation and independent final validation. Importantly, the first two steps were supported by dedicated forms which played a crucial role in standardizing annotations provided by curators.

The pre-screen of manuscripts started with our in-house collection of 24 publications. Next, we expanded the search by repeatedly adding manuscripts cited by manuscripts or referencing manuscripts from our collections. The final collection included 562 manuscripts, out of which 364 were putatively suitable for the database. Although our collection system was laborious, we found it to be more effective than a search of PubMed records based on its annotations (Supplementary Figure S2A).

Next, the database curators manually extracted information on interactions from the suitable manuscripts. It should be emphasized that, in the curation procedure, we did not re-interpret data and conclusions provided by their authors. Curators interpreted the data only if the authors did not provide a description of the results or if the existing description was too limited. To help curators interpret the results, we enhanced the descriptors with the specifics of experimental procedures, which helped them to identify a descriptor level and draw the best final conclusion. Moreover, during the project, we developed a FAQ list of almost 100 questions related to the curation procedure, which helped the curators. Thanks to all these precautions, the curators could provide data of a higher quality.

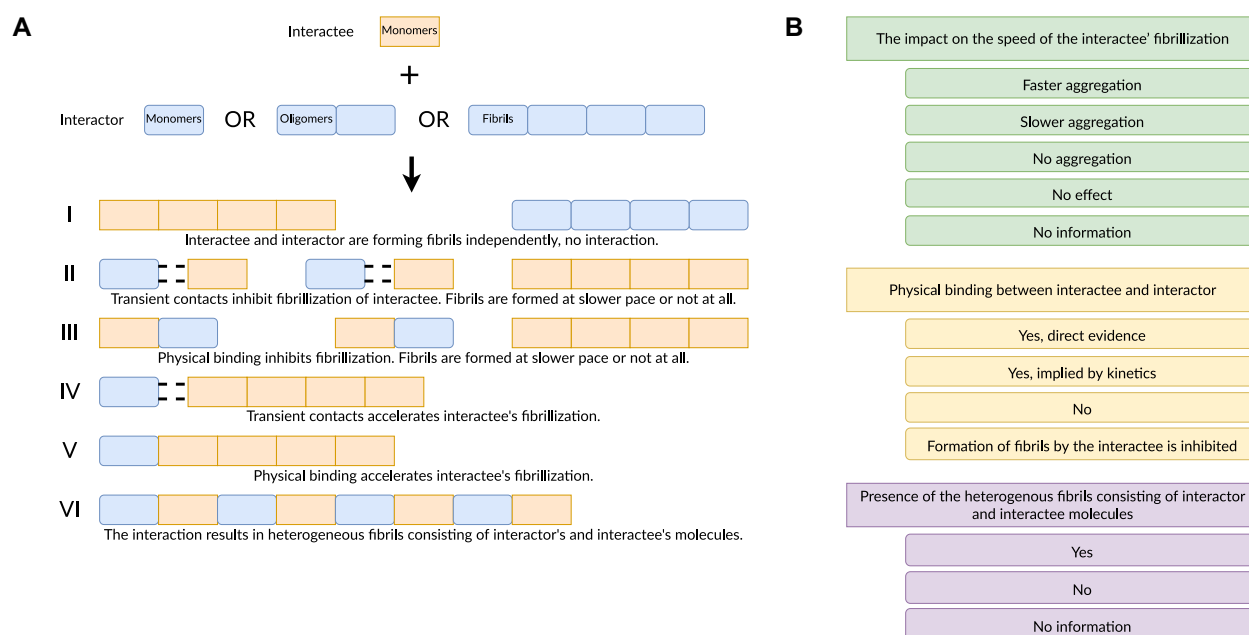


Figure 1. (A) Six scenarios of interactions between amyloid proteins. Orange and blue tiles denotes molecules of amyloid-like proteins participating in interactions. Roman numerals denote different interactions scenarios. (B) Three descriptors of AmyloGraph. Rectangles represent the descriptors. Rectangles with round edges represent the levels of the descriptors. Green, yellow and purple represent descriptors 1, 2 and 3, respectively.

During the initial curation, curators reviewed all collected manuscripts. Curators annotated these interactions using our three descriptors and collected information on sequences of proteins participating in the interaction, focusing on the presence of mutations or other alterations. Curators were also obliged to preserve parts of the publication (in the graphical or textual form) supporting their decisions regarding final description in the records.

After the initial curation, we validated all the collected data to further increase their quality. In this procedure, new curators reviewed the assigned interaction records. The semi-random assignment procedure ensured that the curator who validated a specific record was not involved in its initial curation. Finally, the correct records were accepted in the database. The manual curation resulted in 172 publications and 883 interactions (Supplementary Figure S2B).

Next, we contacted the authors of all 172 publications included in AmyloGraph to obtain their validation of our records. We always tried to reach the corresponding authors or, in case of their unavailability, the publication's first author. The authors were provided with customized links to Google Sheets only containing data from manuscripts they authored. In total, we contacted 122 authors, and 11 authors (9.04%) confirmed 81 interactions (9.17%) in 21 manuscripts (12.14%). Despite our efforts, we could not find correct and up-to-date contact information to authors of three manuscripts. It is important to notice that no interactions were removed or added after the contact with the authors, which implies a high data quality (Supplementary Figure S2C).

The in-depth description of the curation procedure is available in the Supplementary Information, section *Data acquisition*. The full list of 172 manuscripts is available in

the Supplementary Information, section *Supplementary references*.

Implementation

One of the main limitations of web-based tools is their in-built reliance on the external servers which reduces their persistence (16). Therefore, we made AmyloGraph fully deployable and usable even if the main server is no longer available. To do so, we implemented our tool as an R package (17). The package contains also the front-end of our database, available as a Shiny app (18). The local deployment of AmyloGraph only requires a very rudimentary knowledge of R and is described in the AmyloGraph main repository at <https://kotulskalab.github.io/AmyloGraph/>. The AmyloGraph codebase is open and documented in the roxygen2 standard.

DATABASE OVERVIEW

Manually curated data, obtained in the previously described procedure, are available in the AmyloGraph database. Currently, the database includes 883 interactions between 46 proteins reported in 172 manuscripts. Furthermore, one of the main objectives of the database is to present the interactions between amyloid proteins in a standardized manner and user-friendly presentations, such as a graph format (Figure 2A). Here, nodes represent individual amyloid proteins and edges stand for interactions between them. Notably, a single edge represents all interactions between two amyloid proteins. Tooltips of the edges represent digital object identifiers (DOIs) of manuscripts reporting the interactions.

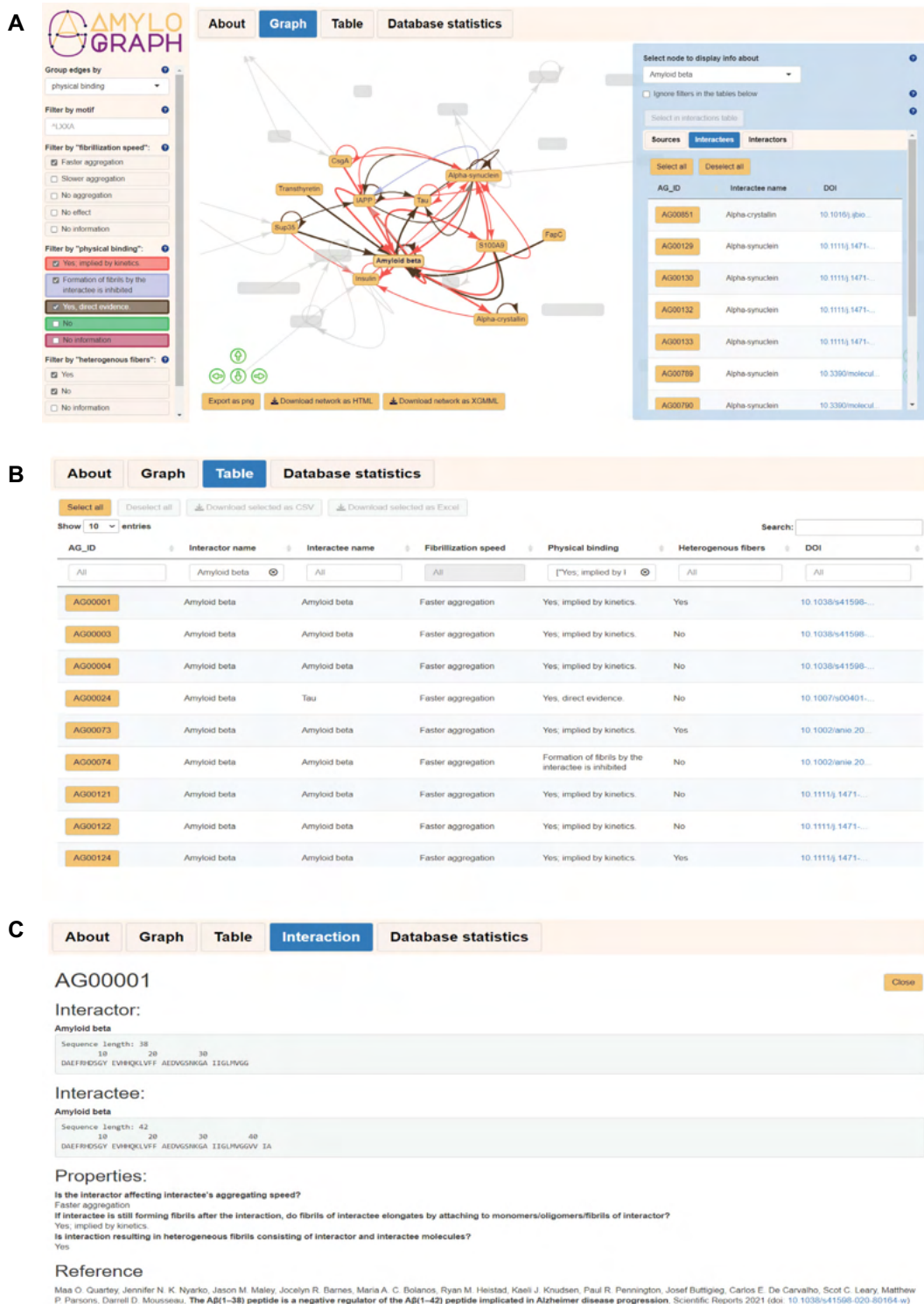


Figure 2. Overview of the AmyloGraph database. (A) Graph view of interactions between amyloid proteins. The interactions (edges of the graph) are colored according to the levels of descriptor 2, 'physical binding'. The panel on the right-hand side represents an overview of the amyloid-β interactions. (B) Tabular view of interactions. The top section of this card contains download options allowing to obtain data in a selected flat-file format. (C) View of a single interaction with the sequential information.

After clicking on a node, a panel on the right-hand side opens. It presents brief information on a protein, its name and links to its UniProt record. If a single amyloid protein in AmyloGraph is associated with several records in the UniProt, we provide links to all of them. This panel also contains two tables presenting all interactees and interactors of the protein.

Aside from the graph, AmyloGraph enables tabular representation of the interaction data (Figure 2B). The table is interactive and searchable. A user can also download selected rows in a flat-table format (.csv or .xlsx). As a result, the downloaded table contains all available information, including the sequences of amyloid proteins participating in the interactions.

Both, graph and tabular representations of the data can be filtered out using filters available on the left-hand side of the user interface. The filters cover all three descriptors. Moreover, a user can color the edges on the graph, according to the levels of a chosen descriptor. The user can also use amino acid sequences to filter the information presented in the graph or tabular form. Here, we implemented a simplified set of regular expression inspired by the POSIX system to facilitate more advanced searches.

The last card of the graphical interface, 'Interaction', opens when the user selects a specific interaction (Figure 2C). This view presents information on a single interaction, including levels of all descriptors and exact sequences of proteins. In case of multi-chain proteins, such as insulin, AmyloGraph presents the sequences of all chains.

To streamline the use of AmyloGraph, we enhanced it with helpers explaining basic functionalities of the database. Moreover, a video tutorial is available, presenting examples of AmyloGraph queries.

AmyloGraph is a FAIR-compliant database (19). All interactions are identifiable by an individual index. They are also linked to original publications using their DOIs provided by the Crossref. Proteins participating in interactions are linked to the UniProt database (20). As recommended by the FAIR guidelines, we extended the existing vocabulary to describe our data by fully providing our standardized methodology.

CONCLUSIONS AND FUTURE DIRECTIONS

AmyloGraph is the first endeavor to present an overview of experimentally verified interactions between amyloid proteins. It has also been the first attempt to standardize reporting of the amyloid-amyloid interactions and present them in the interactive database. We believe that, thanks to our rigorous data curation procedure, we have managed to collect and thoroughly systematize the majority of available information. Even though AmyloGraph is currently the most comprehensive compendium on the interactions between amyloid proteins, we see three areas that require an improvement: constant updates, representation of protein data and extending information of experimental conditions.

The greatest challenge regarding AmyloGraph, which we envisage, will be to keep it updated. We encountered a sudden influx of new publications reporting new interactions during our work on the database. To alleviate this issue,

AmyloGraph offers a submission form for authors involved in relevant research to report their results directly to the database. Thus, we are going to implement a highly structured system for finding publications by annotating records acquired in searches of the PubMed database.

One of other challenges regarding AmyloGraph is representation of protein data. Right now, AmyloGraph is very protein-centric and treats whole families of homologs or variants of a single protein as a single entity. The actual situation is much more complicated as we often deal with fragments of recombinant proteins or even protein grafts (21). In the future, we want to extend AmyloGraph to contain more information about the protein sequence and allow proteins with non-standard amino acids or even non-amino-acid modifications. This change is also necessary to add to AmyloGraph information on the impact of small molecules on amyloid fibrillization.

Another limitation of AmyloGraph, which we plan to alleviate in the next version, is the lack of experimental information. Amyloid assembly process and their interactions are extremely liable to experimental conditions, such as pH (22) or concentration of proteins (23). We are convinced that each record in AmyloGraph should be annotated with more parameters defining the environment of the interaction.

Even considering all limitations described above, we still believe that the current release of AmyloGraph is a valuable tool that provides access to a unique dataset. To our knowledge, AmyloGraph is the first effort to collect and present information on interactions between amyloid proteins in a unified format. AmyloGraph's high accessibility and data quality further enhance its usefulness.

DATA AVAILABILITY

The forms supporting collection of manuscripts, initial curation and validation of manuscripts are available upon request to the corresponding author. The procedure of data acquisition, the in-depth definitions of AmyloGraph descriptors and references to all curated manuscripts, are available in the Supplementary Information and online at <https://kotulskalab.github.io/AmyloGraph/articles/definitions.html>. AmyloGraph is available as an online database (<http://AmyloGraph.com/>).

CODE AVAILABILITY

All AmyloGraph functionalities are also accessible as the R package (<https://github.com/KotulskaLab/AmyloGraph>).

SUPPLEMENTARY DATA

[Supplementary Data](#) are available at NAR Online.

ACKNOWLEDGEMENTS

Access to Wroclaw Centre for Networking and Supercomputing at Wroclaw University of Science and Technology is greatly acknowledged. We also thank Daniel Otzen (Aarhus University, Denmark) and Vytautas Smirnovas (University of Vilnius, Lithuania) for fruitful discussions.

FUNDING

National Science Centre, Poland [2019/35/B/NZ2/03997]; M.B. was supported by the Maria Zambrano grant funded by the European Union-NextGenerationEU.
Conflict of interest statement. None declared.

REFERENCES

- Konstantoulea, K., Louros, N., Rousseau, F. and Schymkowitz, J. (2022) Heterotypic interactions in amyloid function and disease. *FEBS J.*, **289**, 2025–2046.
- Ivanova, M.I., Lin, Y., Lee, Y.-H., Zheng, J. and Ramamoorthy, A. (2021) Biophysical processes underlying cross-seeding in amyloid aggregation and implications in amyloid pathology. *Biophys. Chem.*, **269**, 106507.
- Gil-García, M., Iglesias, V., Pallarès, I. and Ventura, S. (2021) Prion-like proteins: from computational approaches to proteome-wide Analysis. *FEBS Open Biol.*, **11**, 2400–2417.
- Friedland, R.P. and Chapman, M.R. (2017) The role of microbial amyloid in neurodegeneration. *PLoS Path.*, **13**, e1006654.
- Tran, J., Chang, D., Hsu, F., Wang, H. and Guo, Z. (2017) Cross-seeding between A β 40 and A β 42 in Alzheimer's disease. *FEBS Lett.*, **591**, 177–185.
- Wozniak, P.P. and Kotulska, M. (2015) AmyLoad: Website dedicated to amyloidogenic protein fragments. *Bioinformatics (Oxford, England)*, **31**, 3395–3397.
- Louros, N., Konstantoulea, K., De Vleeschouwer, M., Ramakers, M., Schymkowitz, J. and Rousseau, F. (2020) WALTZ-DB 2.0: An updated database containing structural information of experimentally determined amyloid-forming peptides. *Nucleic Acids Res.*, **48**, D389–D393.
- Varadi, M., De Baets, G., Vranken, W.F., Tompa, P. and Pancsa, R. (2018) AmyPro: A database of proteins with validated amyloidogenic regions. *Nucleic Acids Res.*, **46**, D387–D392.
- Rawat, P., Prabakaran, R., Sakthivel, R., Thangakani, M.A., Kumar, S. and Gromiha, M. (2020) CPAD 2.0: A repository of curated experimental data on aggregating proteins and peptides. *Amyloid*, **27**, 128–133.
- Pawliski, S., Le Béche, A. and Delamarche, C. (2008) AMYPdb: A database dedicated to amyloid precursor proteins. *BMC Bioinform.*, **9**, 273.
- Biza, K.V., Nastou, K.C., Tsiolaki, P.L., Mastrokalou, C.V., Hamodrakas, S.J. and Iconomidou, V.A. (2017) The Amyloid interactome: exploring protein aggregation. *PLOS ONE*, **12**, e0173163.
- Bondarev, S.A., Antonets, K.S., Kajava, A.V., Nizhnikov, A.A. and Zhouravleva, G.A. (2018) Protein co-aggregation related to amyloids: methods of investigation, diversity, and classification. *Int. J. Mol. Sci.*, **19**, 2292.
- Ren, B., Zhang, Y., Zhang, M., Liu, Y., Zhang, D., Gong, X., Feng, Z., Tang, J., Chang, Y. and Zheng, J. (2019) Fundamentals of cross-seeding of amyloid proteins: An introduction. *J. Mat. Chem. B*, **7**, 7267–7282.
- Martins, P.M., Navarro, S., Silva, A., Pinto, M.F., Sárkány, Z., Figueiredo, F., Pereira, P.J.B., Pinheiro, F., Bednarikova, Z., Burdukiewicz, M. et al. (2020) MIRRAGGE – Minimum information required for reproducible aggregation experiments. *Front. Mol. Neurosci.*, **13**, 582488.
- Green, J., Goldsbury, C., Mini, T., Sunderji, S., Frey, P., Kistler, J., Cooper, G. and Aebi, U. (2003) Full-length rat amylin forms fibrils following substitution of single residues from human Amylin. *J. Mol. Biol.*, **326**, 1147–1156.
- Veretnik, S., Fink, J.L. and Bourne, P.E. (2008) Computational biology resources lack persistence and usability. *PLoS Comput. Biol.*, **4**, e1000136.
- R Core Team (2022) In: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A. and Borges, B. (2022) shiny: web application framework for R. R package version 1.7.2.
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E. et al. (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018.
- The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
- Azoitei, M.L., Correia, B.E., Ban, Y.-E.A., Carrico, C., Kalyuzhnyi, O., Chen, L., Schroeter, A., Huang, P.-S., McLellan, J.S., Kwong, P.D. et al. (2011) Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. *Science*, **334**, 373–376.
- Pfefferkorn, C.M., McGlinchey, R.P. and Lee, J.C. (2010) Effects of pH on aggregation kinetics of the repeat domain of a functional amyloid, Pmel17. *Proc. Natl. Acad. Sci.*, **107**, 21447–21452.
- Hu, X., Crick, S.L., Bu, G., Frieden, C., Pappu, R.V. and Lee, J.-M. (2009) Amyloid seeds formed by cellular uptake, concentration, and aggregation of the amyloid-beta peptide. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 20324–20329.

PACT - Prediction of Amyloid Cross-interaction by Threading

Jakub W. Wojciechowski^{1,*}, Witold Szczurek¹, Natalia Szulc^{1,2,3}, Monika Szefczyk⁴, and Malgorzata Kotulska^{1,*}

¹Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wrocław University of Science and Technology, 50-370 Wrocław, Poland

²Department of Physics and Biophysics, Wrocław University of Environmental and Life Sciences, Norwida 25, 50-375 Wrocław, Poland

³LPCT, CNRS, Université de Lorraine, F-54000 Nancy, France

⁴Department of Bioorganic Chemistry, Faculty of Chemistry, Wrocław University of Science and Technology, 50-370 Wrocław, Poland

*jakub.wojciechowski@pwr.edu.pl, malgorzata.kotulska@pwr.edu.pl

ABSTRACT

Amyloids are protein aggregates usually associated with their contribution to several diseases e.g., Alzheimer's and Parkinson's. However, they are also beneficially utilized by many organisms in physiological roles, such as microbial biofilm formation or hormone storage. Recent studies showed that an amyloid aggregate can affect aggregation of another protein. Such cross-interactions may be crucial for understanding the comorbidity of amyloid diseases or the influence of microbial amyloids on human amyloidogenic proteins. However, due to demanding experiments, understanding of interaction phenomena is still limited. Moreover, no dedicated computational method to predict potential amyloid interactions has been available until now. Here, we present PACT - a computational method for prediction of amyloid cross-interactions. The method is based on modeling a heterogenous fibril formed by two amyloidogenic peptides. The stability of the resulting structure is assessed using a statistical potential that approximates energetic stability of a model. Importantly, the method can work with long protein fragments and, as a purely physicochemical approach, it relies very little on training data. PACT was evaluated on data collected in the AmyloGraph database and it achieved high values of AUC (0.88) and F1 (0.82). The new method opens the possibility of high throughput studies of amyloid interactions. We used PACT to study interactions of CsgA, a bacterial biofilm protein from several bacterial species inhabiting human intestines, and human Alpha-synuclein protein which is involved in the onset of Parkinson's disease. We show that the method correctly predicted the interactions, performing experimental validation, and highlighted the importance of specific regions in both proteins.

The tool is available as a web server at: <https://pact.e-science.pl/pact/>. The local version can be downloaded from: <https://github.com/KubaWojciechowski/PACT>

Introduction

Pathological misfolding and aggregation of proteins is a hallmark of a number of devastating disorders, including major public health challenges like Alzheimer's and Parkinson's diseases^{1,2}, type II diabetes^{3,4}, as well as some cancers⁵. These diseases not only share a similar molecular mechanism, but they also often co-occur in the same patients. Among others, comorbidities were observed between Alzheimer's disease and type II diabetes^{6,7} and Alzheimer's and Parkinson's diseases⁸. One of the possible explanations of this phenomenon could be related to amyloid cross-interactions. Amyloids are insoluble protein aggregates characterized by exceptional stability due to the tight packing of monomers, resulting in characteristic pattern in X-ray diffraction experiments⁹. Despite significant structural similarities shared by all amyloids, their sequences are surprisingly diverse and have little homology¹⁰. On the other hand, sometimes very similar sequences can result in distinctive structures¹¹. Numerous, both experimental and computational studies, explored mechanisms of amyloid aggregation and their roles in neurodegenerative disorders, including the pivotal role of oligomers formed at early stages of the aggregation process¹². More recent studies have shown that in some cases presence of amyloid aggregates can affect the aggregation rate of other proteins¹³. Later, it was observed that interacting proteins can form heterogeneous fibers consisting of molecules of both interaction partners. Hypothetical structural mechanisms of the cross-seeding, depending on the nature of interactors, are proposed in¹⁴. Aggregation and co-aggregations, observed in the phenomenon of cross-talk, is affected by environmental or experimental conditions. In case of the aggregation enhanced by interaction with another amyloid at conditions hampering the

aggregation, the cross-seeding presumably helps to overcome an energy barrier required for fibrillation (e.g., as observed in BSA protein in the presence of HEWL¹⁵)

The cross-interactions were identified between numerous proteins, including those involved in type II diabetes and neurodegenerative diseases. For example, interactions between Alpha synuclein and human Islet Amyloid Polypeptide (hIAPP)¹⁶. This shed new light on potentially new aspects regarding the origin of comorbidity of these disorders¹⁷. A similar mechanism was found to enhance the virulence HIV virus by increasing its adhesion to host cells¹⁸. Despite the importance of this process, its mechanisms are still poorly understood, although it was shown that polymorphism of an amyloid structure may play a certain role in the aggregation processes^{19,20}. The lacking understanding of this process can be attributed to a limited number of experimental data. As a result, interactions of only a few well-described proteins, such as Amyloid-beta (Abeta), islet amyloid polypeptide, or Alpha-synuclein, have been very extensively studied and they contributed to a majority of data. This may introduce a bias in available data. Despite the difficulties, it was shown that proteins with similar sequences are more likely to interact, however, many counterexamples were also shown¹⁷. The studies highlight the importance of the structural compatibility of amyloid cores.

The main limitation regarding experimental studies of amyloid aggregation and their interactions is that they require expensive and time consuming methods. In practice, biochemical assays based on the binding of Congo Red²¹, Thioflavin T²², and infrared spectroscopy are frequently used. Especially the last method is widely applied, due to its simplicity and efficiency²³. Another approach involves direct observation of fibers using high resolution imaging techniques, such as electron microscopy²⁴ and atomic force microscopy²⁵. Finally, the advancements in NMR spectroscopy made it an important tool for studying aggregation at the molecular levels²⁶. Since different methods rely on different approaches, their results might differ in some cases. More importantly, all of them are expensive and time consuming. Experiments are hampered by difficulties in handling amyloids, including their low solubility, rapid aggregation, and need for their high-purity²⁷. Currently, the use of experimental methods for the identification of all amyloids in genome wide studies would be impossible. To address this problem several computational methods have been proposed based on different approaches (reviewed in²⁸ and²⁹), starting from structural modeling³⁰, statistical analysis of the sequence including FoldAmyloid³¹ and FishAmyloid³², physicochemical models like PASTA 2.0³³, machine learning techniques such as APPNN³⁴ and AmyloGram³⁵. Furthermore, there are methods combining both approaches such as PATH³⁶ and Cordax³⁷. Finally, some methods, like Aggrescan 3D³⁸, utilize information about protein structure. It was also shown that bioinformatics techniques are quite robust and capable of identifying even some misannotated data despite being trained on them³⁹. Unfortunately, none of these methods can predict amyloid cross-interactions.

Here, we present a new computational method PACT (Prediction of Amyloid Cross-interaction by Threading) designed for the identification of potentially interacting amyloid pairs. The method is based on the molecular threading applied to the potential complex of interacting amyloids and the assessment of the stability of obtained molecular models.

Results

The main assumption of the method is that interactions between amyloidogenic fragments that cross-interact, threaded into an amyloid fiber structure, would result in a heterogeneous aggregate that is more stable, thus energetically more favorable than a non-interacting pair. In PACT, we use the Modeller⁴⁰ software for threading a query sequence on the structure of amyloid fiber formed by Islet Amyloid Polypeptide (IAPP)⁴¹. To assess obtained models we proposed *ndope* score, which is a normalized version of DOPE (Discrete Optimized Protein Energy) statistical potential implemented in the Modeller software.

PACT correctly identifies amyloid-prone peptides

In the first step, we focused on the prediction of homoaggregation, which can be considered a special case of heteroaggregation. We compared *ndope* scores obtained for models of potential homoaggregates of amyloidogenic and non-amyloidogenic peptides, for which the sequences were obtained from the AmyLoad database⁴². The majority of models obtained for amyloidogenic peptides showed much lower scores (meaning more stable structures) in comparison with non-amyloids and their first quartiles of the scores were well separated (Fig.1). Differences between both groups were statistically significant. Based on the Mann-Whitney U test we were able to reject the hypothesis that the distributions of both populations were identical ($p = 2.48e - 8$). Considering energy difference, we built a threshold-based classifier. The classification threshold was chosen based on the Receiver Operating Characteristic (ROC) curve, as a point on the curve closest to the point (0,1), representing perfect classification (Fig. 1B). The chosen value of *ndope* score was -242. If used merely for distinguishing amyloids from non-amyloids, such a classifier was able to achieve an Area Under ROC Curve (AUC) of 0.73 and *Accuracy* of 0.77. Moreover, high values of *Sensitivity* (0.73) and *Specificity* (0.86) were obtained. Such results are comparable with state-of-the-art amyloid predictors on the same data set (Table S1).

We also tested if the method is capable of recognizing amyloid propensity in functional amyloids, which pose a major problem for most predictors due to their under-representation in databases of amyloids. We tested the performance of the method on imperfect repeats of CsgA protein from *Escherichia coli* and *Salmonella enterica*²³ (Fig. S1). Aggregation-prone

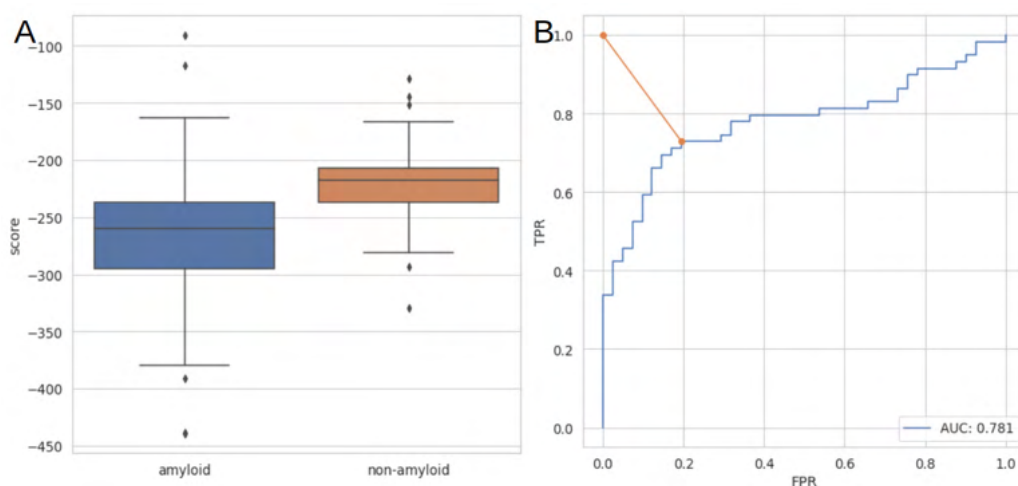


Figure 1. A) Distribution of *ndope* score for models of amyloidogenic and non-amyloidogenic peptides. B) ROC curve for amyloid vs non-amyloid classification. The orange line represents the distance between perfect classification point (0,1) and the chosen threshold

regions of this protein (R1, R3, and R5) scored much lower than non-amyloidogenic regions (R2 and R4) from *Escherichia coli*. On this data, PACT achieved an accuracy of 0.9. Furthermore, the observed difference between *ndope* score for R4 fragments from *Escherichia coli* and *Salmonella enterica* corresponds very well to the difference in their aggregation propensities observed in experimental works²³.

The results showed that the method can accurately predict aggregation-prone peptides of varying lengths. Furthermore, it can be utilized to detect functional amyloids.

PACT predicts amyloid cross-interactions

We used a similar methodology to predict cross-interactions of amyloid peptides, which is the main purpose of the method. The *ndope* scores of heteroaggregates consisting of pairs of peptides whose cross-interactions resulted in faster aggregation were compared with non-amyloidogenic pairs of peptides (Fig.2). A similar analysis was performed for pairs of peptides whose cross-interactions resulted in slower aggregation (Fig. S3). In both cases, models of heterologous aggregates resulting from cross-interactions showed lower values of *ndope* scores than non-amyloids, and well-separated first quartiles of their scores (Fig.2). Furthermore, in both cases, differences between groups were statistically significant (Mann–Whitney U test, $p = 5.54e - 16$ for *faster vs negative* and $p = 2.19e - 15$ *slower vs negative* cases). Therefore, we built the threshold-based classifier using the approach described in the previous section.

To assess the performance and choose the optimal threshold value, ROC curves were calculated for both cases; *faster rate vs negative* (Fig.3) and *slower vs negative* (Fig. S3) on both training and test sets. To minimize the impact of the data choice, we performed k-folds cross-validation with $k=5$ on the training set and calculated several metrics describing the performance of the method (Table 1). The same metrics were then calculated on an independent test set. The same analysis was performed for the case of prediction of interactions resulting in slower aggregation (Table S2). Chosen *ndope* thresholds were very similar in both scenarios, namely -256 and -245 for *faster vs negative* and *slower vs negative* respectively. PACT performed well on both cross-validation and independent test set. It achieved the *Accuracy* of 0.83 and 0.80 on test sets of *faster vs negative* and *slower vs negative* cases, respectively. In all cases, the results obtained on the test set were within the value of one standard deviation range from the mean values obtained with the cross-validation procedures. The method performance was quite similar in both *faster vs negative* and *slower vs negative* scenarios. However, due to the smaller data set size, a larger standard deviation was obtained for *slower vs negative* scenario (Table 1). The results show that the method can predict whether two peptides can cross-interact but cannot distinguish between types of interactions with regard to their rate.

PACT is robust to bias in data

A serious problem with the data regarding interacting amyloids, which is available in the literature and, consequently, our dataset, is the large overrepresentation of interactions concerning the Abeta peptide. This may cause overfitting of the method to this group of sequences. To assess its effect, we analyzed the scores obtained for interactions between different Abeta variants (Fig. S4). The observed scores for Abeta pairs fall within the range of values observed for the remaining pairs and, therefore,

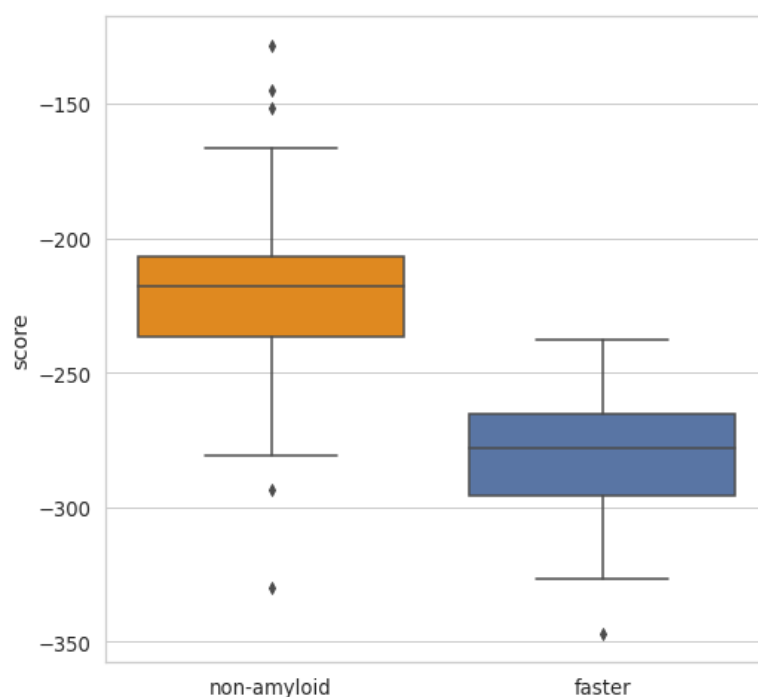


Figure 2. The score *ndope* for models of interacting identical non-amyloidogenic peptides (negative set) and interacting pairs resulting in increased aggregation rates (*faster* set).

Table 1. Performance of PACT on cross-validation and independent test set for classification of non-aggregating and cross-interacting pairs resulting in faster aggregation

| | Acc [std] | Sens [std] | Spec [std] | F1 [std] | MCC [std] |
|------------------|-------------|-------------|-------------|-------------|-------------|
| Cross-validation | 0.90 [0.05] | 0.91 [0.03] | 0.90 [0.08] | 0.90 [0.04] | 0.80 [0.06] |
| Test set | 0.83 | 0.78 | 0.88 | 0.82 | 0.66 |

they should not have a significant effect on the performance of the method. These pairs showed a relatively narrow distribution of the *ndope* values, centered slightly below the *ndope* value of -275 , which is relatively close to the identified classification threshold of -256 , while the remaining interacting pairs showed even lower scores.

Interactions between bacterial amyloids and Alpha-synuclein

In recent years, numerous studies have highlighted the connection between the gut microbiome composition and the onset of many diseases, including neurodegenerative ones such as Alzheimer’s and Parkinson’s diseases⁴³. Despite extensive research, understanding of the molecular mechanisms underlying this connection remains elusive. One possible explanation for this relates to functional amyloids from bacteria and human disease-related amyloids through the cross-interaction theory. The aggregation of bacterial amyloids could speed up the aggregation of disease-related proteins, leading to the disorder⁴⁴. This hypothesis seems consistent with the results obtained by Chen and co-workers, who discovered increased production and aggregation of Alpha-synuclein in rats exposed to bacterial strains producing biofilm-related functional amyloids⁴⁵.

In order to better understand this connection we studied possible interactions between bacterial functional amyloid CsgA and human Alpha-synuclein, whose aggregation is a hallmark of Parkinson’s disease. We used PACT to predict interactions of CsgA protein from five different organisms found in the human microbiome; *Escherichia coli* (EC), *Hafnia alvei* (HA), *Yokenella regensburgei* (YR), *Citrobacter youngae* (CY), and *Cedecea davisae* (CD) with human Alpha-synuclein which were recently studied experimentally by Bhoite and coworkers⁴⁶. It should be noted that CsgA protein was not included in the data set used to develop PACT since it exceeded the maximum length of the template.

The sequence of Alpha-synuclein was divided into overlapping fragments of length 20 amino acids and their interactions with R1-R5 repeats of each of CsgA proteins were tested. Consistently with experimental results, all of the studied CsgA

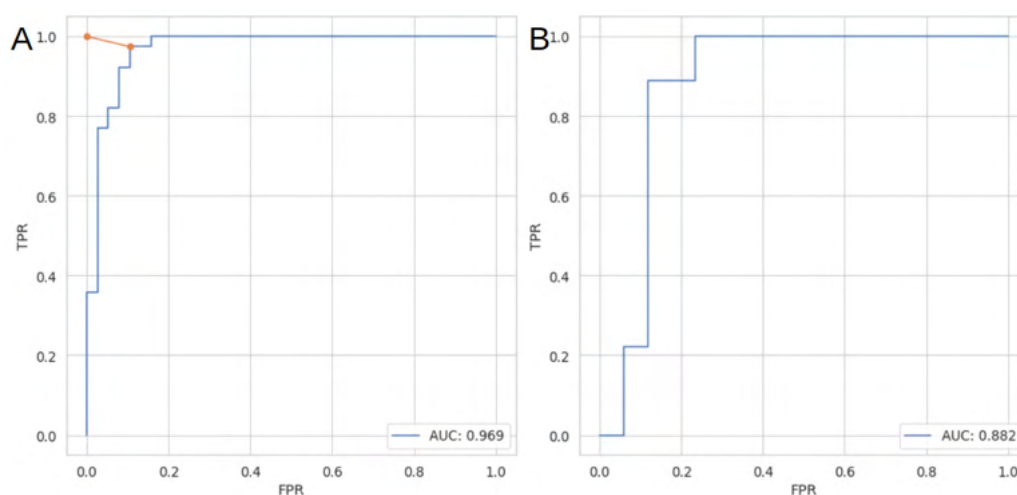


Figure 3. ROC curves for classification of non-aggregating and cross-interacting pairs resulting in faster aggregation on (A) training and (B) test set.

variants were predicted to interact with Alpha-synuclein. Among CsgA proteins' fragments, R1, R3 and R5 were predicted to interact, with R5 showing the best scores (Fig. 4). These results are consistent with our current state of knowledge about CsgA as the most aggregation-prone regions in these proteins are R1, R3 and R5. Furthermore, R5 fragment which showed the lowest *ndope* scores, is located at the protein surface, therefore it can interact without a need for major conformational changes. On the Alpha-synuclein part, the best scoring region was located between positions 32 and 56 (Fig. S5). This region was recently shown to be of crucial importance for the aggregation of the protein^{47,48}.

Mechanism of cross-interactions between CsgA and hIAPP

Finally, we studied interactions of CsgA protein from *Escherichia coli* with human Amylin (hIAPP) and complemented the results with experimental validation (see Supplementary materials). It was previously shown that CsgA could enhance the aggregation of hIAPP¹⁸. We aimed at more detailed characterization of this interaction by identifying which CsgA region is most likely to interact with hIAPP. To do so, interactions between each of CsgA repeats and hIAPP were first modeled. PACT classified positively interactions of hIAPP with R1 and R5, with the scores of -257.39 for R5 and -256.52 for R1. Notably, these fragments are likely to be exposed to the environment, which additionally makes them good candidates for potential interactions. To test the PACT predictions, experimental validation was performed using ThT assay (Methodology and results are presented in section 4 of the Supplementary materials). Obtained results showed stronger fluorescence in both cases and reduced durations of the lag phase and half-time of hIAPP aggregation in the presence of R5 (Fig. S8, Table S5). This could suggest a particular role of R5 fragment in seeding hIAPP, as predicted by PACT.

Code Availability

PACT was implemented as an open-source Python module, available at GitHub repository: <https://github.com/KubaWojciechowski/PACT>. For users' convenience, we prepared a docker container for the application, as well as the web server: <https://pact.e-science.pl/pact/>. For the prediction of cross-interaction we recommend the use of a default score threshold of -256 and for the prediction of homoaggregation -242. The classification result denoted as "1" indicates potential interactions. Apart from the classification, the software returns generated models of aggregates.

Discussion

We proposed the first computational method for predicting amyloid cross-interactions. It is based on a highly interpretable and well-established physicochemical model, which is not heavily dependent on training data. This feature is especially important since the available data contains a strong interest bias towards interactions of a few popular amyloids related to neurodegenerative diseases, for example Aβ. However, in case of our method we carefully studied the effect of this overrepresentation and showed that it does not affect its performance. Furthermore, good performance on functional amyloids, which are very underrepresented in the datasets, suggests that the method is robust and can be effectively used on a wide range of sequences. In total, PACT achieved a high accuracy of 0.83 and 0.80 on the independent test sets of interactions concerning increasing and decreasing aggregation rates. On both sets the method achieved high AUC values of 0.88 and 0.89, and F1

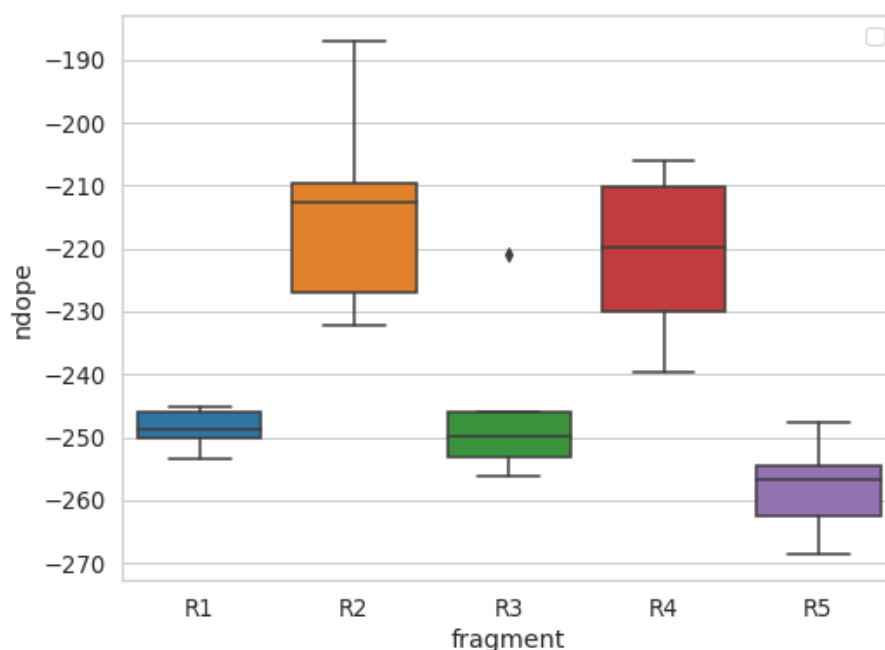


Figure 4. Lowest *ndope* scores for interactions of R1-R5 repeats from each of CsgA proteins with Alpha-synuclein.

values of 0.82 and 0.77, respectively. On the other hand, since both cases were characterized by similar interaction energies, the method cannot distinguish between enhancement and inhibition of aggregation. These results suggest that both processes may be driven by similar mechanisms. The issue was addressed in a recently published work by Louros and coworkers⁴⁹, who applied a somewhat similar approach to study the effect of point mutations on aggregation characteristics.

We used PACT to predict the interactions of bacterial functional amyloid CsgA from different species with human Alpha-synuclein and human amylin. Although these interactions were not included in the training data set, our results are in good agreement with recently published data regarding these pairs of proteins. Importantly, they also indicate which regions can drive the cross-interactions between both proteins. The identification of potentially interacting regions can provide important insights into the possible mechanism of the process and guide future experiments.

Apart from the identification of amyloid cross-interactions, the proposed method is also capable of reliably predicting amyloid-prone regions in proteins with comparable accuracy to state-of-the-art techniques. Furthermore, it overcomes their major limitations regarding the identification of functional amyloids. Unlike most of the currently available amyloid predictors, it does not rely on the scanning of a query sequence with a very short sliding window.

High-throughput identification of amyloid cross-interactions is an important step towards our understanding of its mechanisms. It can allow for a better understanding of the principles governing the process and can also be used to identify novel cases of amyloid interactions. Such capabilities can shed light on possible mechanisms responsible for the comorbidity of devastating disorders.

Methods

The main assumption of the method is that interactions between amyloidogenic fragments that cross-interact, threaded into an amyloid fiber structure, would result in a heterogeneous aggregate that is more stable, thus energetically more favorable than a non-interacting pair. A somewhat similar assumption was successfully applied in our previous work to predict the aggregation of short amyloidogenic fragments³⁶, although the current approach differs in other aspects of the method and the objectives. In PACT, we use the Modeller software for threading a query sequence on the structure of amyloid fiber formed by Islet the Amyloid Polypeptide (IAPP)⁴¹. To assess obtained models we have proposed *ndope* score, which is a normalized version of the DOPE statistical potential implemented in the Modeller software.

Data sets

To build and test the method we used the following datasets:

- the set of 86 amyloidogenic (*amyloid*) and 55 non-amyloidogenic (*non-amyloid*) peptides of lengths between 14 and 45 from the AmyLoad database⁴².
- the set of 119 pairs of peptides, which enhance (*faster* dataset) and 73 which slow down (*slower* dataset) the aggregation of each other. Both from AmyloGraph database⁵⁰. After the removal of identical records 57 and 55 pairs of peptides which enhance and slow down the aggregation of each other respectively.

The first two sets (*amyloid* and *non-amyloid*) were used to test the method on cases of homoaggregation i.e., identifying amyloid-prone peptides.

For the prediction of cross-interactions, we used *faster*, *slower* and non-amyloid sets. The use of the set of non-aggregating peptides as the negative set in the interaction study was caused by the lack of a sufficient number of negative examples of non-interacting amyloid pairs. This is a common problem in studies of protein-protein interactions since negative results are rarely published, which often creates a strong bias in biological data⁵¹. An analysis of this dataset reveals that it is mostly composed of peptides with strong beta propensity used by authors of the Tango method⁵². The proteins from this set could be mistaken for amyloid proteins by modeling methods, therefore they provide the best available negative dataset concerning amyloidogenicity. Importantly, due to the length restrictions, CsgA protein, which was used in our validation studies, was not included in the data sets used in the development of PACT.

Datasets used in this study are available at GitHub repository:

<https://github.com/KubaWojciechowski/PACT>

Modeling

A query pair of sequences were threaded on the structure of amyloid fiber formed by Islet Amyloid Polypeptide (IAPP)⁴¹. In order to allow the method to deal with sequences of varying lengths, sequences shorter than the sequence of a template use only the main part of the template structure. In such cases, a shorter sequence is aligned to the middle of the template sequence (Fig. 5A). This choice can be justified considering that most of the currently known amyloid fragments, which are longer than a few amino acids, share a similar beta-sheet turn architecture, commonly known as the beta arch. This assumption was previously successfully applied by Ahmed and coworkers to build the ArchCandy method for amyloidogenic region prediction⁵³. PACT allows sequences to be marginally longer than the template and, as a result, can be used to study cross-interactions between peptides of lengths between 14 and 45. For each of the tested pairs, 10 different models, consisting of two chains of each interacting peptide (Fig. 5B) were built using Modeller 9.24 model-multichain.py procedure with default parameters⁴⁰. Then, the model with the lowest DOPE value was chosen for further analysis. Since the dataset consisted of fragments of varying lengths, we proposed a normalized DOPE score (*ndope*) defined as follows:

$$ndope = \frac{DOPE}{L} \quad (1)$$

where L is an average length of sequences used to build a given model. Then, *ndope* scores were compared between amyloids and non-amyloids, as well as between pairs of amyloids interacting with non-amyloids.

To choose the *ndope* threshold for the classification, Receiver Operating Characteristics (ROC) curve was calculated by applying different score thresholds and recording False Positive Ratio (FPR) and True Positive Ratio (TPR). The threshold closest to the (0,1) point (representing perfect classification) was chosen. The whole procedure was schematically summarized in Fig. 6.

We also tested a variant of the method which utilized three different structural templates (PDB: 2nnt, 2e8d), however, it did not improve the accuracy of the method but significantly increased the computational time. Therefore, this approach was finally abandoned.

Assessment of performance and data analysis

All the data analysis was performed using Python 3.8 with Matplotlib⁵⁴, NumPy⁵⁵, Pandas⁵⁶, Scikit-Learn⁵⁷, and Seaborn⁵⁸ packages.

To test the performance of the proposed method, a dataset was randomly split into a training set and test set, which consisted of 30% of the data. Additionally, k-folds cross-validation (with k=5) was performed on the training data. Area Under ROC curve (AUC), *Accuracy* (ACC), *Sensitivity* (Sens), *Specificity* (Spec) and Matthew Correlation Coefficient (MCC) were used to assess the performance of the method

Effect of Amyloid-beta variants

For the analysis of the effect of over-represented amyloid beta pairs, we divided the *faster* data set into two subsets: one containing only pairs where both interacting peptides were variants of amyloid beta (16 pairs) (*abeta*), and the set of remaining pairs (39 pairs) (*no abeta*).

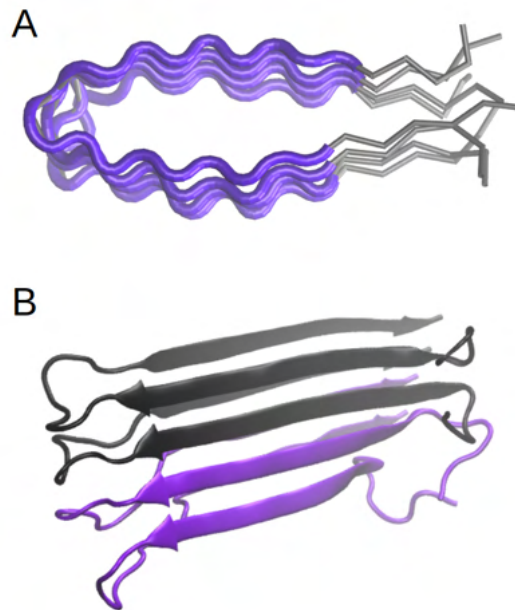


Figure 5. Schematic representation of the modeling procedure. A) In case when a query sequence is shorter than the template, only a part of it is used in modeling. B) The model of heterogenous fibril consists of two chains of each interacting peptide.

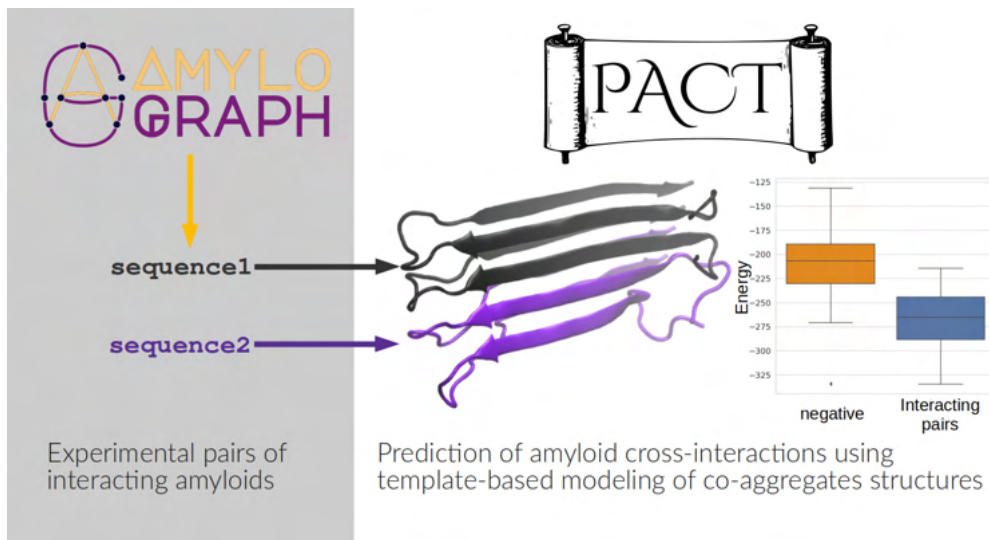


Figure 6. Schematic procedure of PACT

Interactions between bacterial amyloids and Alpha-synuclein

Modeling the interactions between Alpha-synuclein and CsgA proteins was performed using human Alpha-synuclein sequence (Uniprot id: P37840) and CsgA protein from five different organisms found in human microbiome; *Escherichia coli* (EC) (Uniprot id: P28307), *Hafnia alvei* (HA) (Uniprot id: G9Y7N6), *Yokenella regensburgei* (YR) (Uniprot id: A0A6H0K4L9), *Citrobacter youngae* (CY) (Uniprot id: A0A549VPM7), and *Cedecea davisae* (CD) (Uniprot id: S3IYN9). The sequence of Alpha-synuclein was divided into overlapping subsequences of lengths 20. This window length was chosen because it is similar to the length of repeated units in CsgA protein, responsible for its aggregation. CsgA variants were split into five non-overlapping fragments R1-R5 corresponding to five imperfect repeats observed in their sequences. Interactions of each of CsgA fragments with all Alpha-synuclein fragments were studied.

Data availability

Datasets used in this study are available at the GitHub repository:

<https://github.com/KubaWojciechowski/PACT>

References

1. Ross, C. A. & Poirier, M. A. Protein aggregation and neurodegenerative disease. *Nat. medicine* **10**, S10–S17 (2004).
2. Tang, Y., Zhang, D., Gong, X. & Zheng, J. A mechanistic survey of alzheimer's disease. *Biophys. Chem.* **281**, 106735 (2022).
3. Haque, E. *et al.* Protein aggregation: A new challenge in type-ii diabetes. *Adv Biotech & Micro* (2017).
4. Milardi, D. *et al.* Proteostasis of islet amyloid polypeptide: a molecular perspective of risk factors and protective strategies for type ii diabetes. *Chem. Rev.* **121**, 1845–1893 (2021).
5. Yang-Hartwich, Y., Bingham, J., Garofalo, F., Alvero, A. B. & Mor, G. Detection of p53 protein aggregation in cancer cell lines and tumor samples. In *Apoptosis and Cancer*, 75–86 (Springer, 2015).
6. Janson, J. *et al.* Increased risk of type 2 diabetes in alzheimer disease. *Diabetes* **53**, 474–481 (2004).
7. Sims-Robinson, C., Kim, B., Rosko, A. & Feldman, E. L. How does diabetes accelerate alzheimer disease pathology? *Nat. Rev. Neurol.* **6**, 551–559 (2010).
8. Rajput, A., Rozdilsky, B. & Rajput, A. Alzheimer's disease and idiopathic parkinson's disease coexistence. *J. geriatric psychiatry neurology* **6**, 170–176 (1993).
9. Eisenberg, D. S. & Sawaya, M. R. Structural studies of amyloid proteins at the molecular level. *Annu. Rev. Biochem.* **86**, 69–95 (2017).
10. Sipe, J. D. *et al.* Amyloid fibril protein nomenclature: 2010 recommendations from the nomenclature committee of the international society of amyloidosis. *Amyloid* **17**, 101–104 (2010).
11. Foo, C. K., Ohhashi, Y., Kelly, M. J., Tanaka, M. & Weissman, J. S. Radically different amyloid conformations dictate the seeding specificity of a chimeric sup35 prion. *J. molecular biology* **408**, 1–8 (2011).
12. Nguyen, P. H. *et al.* Amyloid oligomers: A joint experimental/computational perspective on alzheimer's disease, parkinson's disease, type ii diabetes, and amyotrophic lateral sclerosis. *Chem. reviews* **121**, 2545–2647 (2021).
13. Lundmark, K., Westermarck, G. T., Olsén, A. & Westermarck, P. Protein fibrils in nature can enhance amyloid protein a amyloidosis in mice: Cross-seeding as a disease mechanism. *Proc. Natl. Acad. Sci.* **102**, 6098–6102 (2005).
14. Ivanova, M. I., Lin, Y., Lee, Y.-H., Zheng, J. & Ramamoorthy, A. Biophysical processes underlying cross-seeding in amyloid aggregation and implications in amyloid pathology. *Biophys. Chem.* **269**, 106507 (2021).
15. Nirwal, S., Bharathi, V. & Patel, B. K. Amyloid-like aggregation of bovine serum albumin at physiological temperature induced by cross-seeding effect of hewl amyloid aggregates. *Biophys. Chem.* **278**, 106678 (2021).
16. Horvath, I., Rocha, S. & Wittung-Stafshede, P. In vitro analysis of α -synuclein amyloid formation and cross-reactivity. *Amyloid Proteins: Methods Protoc.* 73–83 (2018).
17. Ren, B. *et al.* Fundamentals of cross-seeding of amyloid proteins: an introduction. *J. Mater. Chem. B* **7**, 7267–7282 (2019).
18. Hartman, K. *et al.* Bacterial curli protein promotes the conversion of pap248-286 into the amyloid sevi: cross-seeding of dissimilar amyloid sequences. *PeerJ* **1**, e5 (2013).
19. Lucas, M. J. *et al.* Cross-seeding controls $a\beta$ fibril populations and resulting functions. *The J. Phys. Chem. B* **126**, 2217–2229 (2022).
20. Rahimi Araghi, L. & Dee, D. R. Cross-species and cross-polymorph seeding of lysozyme amyloid reveals a dominant polymorph. *Front. molecular biosciences* **7**, 206 (2020).
21. Howie, A. J. & Brewer, D. B. Optical properties of amyloid stained by congo red: history and mechanisms. *Micron* **40**, 285–301 (2009).
22. Nielsen, L., Frokjaer, S., Brange, J., Uversky, V. N. & Fink, A. L. Probing the mechanism of insulin fibril formation with insulin mutants. *Biochemistry* **40**, 8397–8409 (2001).
23. Szulc, N. *et al.* Variability of amyloid propensity in imperfect repeats of csga protein of salmonella enterica and escherichia coli. *Int. journal molecular sciences* **22**, 5127 (2021).

24. Shirahama, T. & Cohen, A. S. High-resolution electron microscopic analysis of the amyloid fibril. *The J. cell biology* **33**, 679–708 (1967).
25. Wang, Z. *et al.* Afm and stm study of β -amyloid aggregation on graphite. *Ultramicroscopy* **97**, 73–79 (2003).
26. Cawood, E. E., Karamanos, T. K., Wilson, A. J. & Radford, S. E. Visualizing and trapping transient oligomers in amyloid assembly pathways. *Biophys. chemistry* **268**, 106505 (2021).
27. Gąsior-Głogowska, M. E., Szulc, N. & Szeftczyk, M. Challenges in experimental methods. In *Computer Simulations of Aggregation of Proteins and Peptides*, 281–307 (Springer, 2022).
28. Kotulska, M. & Wojciechowski, J. W. Bioinformatics methods in predicting amyloid propensity of peptides and proteins. In *Computer Simulations of Aggregation of Proteins and Peptides*, 1–15 (Springer, 2022).
29. Navarro, S. & Ventura, S. Computational methods to predict protein aggregation. *Curr. Opin. Struct. Biol.* **73**, 102343 (2022).
30. Thompson, M. J. *et al.* The 3d profile method for identifying fibril-forming segments of proteins. *Proc. Natl. Acad. Sci.* **103**, 4074–4078 (2006).
31. Garbuzynskiy, S. O., Lobanov, M. Y. & Galzitskaya, O. V. Foldamyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* **26**, 326–332 (2010).
32. Gąsior, P. & Kotulska, M. Fish amyloid—a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of aminoacids. *BMC bioinformatics* **15**, 1–8 (2014).
33. Walsh, I., Seno, F., Tosatto, S. C. & Trovato, A. Pasta 2.0: an improved server for protein aggregation prediction. *Nucleic acids research* **42**, W301–W307 (2014).
34. Família, C., Dennison, S. R., Quintas, A. & Phoenix, D. A. Prediction of peptide and protein propensity for amyloid formation. *PLoS one* **10**, e0134679 (2015).
35. Burdukiewicz, M. *et al.* Amyloidogenic motifs revealed by n-gram analysis. *Sci. reports* **7**, 1–10 (2017).
36. Wojciechowski, J. W. & Kotulska, M. Path-prediction of amyloidogenicity by threading and machine learning. *Sci. Reports* **10**, 1–9 (2020).
37. Louros, N., Orlando, G., De Vleeschouwer, M., Rousseau, F. & Schymkowitz, J. Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities. *Nat. communications* **11**, 1–13 (2020).
38. Kuriata, A. *et al.* Aggrescan3d (a3d) 2.0: prediction and engineering of protein solubility. *Nucleic acids research* **47**, W300–W307 (2019).
39. Szulc, N. *et al.* Bioinformatics methods for identification of amyloidogenic peptides show robustness to misannotated training data. *Sci. reports* **11**, 1–11 (2021).
40. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. molecular biology* **234**, 779–815 (1993).
41. Luca, S., Yau, W.-M., Leapman, R. & Tycko, R. Peptide conformation and supramolecular organization in amylin fibrils: constraints from solid-state nmr. *Biochemistry* **46**, 13505–13522 (2007).
42. Wozniak, P. P. & Kotulska, M. Amyload: website dedicated to amyloidogenic protein fragments. *Bioinformatics* **31**, 3395–3397 (2015).
43. Cani, P. D. Human gut microbiome: hopes, threats and promises. *Gut* **67**, 1716–1725 (2018).
44. Friedland, R. P. & Chapman, M. R. The role of microbial amyloid in neurodegeneration. *PLoS pathogens* **13**, e1006654 (2017).
45. Chen, S. G. *et al.* Exposure to the functional bacterial amyloid protein curli enhances alpha-synuclein aggregation in aged fischer 344 rats and caenorhabditis elegans. *Sci. reports* **6**, 1–10 (2016).
46. Bhoite, S. S., Han, Y., Ruotolo, B. T. & Chapman, M. R. Mechanistic insights into accelerated α -synuclein aggregation mediated by human microbiome-associated functional amyloids. *J. Biol. Chem.* 102088 (2022).
47. Ulamec, S. M. *et al.* Single residue modulators of amyloid formation in the n-terminal p1-region of α -synuclein. *Nat. communications* **13**, 1–16 (2022).
48. Gallardo, J., Escalona-Noguero, C. & Sot, B. Role of α -synuclein regions in nucleation and elongation of amyloid fiber assembly. *ACS chemical neuroscience* **11**, 872–879 (2020).

49. Louros, N. *et al.* Mapping the sequence specificity of heterotypic amyloid interactions enables the identification of aggregation modifiers. *Nat. communications* **13**, 1–20 (2022).
50. Burdukiewicz, M. *et al.* Amylograph: a comprehensive database of amyloid–amyloid interactions. *Nucleic Acids Res.* (2022).
51. Nowakowska, A. W. & Kotulska, M. Topological analysis as a tool for detection of abnormalities in protein-protein interaction data. *Bioinformatics* (2022).
52. Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. biotechnology* **22**, 1302–1306 (2004).
53. Ahmed, A. B., Znassi, N., Château, M.-T. & Kajava, A. V. A structure-based approach to predict predisposition to amyloidosis. *Alzheimer's & Dementia* **11**, 681–690 (2015).
54. Hunter, J. D. Matplotlib: A 2d graphics environment. *Comput. science & engineering* **9**, 90–95 (2007).
55. Harris, C. R. *et al.* Array programming with numpy. *Nature* **585**, 357–362 (2020).
56. McKinney, W. *et al.* Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, vol. 445, 51–56 (Austin, TX, 2010).
57. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. machine Learn. research* **12**, 2825–2830 (2011).
58. Waskom, M. L. Seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

Acknowledgements

We would like to thank the team of AmyloGraph developers for their effort in preparing the database used in this project and Alicja W. Nowakowska for valuable discussions of the manuscript.

This work was partially supported by the National Science Centre, Poland, Grant 2019/35/B/NZ2/03997.

Access to Wrocław Centre for Networking and Supercomputing is greatly acknowledged.

Author contributions statement

J.W.W. and M.K. developed the concept. J.W.W. and W.S. implemented the algorithms. N.S. and M.S. performed experimental validation. J.W.W. and M.K. analyzed data and wrote the manuscript.

Supplementary materials

PACT - Prediction of Amyloid Cross-interaction by Threading

Jakub W. Wojciechowski¹, Witold Szczurek¹, Natalia Szulc^{1,2}, Monika Szefczyk³, Małgorzata Kotulska¹

¹Department of Biomedical Engineering, Wrocław University of Science and Technology
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

²Department of Physics and Biophysics, Wrocław University of Environmental and Life Sciences, Norwida 25, 50-375 Wrocław, Poland

³Department of Bioorganic Chemistry, Faculty of Chemistry, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

1. Classification of amyloid vs non-amyloid peptides

To assess PACT performance for recognizing amyloids from non-amyloids, we calculated the same metrics for three other amyloidogenicity predictors: FoldAmyloid, AmyloGram and PACT (Table S1). PACT performance was similar to these methods, which shows that it can be used also for prediction of amyloid-prone peptides.

Table S1 Performance of PACT on the set of aggregating and non-aggregating peptides of lengths between 14 and 45 amino acids from AmyLoad daTablease.

| Method | Accuracy | Sensitivity | Specificity | F1 | MCC |
|-------------|----------|-------------|-------------|------|------|
| PACT | 0.77 | 0.73 | 0.85 | 0.81 | 0.55 |
| PATH | 0.67 | 0.56 | 0.85 | 0.68 | 0.41 |
| AmyloGram | 0.81 | 0.83 | 0.78 | 0.86 | 0.59 |
| FoldAmyloid | 0.75 | 0.73 | 0.78 | 0.80 | 0.49 |

2. Performance on functional amyloids

Our experience shows that most of amyloidogenicity predictors perform poorly on functional amyloids, which are underrepresented in available daTableases. To test if our method can be used on functional amyloids we tested it on R1-R5 imperfect repeats from CsgA protein from *E. coli* and *S. enterica*, which we studied previously (Szulc et al. 2021). Figure S2 shows calculated *ndope* scores for these fragments. Aggregation prone regions of R1, R3, and R5 scored much lower than non-amyloidogenic regions, R2 and R4, from *E. coli*. On this data, PACT achieved the Accuracy of 0.9. Furthermore, the observed a difference in

ndope score for R4 fragments from *E. coli* and *S. enterica*, which corresponds very well to the difference in their aggregation propensity.

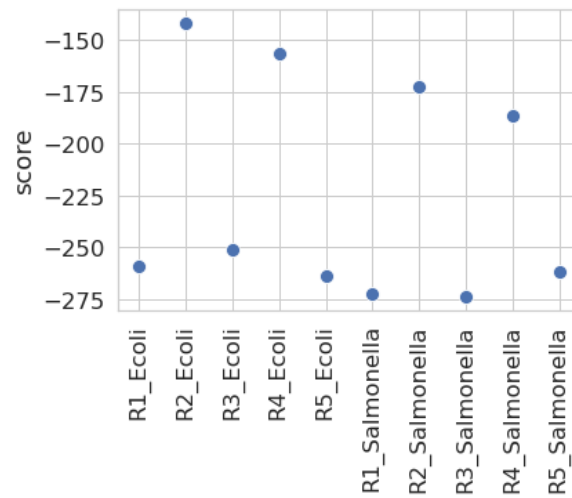


Fig. S1 *ndope* scores for R1-R5 imperfect repeats of CsgA protein from *E. coli* and *S. enterica*.

3. Prediction of cross-interactions

In the next step, we tested the performance of the method on pairs of interacting amyloids. We tested pairs whose interactions resulted both in increased and decreased aggregation speed. The first case is described in more detail in the main text. Here we show the results for the case of interactions resulting in slower aggregation (Fig. S2).

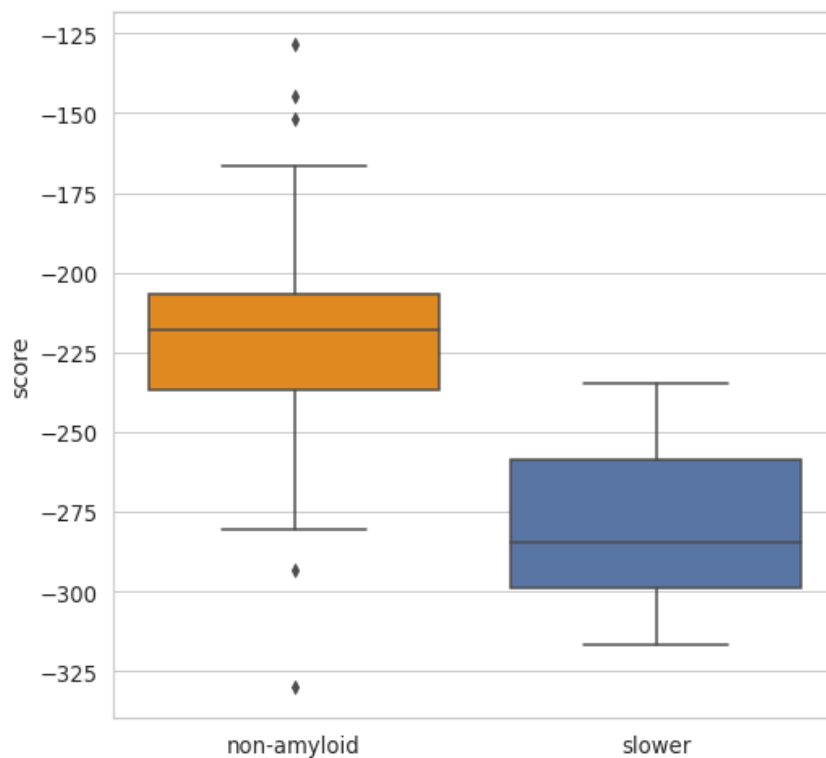


Fig. S2 Normalized DOPE score for models of non-amyloidogenic peptides and pairs of interactions resulting in decreased aggregation rates.

Same as in the case of prediction of homoaggregation, the threshold based classifier was built, but this time the data set was first split into training and test sets. On the training set k-folds cross-validation was performed with k=5. Then we used the whole training set to find the threshold again and tested the method on the independent test set. ROC curves were calculated for both training and test sets (Fig. S3), *ndope* threshold of -245 was found and other metrics were calculated (Table S2). The results similar to those obtained for faster aggregation were obtained.

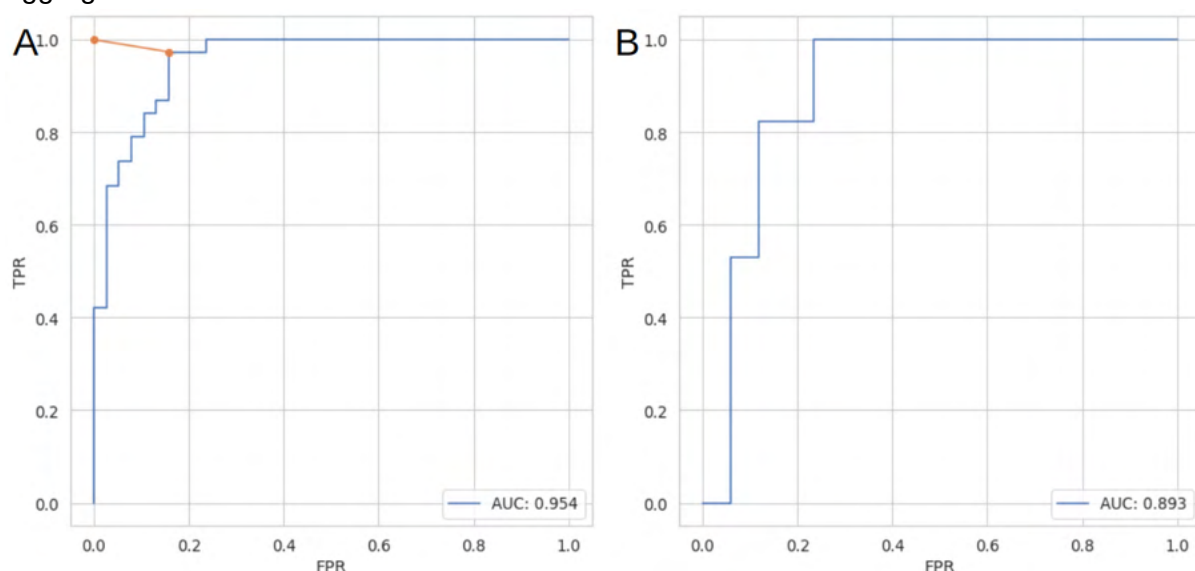


Fig. S3 ROC curves for classification of non-aggregating and cross interacting pairs resulting in slower aggregation on (A) training and (B) test set.

Table S2 Performance of PACT on cross-validation and independent test set for classification of non-aggregating and cross-interacting pairs resulting in their slower aggregation.

| | Accuracy [std] | Sensitivity [std] | Specificity [std] | F1 [std] | MCC [std] |
|------------------|----------------|-------------------|-------------------|-------------|-------------|
| Cross-validation | 0.90 [0.07] | 0.91 [0.17] | 0.86 [0.13] | 0.89 [0.09] | 0.81 [0.12] |
| Test set | 0.79 | 0.71 | 0.88 | 0.77 | 0.59 |

Next, we tested how interest bias of authors of the publications influences the performance of PACT. To do so, highly overrepresented interactions of different variants of Abeta were closely studied. The obtained scores for Abeta pairs were within the same range as values for the remaining pairs, and therefore should not have a significant effect on the performance of the method. These pairs obtained quite similar *ndope* values, centered slightly below *ndope* value of -275, which is relatively close to identified classification threshold for *faster vs negative* scenario of -256.

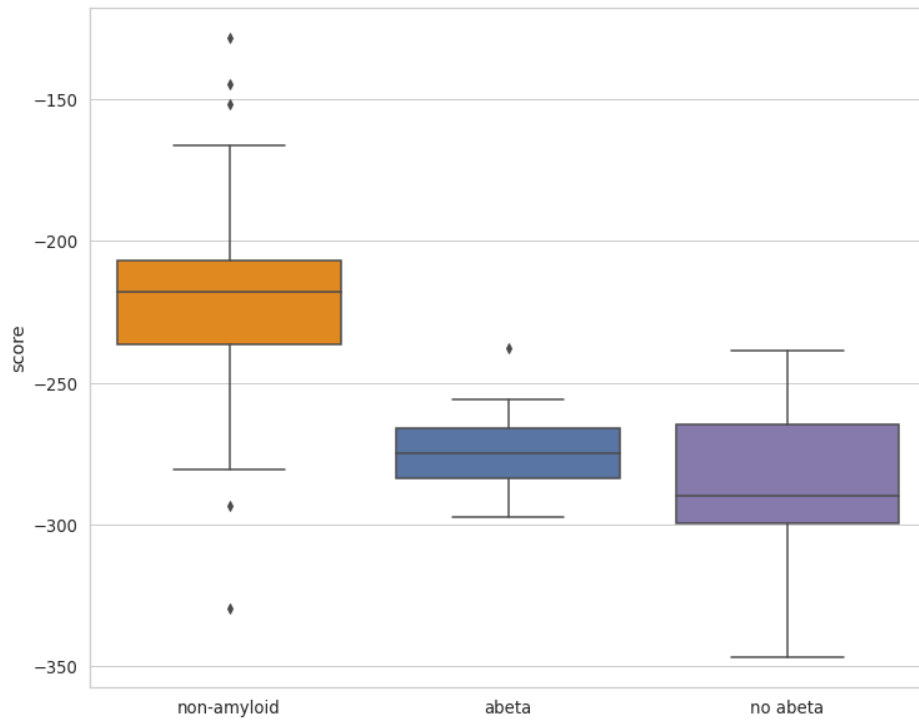


Fig. S4 Normalized DOPE score for models of non-amyloidogenic peptides and pairs of interactions resulting in increased aggregation rates where both partners belong to Abeta variants and the remaining pairs.

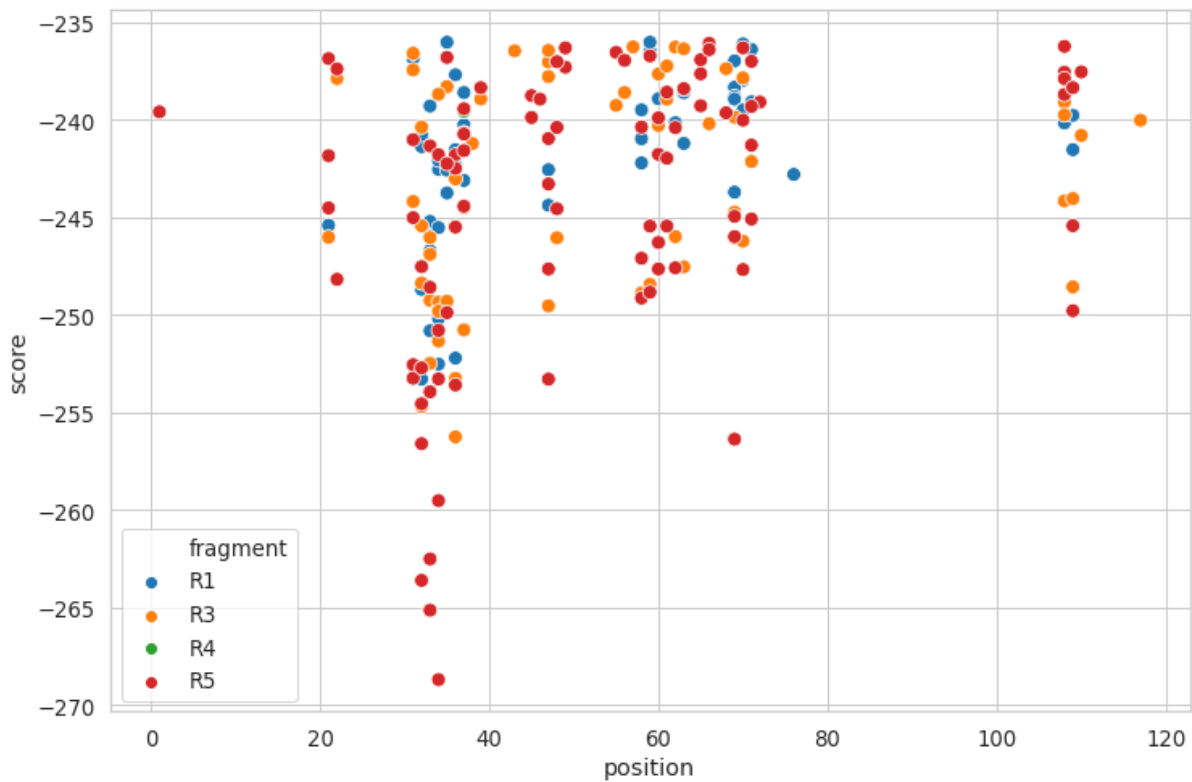


Fig. S5 Scores for interactions of CsgA fragments with Alpha-synuclein fragments. Each dot represents the starting position of 20 amino acid fragment of the sequence.

4. Experimental validation

Table S3 Peptides analytical data.

| Name | Sequence | Formula | Calculated M/z | Experimental M/z | Analytical HPLC t, [min] |
|-------|--|--|---|---|--------------------------|
| R1 | H-SELNIYQYGGNSALALQTDARN-NH ₂ | C ₉₄ H ₁₅₃ N ₂₉ O ₃₆ | [(M+2H)/2] 1228.1060 [(M+3H)/3] 819.0733 | [(M+2H)/2] 1228.1328 [(M+3H)/3] 819.0726 | 14.599 |
| R5 | H-SDLTITQHGGNGADVQGSDD-NH ₂ | C ₈₇ H ₁₃₈ N ₂₈ O ₃₆ | [(M+2H)/2] 1994.9397 [(M+3H)/3] 997.9738 | [(M+2H)/2] 1994.2070 [(M+3H)/3] 997.5491 | 12.273 |
| hIAPP | H-KCNTATCATQRLANFLVHSSNFGAILSSTNVGSNTY-NH ₂ | C ₁₆₅ H ₂₆₁ N ₅₁ O ₅₅ S ₂ | [(M+3H)/3] 1302.8 [(M+4H)/4] 977.3 | [(M+3H)/3] 1302.5 [(M+4H)/4] 977.3 | 10.230 |

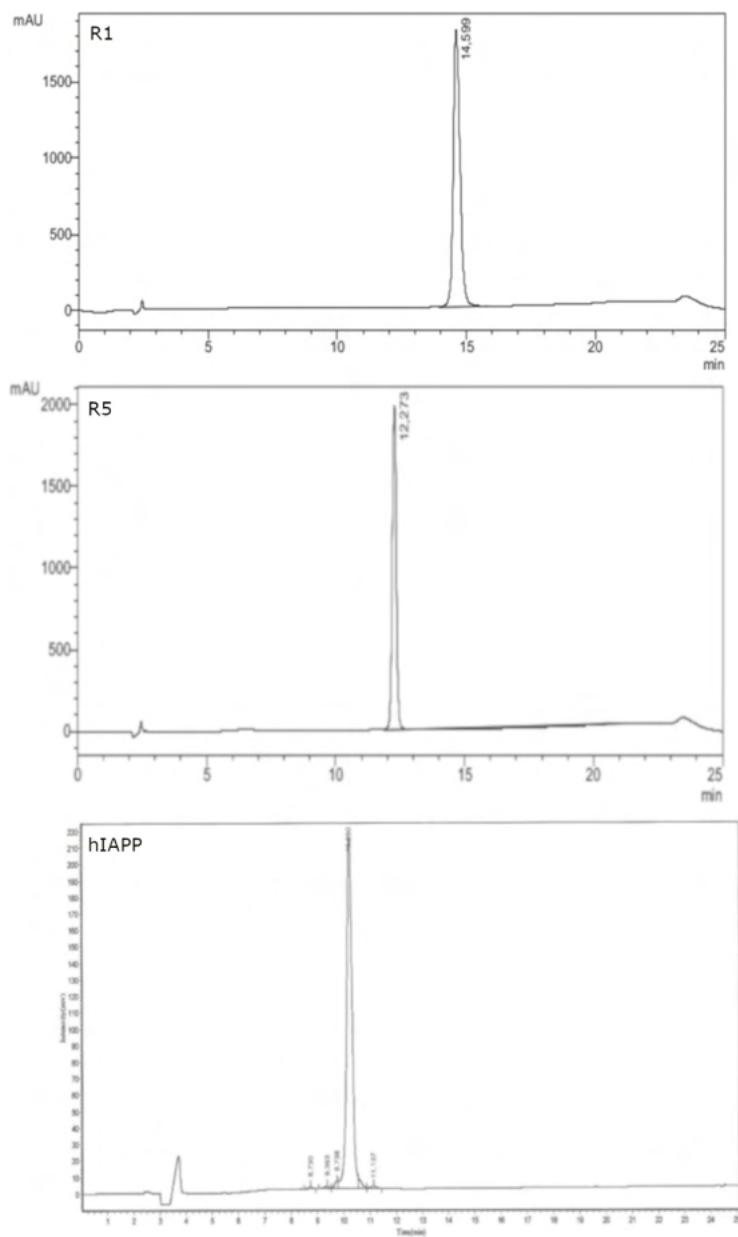


Fig. S6 Analytical HPLC chromatograms of the studied peptides.

Materials and Methods

Peptide synthesis. All commercially available reagents and solvents were purchased from Merck and used without further purification. Peptides R1 and R5 were synthesized with an automated solid-phase peptide synthesizer (Liberty Blue, CEM) using H-Rink amide ChemMatrix resin 35-100 mesh particle size (loading: 0.59 mmol/g). Fmoc deprotection was obtained using 20% piperidine in DMF for 1 min at 90 °C. A single-coupling procedure was achieved with 0.5 M solution of *N,N'*-diisopropylcarbodiimide (DIC) and 0.5 M solution of Oxyma Pure Novabiochem® in DMF for 4 min at 90 °C. Cleavage of the peptides from the resin was accomplished with the mixture of TFA/TIS/H₂O (95:2.5:2.5) after 3 h of shaking. The crude peptide was precipitated with ice-cold Et₂O and centrifuged (7 000 rpm, 10 min, 4 °C). Peptides were purified using preparative HPLC (Knauer AZURA ASM 2.1L) with a C18 column (Thermo Scientific, Hypersil Gold 12 μm, 250 mm × 20 mm) with water/acetonitrile (0.05% TFA) eluent system. hIAPP was purchased from ProteoGenix, see Table S3.

Analytical high-performance liquid chromatography (HPLC). Analytical HPLC for R1 and R5 was performed using column ReproSil Saphir C18 100Å 5μ 4.6 × 150 mm; detection wavelength 222 nm; eluent system: A = H₂O+0.05% TFA, B = CH₃CN+0.05% TFA, gradient: t=0–20 min, 90%–0% A; t=20–22 min, 0% A; t=22–25 min, 0%–90% A, see Fig. S6). Analytical HPLC for hIAPP was provided by ProteoGenix and performed on PLRP-S column 100Å 4.6 × 250 mm; detection wavelength 220 nm; eluent system: A = CH₃CN+0.1 % TFA, B = H₂O+0.1% TFA, gradient: t=0–25 min, 10%–90% A; t=25–30 min, 100%–0% A).

Mass spectrometry (MS). Peptides R1 and R5 were studied by WATERS LCT Premier XE System consisting of a high resolution mass spectrometer with a time of flight (TOF) using electrospray ionization (ESI). MS analysis for hIAPP was provided by ProteoGenix.

Circular dichroism (CD). CD spectra were recorded on JASCO J-1500 at 20 °C between 250 and 190 nm in water with following parameters: 0.2 nm resolution, 1.0 nm band width, 20 mdeg sensitivity, 0.25 s response, 50 nm/min scanning speed, 5 scans, 0.1 cm cuvette path length. The CD spectra of the 10 mM PBS buffer pH 7.4 was recorded and subtracted from the raw data. The peptides were dissolved in hexafluoroisopropanol (HFIP), then mixed for 3 hours to obtain monomers. HFIP was evaporated overnight in a desiccator, then the samples were dissolved in a PBS buffer to obtain peptide concentration of 100 μM. Then, a filtration process was conducted, and the resulting filtrate was employed for subsequent experimentation. The CD intensity is given as mean residue ellipticity (θ [deg × cm² × dmol⁻¹]). The spectra were smoothed using the Savitzky–Golay filter (polynomial order 2, widow size 19) applied in the SciPy package.

Thioflavin T (ThT) fluorescence assay. Kinetic measurements were carried out in a 96-well BRANDplate® on a CLARIOstar Plus, BMG LABTECH, at 20 °C, using wavelengths of 440±15 nm and 480±20 nm, for ThT excitation and emission respectively. Additionally, the plate was shaken for 30 s at the interval of 30 min during 24 hours of measurements. Final concentrations were 50 μM of ThT and 100 μM of each monomerized peptide. Peptides were monomerized according to the procedure described in the CD section. The experiment

was performed in the duplicate. The obtained fluorescence values were normalized to the fluorescence maximum in the 0–1 range.

Results

An experimental validation was conducted to confirm the predictions of cross-interactions obtained by PACT. Peptides for the studies were chosen based on the predicted energies and availability and included fragments R1 and R5 of the CsgA protein from *E. coli* species, known for their functional amyloid properties¹, as well as hIAPP, an amyloid associated type 2 diabetes². The interactions between the N-terminal (R1) and C-terminal (R5) fragments of CsgA protein and hIAPP were investigated using circular dichroism (CD) and Thioflavin T (ThT) fluorescence assays. The studies were undertaken to demonstrate cross-interaction of R1 and R5 fragments with hIAPP, as predicted computationally with PACT.

CD spectra of all samples exhibited a minimum at approximately 200 nm upon dissolution (see Fig. S7 and Table S4) indicating a random coil formation. Throughout the time course of the experiment, only a slight shift towards lower wavenumbers (approximately 1-2 nm) was observed for all samples. Despite that the secondary structure characteristic observed still resembled those of a random coil, the reduced spectral intensity and higher voltage on the photomultiplier indicate occurrence of the temporal aggregation.

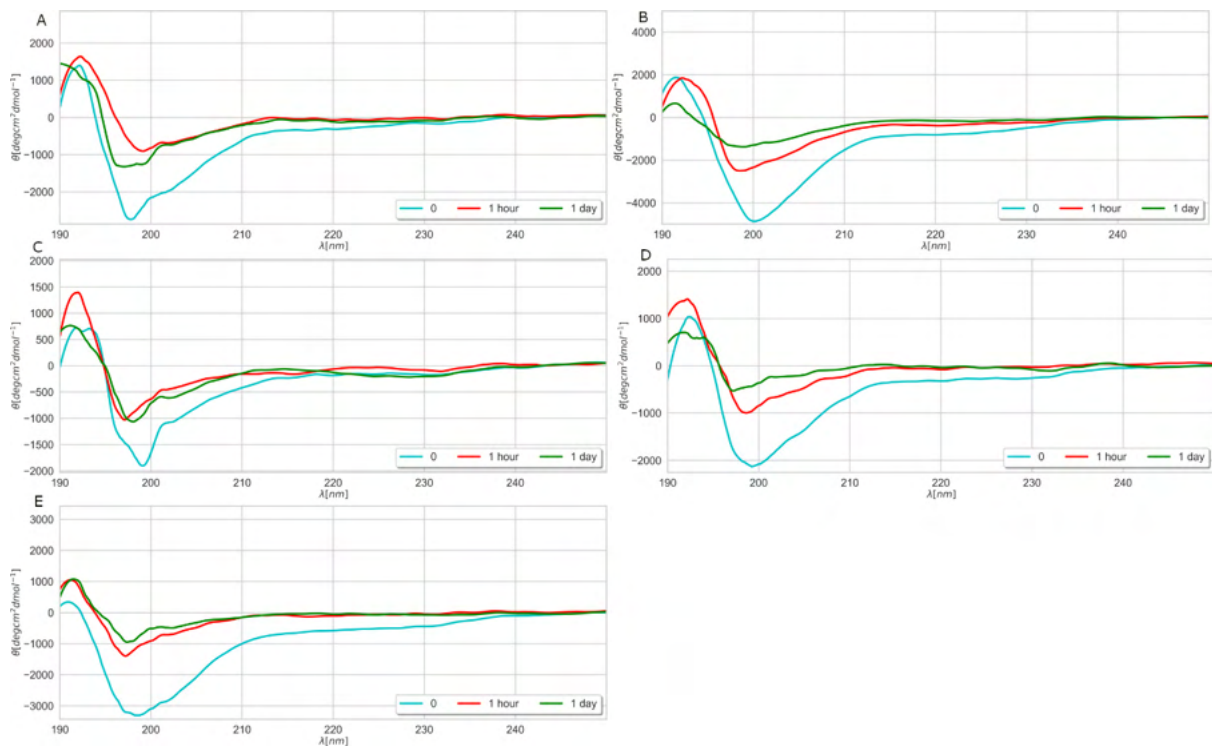


Fig. S7 Far-UV CD spectra of the studied samples: (A) fragment R1 of CsgA protein of *E. coli* species, (B) fragment R1 of CsgA protein of *E. coli* species + hIAPP, (C) fragment R5 of CsgA protein of *E. coli* species, (D) fragment R5 of CsgA protein of *E. coli* species + hIAPP, (E) hIAPP, on the day of dissolving, after one hour and after one day in the PBS buffer. C_{pep}=100 μM.

Table S4 Changes in the positions of CD spectra minima (given in nm) of the studied samples within time in the PBS buffer. C_{pep}=100 μM.

| sample time | R1 | R5 | hiAPP | R1+hiAPP | R5+hiAPP |
|-------------|-------|-------|-------|----------|----------|
| 0 | 197.8 | 199.2 | 198.6 | 200 | 199.2 |
| 1 hour | 199.2 | 197.2 | 197.2 | 198.6 | 198.6 |
| 1 day | 197.0 | 198.0 | 197.4 | 198.2 | 197.2 |

Comparative analysis revealed an acceleration of aggregation for the R1+hiAPP and R5+hiAPP samples compared to the individual peptides (see Fig. S8 and Table S5), confirming cross-interactions between the peptides. This acceleration was accompanied by a reduction in the lag phase duration from hours, observed in the case of terminal fragments of the CsgA protein, to minutes upon addition of hiAPP to the mixture. Additionally, the heterogeneous mixture of R5+hiAPP showed a faster rate of aggregation compared to R1+hiAPP (see Table S5). The R5+hiAPP mixture exhibited a stronger cross-interaction effect than R1+hiAPP. The PACT score indicated a lower energy (-257.39) for the R5-hiAPP interaction compared to R1+hiAPP (-256.52), which is consistent with the ThT results. The difference between these two heterogeneous mixtures was minimal based on the half-time, lag time, and the increase in normalized fluorescence value.

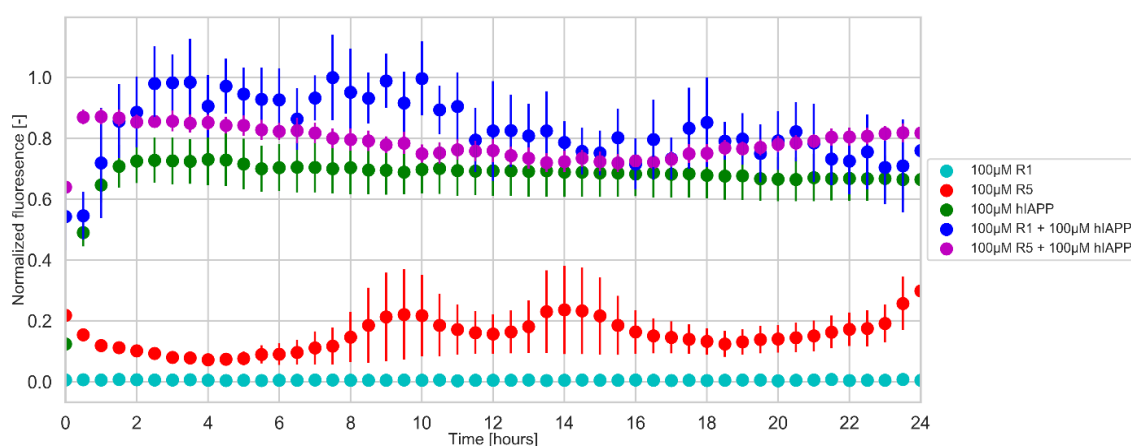


Fig. S8 ThT curves for the studied samples following the aggregation process in the PBS buffer. C_{pep}=100 μM.

Table S5 Parameters obtained by fitting aggregation kinetics to studied peptides. C_{pep}=100 μM. Where, the symbol '-' denotes that no half-time calculation was performed due to the absence of a fitted curve (only the lag phase was observed, representing the monomeric state of the peptide).

| sample parameter | R1 | R5 | hiAPP | R1+hiAPP | R5+hiAPP |
|------------------|----------|----------|------------|------------|------------|
| lag phase | 24 hours | 23 hours | 30 minutes | 30 minutes | 10 minutes |
| half time | - | 24 hours | 50 minutes | 1 hour | 30 minutes |

Bibliography

1. Wang, X., Hammer, N. D. & Chapman, M. R. The Molecular Basis of Functional Bacterial Amyloid Polymerization and Nucleation. *J Biol Chem* **283**, 21530 (2008).
2. Hull, R. L., Westermark, G. T., Westermark, P. & Kahn, S. E. Islet Amyloid: A Critical Entity in the Pathogenesis of Type 2 Diabetes. *J Clin Endocrinol MeTable* **89**, 3629–3643 (2004).