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 szerokich 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 agregacje inych 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 agregacje 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 agregacje (*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.

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. Sprawdzono również możliwość zbudowania modelu uczenia maszynowego wykorzystującego obliczone parametry do predykcji regionów amyloidowych. Metoda ta została następnie istotnie rozwinięta. Sprawdzono 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 natrafiliśmy 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 przesuwnym, 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 przyjrzeliśmy 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 motywach 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 rozsianych 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/

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.

26.06.2023

Jakub Wojciech