Algorytmy i modele do analizy struktur białkowych

Aleksandra Irena Jarmolińska
Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski
Czerwiec 2019

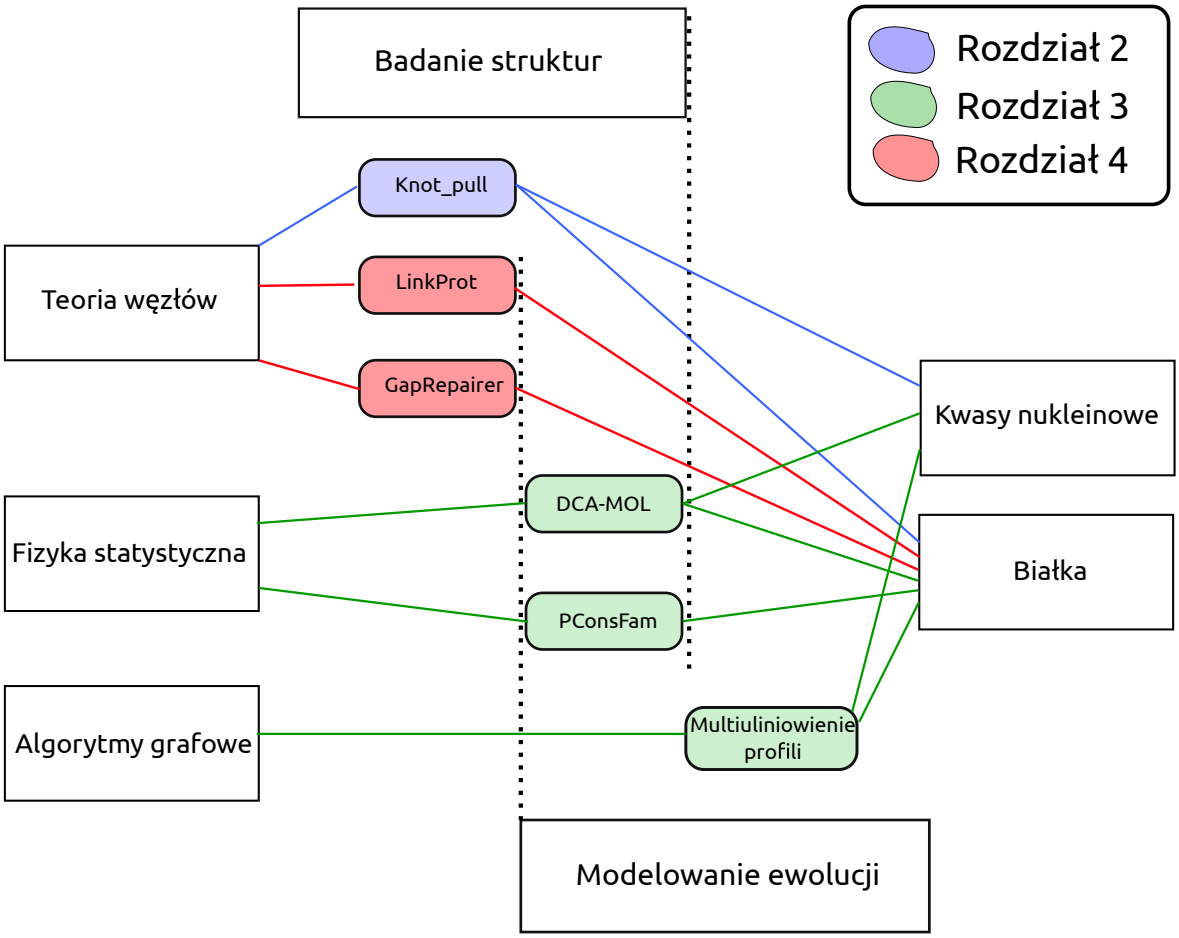
Białka często są określane mianem „cegiełek życia", ale jest to niedopowiedzeniem. Całe funkcjonowanie organizmów żywych zależy od właściwego działania białek. Nadal pozostało jednak wiele tajemnic z nimi związanych – w tym to, w jaki sposób sekwencja zakodowana w genomie wymusza zwinięcie białka do określonej struktury i dlaczego ten kształt jest niezbędny do jego prawidłowego funkcjonowania. Jednym ze szczególnie ciekawych aspektów zwijania białek jest obecność, znalezionych w ok. 2% znanych struktur białkowych (według baz danych KnotProt (Jamroz *et al.*, 2014) i LinkProt (Dabrowski-Tumanski *et al.*, 2016)) nietrywialnych topologii łańcucha głównego białka. Owa nietrywialność odnosi się do węzłów i splotów w rozumieniu teorii węzłów, jednak z pewnymi dodatkowymi założeniami wynikającymi z otwartego charakteru łańcuchów białkowych, które omawiamy dalej w pracy.

Niniejsza rozprawa opisuje różne algorytmy i metody przydatne w badaniu białek na różnych poziomach organizacji (Rys. 1) - począwszy od algorytmu wykrywającego rodzaj nietrywialnej struktury (przedstawionej w Rozdziale 2), poprzez algorytmy do uliniawiania wielu sekwencji i metody wiążące wariancję sekwencji białka z jego strukturą (Rozdział 3), aż po narzędzia i bazy danych węzłów i splotów w wielu łańcuchach (przedstawione w Rozdziale 4).

Wykrywanie węzłów w cząsteczkach biologicznych

Znalezienie struktur węzłopodobnych w molekułach biologicznych jest nietrywialnym zadaniem - większość biopolimerów to otwarte łańcuchy, a nie zamknięte krzywe (zgodne z oczekiwaniami matematycznej definicji węzła). Tak więc, na ogół, gdy pojawia się wyrażenie „węzeł białkowy", mamy na myśli węzeł „zdroworozsądkowy", taki jaki można by zrobić przy użyciu sznurka – struktura pociągnięta za oba końce nie stanie się linią prostą. Wartym uwagi wyjątkiem są tutaj cząsteczki DNA, które mogą być koliste (np. plazmidy) i które mogą być w rzeczywistości dość łatwo zawęźlane i rozwęźlane przez topoizomerazy (enzymy, których jedyną funkcją jest umożliwienie „przenikania się" łańcucha DNA). Te cechy sprawiają, że węzły i sploty na kolistych cząsteczkach DNA są znane i badane już od czterech dekad (Macgregor and Vlad, 1972; Sumners, 1995).

Znacznie bardziej złożonym tematem są węzły na otwartych biopolimerach – takich jak białka, RNA i chromatyna (otwarte łańcuchy DNA). Znalezienie ich wymaga mniej rygorystycznego podejścia do definicji matematycznych, ponieważ takie cząsteczki muszą najpierw zostać domknięte, aby umożliwić użycie narzędzi stosowanych w teorii węzłów. Zazwyczaj



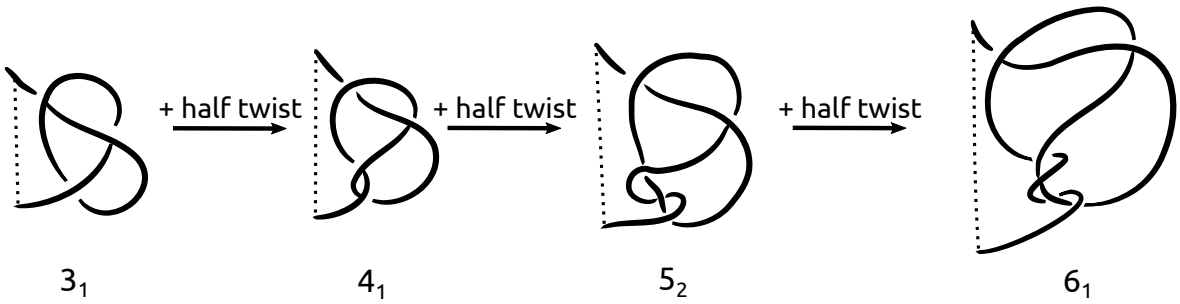
Rysunek 1: Diagram przedstawiający różne tematy poruszane w pracy.

takie podejście polega na wielokrotnym wydłużaniu końców struktury w losowych kierunkach, łączeniu ich na powierzchni dużej kuli otaczającej cząsteczkę i obliczaniu typu węzła dla każdego tak powstałego domknięcia łańcucha osobno. W efekcie otrzymuje się prawdopodobieństwo każdego znalezionego typu zapętlenia (Mansfield, 1994), przy często spotykanym założeniu, że prawdopodobieństwo powyżej 40% oznacza obecność węzła (Jamroz *et al.*, 2014).

Wśród polimerów biologicznych najbardziej interesujące – ze względu na najbardziej zróżnicowany charakter elementów łańcucha – wydają się być białka. Chociaż można łatwo wykazać, że zapadanie się długiego łańcucha zazwyczaj prowadzi do zawęźlonej struktury, (Levitt, 1976; Némethy and Scheraga, 1977; Skolnick and Kolinski, 1991; Chan and Dill, 1993), jako najbardziej korzystnego upakowania dla polimeru, to przez lata sądzono, iż proces fałdowania się białka wymyka się tej tendencji (Bryant *et al.*, 1974). Chociaż uważa się, że zwijanie białek jest kierowane głównie przez zapadnięcie hydrofobowe, różne interakcje (zarówno przyciągające, jak i odpychające) pomiędzy aminokwasami w łańcuchu komplikują ten proces – łańcuch białkowy nie jest „gładki". Z tego powodu węzły powinny powstawać blisko swojej natywnej pozycji w strukturze – dając splątaniom w białkom dodatkowy para-

metr: głębokość. Jest ona zdefiniowana jako minimalna liczba aminokwasów, które muszą zostać usunięte z któregokolowiek końca łańcucha, by rozplątać strukturę. Dopełnieniem tej miary jest zaciśnięcie węzła – liczba aminokwasów tworzących jego rdzeń, czyli minimalny fragment struktury na którym można wykryć węzeł.

Ze względu na fakt, że splątanie nie może przesuwać się wzdłuż łańcucha białkowego, przyjmuje się, że najtrudniejszym i najbardziej czasochłonnym etapem zwijania jest przeciskanie końca łańcucha przez pętlę. Potwierdza to fakt, że wszystkie dotychczas poznane węzły w białkach mogą powstać poprzez tylko jedno przejście przez (potencjalnie wielokrotnie skręconą) pętlę. (Sułkowska *et al.*, 2012; Taylor, 2007). Na przykład - chociaż w rodzinie deubikwitynaz można znaleźć strukturę podobną do węzła 5_2, nie ma znanych struktur podobnych do węzła 5_1 (patrz Rys. 2).



Rysunek 2: Węzły typu „twist", to węzły które mogą powstać przez tylko jedno przeciągnięcie końca łańcucha przez skręconą pętlę.

W Rozdziale 2 prezentujemy algorytm „knot_pull" – nowe narzędzie do analizy topologii w otwartych łańcuchach, takich jak białka, RNA i DNA (chromatyna). Powstało ono, by ominąć pewne ograniczenia narzucane przez obecnie stosowane metody.

Matematyczna definicja węzłów opisuje je jako „zamknięte krzywe"(jednowymiarowe okręgi) zanurzone w trójwymiarowej przestrzeni Euklidesowej. Węzły klasyfikowane są według złożoności, określanej liczbą przecięć (punktów podwójnych) w ich rzucie na płaszczyznę (diagramie).

Definicja 1 (Diagram splotu)**.** *Ortogonalny rzut węzła lub splotu na płaszczyznę, mający skończoną liczbę punktów wielokrotnych (punktów podwójnych, w miejcu poprzecznego przecięcia się linii), jest diagramem splotu $\mathcal{D}$ – nieskierowanym grafem płaskim, spełniającym poniższe kryteria:*

1. *pętle to spójne składowe grafu pozbawione wierzchołków (czyli rozłączne z resztą grafu);*
2. *każdy z końców krawędzi nienależącej do pętli prowadzi do jednego wierzchołka (może być ten sam), i każdy jest opisany jako idący górą lub idący dołem w danym wierzchołku;*
3. *do każdego wierzchołka prowadzą dwie krawędzie „górne" i dwie „dolne", poprowadzone naprzemiennie.*

Diagram węzła to diagram splotu zawierający tylko jedną spójną składową. Wierzchołki w diagramie splotu nazywane są przecięciami.

Obecnie, większość programów (Tubiana *et al.*, 2018; Lua, 2012; Jamroz *et al.*, 2014) do wykrywania węzłów w cząsteczkach biologicznych działa według tego samego schematu:

1. wygładzanie (upraszczanie) łańcucha, które prowadzi do uzyskania krzywej o tej samej topologii, ale mniejszej liczbnie przecięć w rzucie na płaszczyznę;

2. krzywa jest domykana na powierzchni dużej (*implicite* nieskończonej) kuli otaczającej strukturę – może to generować błędy, gdyż takie domykanie może wprowadzać dodatkowe przecięcia do diagramu. Dlatego powtarza się ten krok wielokrotnie, i liczy statystykę tak tworzonych topologii.

3. Domknięty łańcuch rzutowany jest na płaszczyznę, i liczony jest niezmiennik węzła (np. wielomian Alexandera (Alexander, 1928), albo wielomian HOMFLY-PT (opisany poniżej, Równanie (1)). Przy wielokrotnym domykaniu struktury, oblicza się statystykę niezmienników by określić prawdopodobieństwo, że struktura ma dany typ węzła.

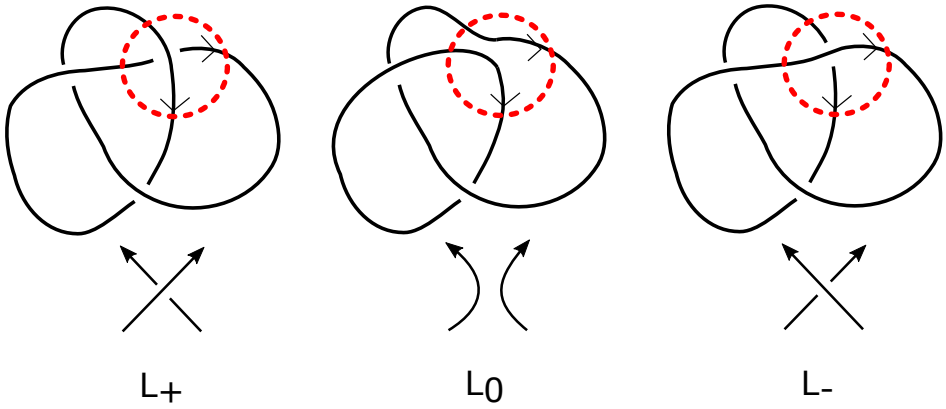
Węzły zazwyczaj rozpoznawane są przez ich niezmienniki. Niezmiennik, to dowolna właściwość, którą można określić dla każdego węzła, i jest stała dla równoważnych (izomorficznych) węzłów. Warto jednak zauważyć, że nie działa to w drugą stronę – odmienne węzły mogą mieć taki sam niezmiennik (np. w wypadku wielomianu Alexandera, istnieje węzeł o 8 przecięciach mający taki sam niezmiennik jak jeden z węzłów o 6 przecięciach). Najczęściej używanymi niezmiennikami są wielomiany, w szczególności wielomian HOMFLY-PT (Freyd *et al.*, 1990; Przytycki and Traczyk, 1988), stosowany także do splotów. Współczynniki wielomianu liczone są przez modyfikacje kierunków przecięć. Wielomian HOMFLY-PT rozszerza wielomiany Alexandera i Jonesa (Jones, 1985), i można go przekształcić do obu. Jest on określony poprzez zależności określane jako relacje *skein* (Rys. 3), które określają liniowe zależności między wielomianami dla splotów różniących się tylko jednym przecięciem (w wypadku prostszych wielomianów zależności te pozwalają na obliczenie ich współczynników wprost, poprzez rekursję).

Mając diagramy splotów L_-, L_+, L_0 przedstawione na Rys. 3, wielomian HOMFLY-PT definiujemy:

$$P_U(l,m) = 1 \tag{1}$$
$$lP_{L_+}(l,m) + l^{-1}P_{L_-}(l,m) + mP_{L_0}(l,m) = 0,$$

gdzie U to węzeł trywialny (okrąg), a l i m to współczynniki wielomianu służące do rozróżniania typów węzłów. Najważniejsze właściwości tego wielomianu to:

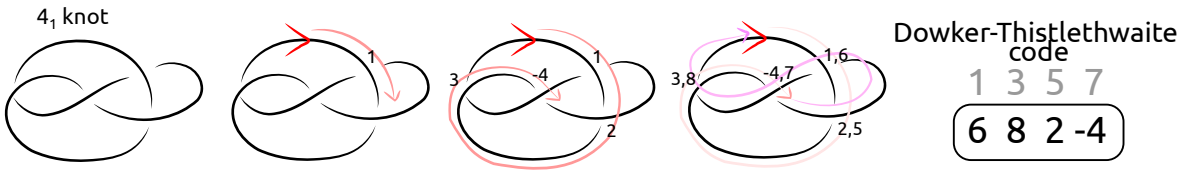
- wielomian HOMFLY-PT dla węzła złożonego to iloczyn wielomianów jego składników;



Rysunek 3: Relacje *skein* są określane dla trzech diagramów splotów różniących się jednym przecienciem. Każdy diagram powinien mieć inną konfigurację wspomnianego przecięcia, możliwe ułożenia pokazane są na ilustracji.

- wielomian HOMFLY-PT rozróżnia węzły rózniące się tylko chiralnością: $P_K(l,m) = P_{Odbicie\ lustrzane(K)}(l^{-1},m)$.

W pracy proponujemy aby do opisu typu węzła w cząsteczce używać kodu Dowkera-Thistlethwaite'a (DT) (Dowker and Thistlethwaite, 1983). Aby określić tę notację dla rzutu węzła na płaszczyźnie, zaczynając w dowolnie wybranym punkcie na krzywej, przemieszczamy się po niej numerując kolejno napotkane przecięcia. Dla poprawnego węzła, w chwili powrotu do punktu wyjścia, wszystkie przecięcia będą ponumerowane dwukrotnie, jedną parzystą i jedną nieparzystą liczbą (Rys. 4). Aby uwzględnić również chiralność struktury, odpowiednia wartość jest również oznaczana jako dodana przy przechodzeniu w przecięciu nad lub pod innym fragmentem krzywej (ujemna wartość parzysta oznacza, że krzywa przechodziła w danym przecięciu górą). Zapis jest dodatkowo skracany, poprzez posortowanie par liczb rosnąco według nieparzystych – porządek liczb parzystych wyznacza wtedy typ węzła. Jednakże należy pamiętać, że notacja DT nie jest niezmiennikiem – jeden diagram węzła moze mieć kilka różnych notacji.



Rysunek 4: Określanie kodu Dowkera-Thistlethwaite'a na diagramie węzła 4_1.

Kod DT zawiera o wiele więcej informacji na temat struktury którą opisuje, niż niezmiennik tego węzła, ponieważ zależy on od wyboru punktu startowego i kierunku numerowania. W cząsteczkach biologicznych obie te decyzje są narzucone strukturą – w białkach zaczynamy w N końcu i idziemy w stronę C końca. Jako, że liczba przecięć w diagramie węzła nawet uproszczonej struktury może być znacząca, w pracy proponujemy algorytm upraszczania kodu DT poprzez przekształcenia oparte na ruchach Reidemeistera (Reidemeister, 1927).

W pracy przedstawiamy nowy algorytm do wygładzania otwartych polimerów, który pozwala na łatwą wizualizację ich topologii, oraz łatwiejsze obliczenie kodu DT.

Modelowanie ewolucji sekwencji białek

Najlepszym sposobem by uzyskać jak najwięcej informacji o białku na podstawie jego sekwencji jest porównanie jej z innymi. Znaczenie samej kolejności aminokwasów nie jest jeszcze poznane wystarczająco by można było określić coś poza podstawowym rozróżnieniem na regiony hydrofobowe i hydrofilowe (chociaż już taka informacja pozwala przypuszczać czy dany fragment jest na powierzchni białka (Callaway, 1994)). Wszystkie inne cechy, takie jak przewidywana struktura drugorzędowa, albo podział na domeny, można określić tylko poprzez porównanie z już znanymi – na podstawie rozwiązanych struktur – statystykami i motywami.

Cała różnorodność fenotypów w toku ewolucji, zarówno w skali makro- jak i mikroskopowej, powstała w wyniku zaledwie kilku procesów molekularnych. Rearanżacje genomowe – takie jak duplikacje – są niezbędne dla powstania nowych białek. Dopóki przynajmniej jedna kopia zduplikowanego genu działa poprawnie, pozostałe mogą mutować swobodniej, na przykład zmienić lub tymczasowo (w skali ewolucji) stracić dotychczasową funkcję. Głównym motorem trwałych zmian genetycznych[1] są mutacje (w tym insercje i delecje, zwane łącznie indelami) pojedynczych nukleotydów. Zmiana jednej zasady azotowej w genie kodującym może spowodować kaskadę zmian w kolejnych etapach powstawania białka (wskazanych przez centralny dogmat biologii molekularnej). Jeżeli nie jest to cicha mutacja – czyli aminokwas kodowany przez zmienioną trójkę nukleotydową się zmieni – zmieni się też sekwencja białka. To z kolei może wprost doprowadzić do utraty funkcji, jeśli na przykład był to aminokwas potrzebny do wiązania ligandu, lub zmienić kształt ostatecznej struktury, co z kolei może zaowocować utratą funkcji, albo nawet uniemożliwić poprawne zwinięcie się białka.

Uliniowienie sekwencji

Pierwszym krokiem w porównywaniu dwóch sekwencji tej samej długości, w postaci napisów, jest określenie odległości edycyjnej, na przykład w jakim procencie dane sekwencje są identyczne. Dla sekwencji różnej długości policzenie takich statystyk wymaga najpierw uliniowienia ich względem siebie (tradycyjnie w sposób optymalizujący wybraną statystykę).

Uliniowienie to dopasowanie do siebie sekwencji – znalezienie odpowiadających sobie pozycji – w praktyce tworzone przez podpisanie pod sobą sekwencji, uzupełniając je w razie potrzeby symbolem przerwy (odpowiadającym indelom) w sposób optymalizujący funkcję oceny liczoną po kolejnych kolumnach. Uliniowienia tworzy się w oparciu o poniższe kryteria (Claverie and Notredame, 2006):

[1] Ograniczamy się tu do mutacji w genomie, które mogą zostać przekazane potomstwu.

- podobieństwo ewolucyjne, w którym uliniowione do siebie aminokwasy pochodzą od tego samego aminokwasu w sekwencji przodka;
- podobieństwo strukturalne, gdzie uliniawia się pozycje sekwencji, które znajdują się w zbliżonym miejscu w strukturze trójwymiarowej cząsteczki;
- podobieństwo funkcyjne, oparte na pełnieniu tej samej roli w białku.

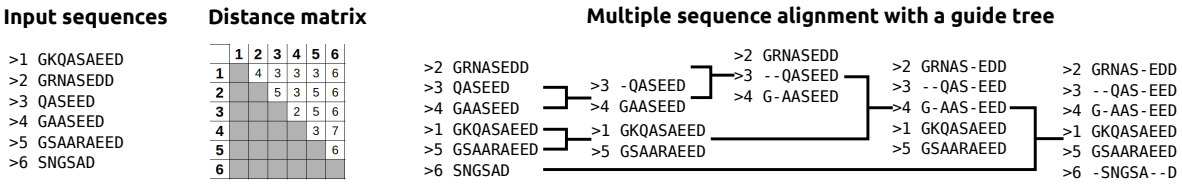
Dla spokrewnionych białek, te kryteria są niemal równoważne, ale żadnego nie można stwierdzić wyłącznie na podstawie sekwencji. Ponadto, trzy sekwencje mające taka samą liczbę zachowanych (identycznych) pozycji nadal mogą się różnić ogólnym stopniem podobieństwa. Założenie, że niektóre zmiany aminokwasów są łatwiej akceptowane (na przykład mutacja do aminokwasu o podobnych właściwościach prawdopodobnie nie zmieni znacząco działania cząsteczki) doprowadziły do stworzenia macierzy podobieństwa dla aminokwasów (takich jak BLOSUM (Henikoff and Henikoff, 1992) i PAM (Dayhoff *et al.*, 1978)), które oceniają jak korzystna jest dana mutacja (z zaznaczeniem, ze najkorzystniejszy jest brak zmiany, zwłaszcza dla aminokwasów o bardziej specyficznych właściwościach). Znalezienie optymalnego uliniowienia jest złożone obliczeniowo – dla dwóch sekwencji o długości M i N złożoność czasowa to $\mathcal{O}(N \times M)$, a pamięciowa to $\mathcal{O}(N \times M)$ (przy odpowiedniej implementacji da się je ograniczyć do opowiednio $\mathcal{O}(N \cdot \max(1, \frac{M}{\log(N)}))$ (Arlazarov *et al.*, 1970; Masek and Paterson, 1980) i $\mathcal{O}(\min(N, M))$ (Hirschberg, 1975)).

Niektóre algorytmy dopasowania dwóch sekwencji można uogólnić do większej ich liczby, dającej uliniowienie wielu sekwencji (ang. *multiple sequence alignment*, MSA). Jest to jednak możliwe tylko w przypadku zastosowania algorytmów programowania dynamicznego. Są one oparte na macierzach, w których kolejne wymiary mają rozmiar odpowiadający długości kolejnych sekwencji. Sprawia to, że globalna optymalizacja multiuliniowienia dla danych rzeczywistych jest zbyt czasochłonna (wymaga uzupełnienia, a następnie znalezienia optymalnej ścieżki w macierzy o rozmiarze L^N, gdzie N to liczba sekwencji, a L ich długość).

Na podstawie multiuliniowienia można stworzyć profil sekwencji, który opisuje różnorodność poszczególnych kolumn na przykład przy użyciu Ukrytych Modeli Markowa (ang. *Hidden Markov Model*, HMM), czyli automatów skończonych, zawierających stan *emisji* dla każdej kolumny uliniowienia. HMM dopasowany do danego multiuliniowienia (na przykład przy użyciu algorytmu Bauma-Welcha) przedstawia przewidywany wzór generujący znane sekwencje, a nie tylko ich zliczenia (jak dzieje się w przypadku macierzy wag – innej formy liczenia profili sekwencji). Sprawia to, że Ukryte Modele Markowa są o wiele skuteczniejsze w wykrywaniu dalekiej homologii (Madera and Gough, 2002).

Multiuliniowienia liczy się zazwyczaj przy użyciu metod heurystycznych, na przykład przy użyciu drzew podobieństwa opartych na podobieństwie wszystkich par sekwencji spośród badanych, w których w kolejnych węzłach drzewa łączy się pod-uliniowienia policzone w

węzłach poniżej (Sievers and Higgins, 2014) (Rys. 5).



Rysunek 5: Tworzenie multiuliniowienia sekwencji przy uzyciu drzewa podobieństwa

Inną metodą tworzenia multiuliniowień jest maksymalizacja zgodności z uliniowieniami par sekwencji każda z każdą (które można policzyć w czasie wielomianowym). Można to zrobić przez znalezienie śladu o maksymalnej wadze (ang. *maximum weight trace* (Kececioglu, 1993)) w grafie $\mathcal{G} = (V, E, \prec)$ przedstawiającym zbiór uliniowień.

Definicja 2. *Graf $\mathcal{G} = (V, E, \prec)$ jest grafem uliniowień dla zbioru $\mathcal{S}$ sekwencji, jeżeli jego wierzchołki V odpowiadają pozycjom w sekwencjach w S, z porządkiem w każdej sekwencji S_i wyznaczonym przez relację $\prec$ dla pozycji $s_i, s_j \in S_i$: $s_i \prec s_j \iff i + 1 = j$. Oznacza to, że relacja $\prec$ jest spełniona wtedy i tylko wtedy, gdy s_i jest w sekwencji bezpośrednio przed s_j. Krawędzie E to nieskierowane, ważone połaczenia między wierzchołkami (pozycjami), które zostały do siebie uliniowione.*

Ścieżka w grafie $\mathcal{G}$ to zbiór pozycji, które powinny znaleźć się w jednej kolumnie multiuliniowienia. Zatem rozdzielając graf na jego spójne składowe, wyznaczamy kolumny, z zastrzeżeniem, że tak stworzone multiuliniowienie jest poprawne tylko, jeżeli jego kolumny można ustawić w porządku liniowym wyznaczonym przez relację $\prec'$, która dla wspólnych składowych A i B:

$$A \prec' B \iff (\exists a \in A)(\exists b \in B) : x \prec y.$$

Ślad w grafie uliniowień $\mathcal{G}$ jest zatem zbiorem krawędzi $T \subseteq E$, dla którego spójne składowe są acykliczne względem relacji $\prec'$. W grafie $\mathcal{G}$ z krawędziami ważonymi funkcją w ślad o maksymalnej wadze znajduje się przez maksymalizację $\sum_{e \in T} w(e)$.

W Rozdziale 3 proponujemy dwa nowe algorytmy heurystyczne służące do znajdowania śladu o największej wadze (i wynikającego z niego multiuliniowienia), oba tworzące kolumny uliniowienia (spójne składowe grafu) przy użyciu zmodyfikowanego algorytmu Dijkstry do znajdowania drzew o najkrótszych ścieżkach. Pierwszy opiera się na zachłannej ekstrakcji kolumn jedna po drugiej (podejście „wgłąb"), drugi na oddolnym klastrowaniu wierzchołków do uzyskania minimalnej liczby poprawnych składowych (podejście „wszerz"). Tworzenie multiuliniowienia w ten sposób ma jedną istotną przewagę w stosunku do metod programowania dynamicznego – o wiele słabsze ograniczenia nałożone na dane wejściowe. W szczególności, sekwencjami, które uliniawiamy, mogą być profile sekwencyjne. Korzystając z tej własności, w dalszej części rozdziału przedstawiamy pierwszą analizę ewolucji białek ze slipknotami (takich, w których węzeł jest tylko na części łańcucha).

Multiuliniowienia sekwencyjne zawierają wyłącznie współczesne sekwencje (jako, że ciężko o historyczne dane molekularne), ale ich różnorodność pozwala na pewną intuicję odnośnie historii ewolucyjnej na przykład rodziny białek.

W szczególności aminokwasy w białku nie są zawieszone w próżni – oddziaływuja ze sobą, i zmiana jednego z partnerów w takiej interakcji może wpłynąć na ewolucję drugiego. Ten proces jest podłożem dla badania związków między sekwencją, a strukturą białek metodami koewolucyjnymi, takimi jak analiza sprzężeń bezpośrednich (ang. *Direct Coupling Analysis* (Weigt *et al.*, 2009; Morcos *et al.*, 2011), DCA).

Dla zadanego multiuliniowienia DCA oblicza gęsty model statystyczny prawdopodobieństw wystąpienia różnych typów aminokwasów, który pozwala na obliczenie współczynników bezpośredniej (bez udziału reszty sekwencji) korelacji dla wszystkich par pozycji (kolumn) w uliniowieniu.

Jest to struktura wnioskowania statystycznego oparta na modelu Pottsa, opisującym zachowanie q typów spinów na siatce. Dla uliniowienia o N kolumnach daje to model na siatce o wymiarach $N \times N$ z $q = 21$ spinami (odpowiadającymi znakom uliniowienia – 20 aminokwasów i symbol przerwy). W przypadku uliniowienia, każdy spin to tak naprawdę macierz współwystępowania aminokwasów we wskazanych kolumnach, o wymiarach $q \times q$. Model pozwala na obliczenie dla dowolnej sekwencji prawdpodobieństwa jej przynależności do sekwencji opisanych modelem, jak również parametrów określajacych siłę bezpośrednich zależności między kolumnami (pozycjami).

Druga część Rozdziału 3. prezentuje dokładniej zastosowania DCA w badaniu białek. Opisujemy tam DCA-MOL, narzędzie do łatwej analizy zależności koewolucyjnych na znaznych strukturach (Jarmolinska *et al.*, 2019b). Następnie przedstawiamy PConsFam, bazę danych struktur białkowych wymodelowanych na podstawie wyników z metody DCA (Lamb *et al.*, 2019). Na koniec pokazujemy jak zastosowanie kontaktów znalezionych przy użyciu DCA może ułatwić symulacje zwijania białek (Dabrowski-Tumanski *et al.*, 2015).

Bazy danych i narzędzia algorytmiczne do badań topologii w białkach

Rozdział 4 opisuje nasze pozostałe prace w dziedzinie komputerowych badań białek. Od niedawna wiemy, że nietrywialność topologiczna białek nie musi ograniczać się do jednego łańcucha (Dabrowski-Tumanski and Sulkowska, 2017) – znalezione zostały struktury zawierające sploty zbudowane z róznych łańcuchów. W tym rozdziale opisujemy internetową, samoaktualizującą się bazę danych, zbierajacą informacje na temat splotów w łańcuchach wszystkich opublikowanych struktur – LinkProt (Dabrowski-Tumanski *et al.*, 2016).

Pomimo ciągłych postępów technologicznych w technikach pozwalających na określanie struktur białek, nadal wielu z nich nie udało się określić w całości. W pewnych zastosowaniach jest to niewielka przeszkoda, ale niepełne – zawierające dziury – struktury nie mogą

zostać wykorzystane na przykład w symulacjach dynamiki molekularnej. Istnieją różne narzędzia pozwalające na wypełnienie takich dziur, ale często mają one dużo ograniczeń, lub też automatycznie odrzucają, jako niepoprawne, struktury o nietrywialnej topologii. By wyjść naprzeciw tej potrzebie stworzyliśmy serwer GapRepairer (Jarmolinska *et al.*, 2018), pozwalający na naprawianie niepełnych struktur z uwzględnieniem topologii.

Wreszcie, przy użyciu gruboziarnistych symulacji dynamiki molekularnej proponujemy możliwe ścieżki zwijania dla kilku niedawno poznanych struktur białek z węzłami (Jarmolinska *et al.*, 2019a).

Publikacje opisane w Rozdziale 2

Jarmolinska, A. I., Gambin, A., Sulkowska, J. I. (2019). Knot_pull - python package for biopolymer smoothing and knot detection. *Bioinformatics (under review)*

Publikacje opisane w Rozdziale 3

Jarmolinska, A. I., Zhou, Q., Sulkowska, J. I. and Morcos, F. (2019b). Dca-mol: A pymol plugin to analyze direct evolutionary couplings. *Journal of Chemical Information and Modeling*, **59** (2), 625-629.

Lamb, J.*, **Jarmolinska, A. I.***, Michel, M.*, Menéndez-Hurtado, D., Sulkowska, J. I. and Elofsson, A. (2019). Pconsfam: An interactive database of structure predictions of pfam families. *Journal of Molecular Biology*, **431** (13), 2442-2448.

Dabrowski-Tumanski, P., **Jarmolinska, A.I.** and Sulkowska, J. I. (2015). Prediction of the optimal set of contacts to fold the smallest knotted protein. *Journal of Physics: Condensed Matter*, **27** (35), 354109.

Publikacje opisane w Rozdziale 4

Jarmolinska, A. I., Kadlof, M., Dabrowski-Tumanski, P. and Sulkowska,J. I. (2018). GapRepairer: a server to model a structural gap and validate it using topological analysis. *Bioinformatics*, **34** (19), 3300-3307.

Jarmolinska, A. I., Perlinska, A. P., Runkel, R., Trefz, B., Ginn, H. M., Virnau, P.and Sulkowska, J. I.(2019). Proteins' knotty problems. *Journal of Molecular Biology*, **431** (2), 244-257.

Dabrowski-Tumanski, P.*, **Jarmolinska, A. I.***, Niemyska, W.*, Rawdon, E. J., Millett, K. C. and Sulkowska, J. I. (2016). Linkprot: A database collecting information about biological links. *Nucleic Acids Research*, **45** (D1), D243–D249.

Inne publikacje

Sulkowska, J. I., Niewieczerzal, S., **Jarmolinska, A. I.**, Siebert, J. T., Virnau, P. and Niemyska, W. (2018). Knotgenome: a server to analyze entanglements of chromosomes. *Nucleic Acids Research*, **46** (W1), W17-W24.

Bibliografia

Alexander, J. W. (1928). Topological invariants of knots and links. *Transactions of the American Mathematical Society*, **30** (2), 275–306.

Arlazarov, V. L., Dinitz, Y. A., Kronrod, M. and Faradzhev, I. (1970). On economical construction of the transitive closure of an oriented graph. In *Doklady Akademii Nauk*, Russian Academy of Sciences, vol. 194, pp. 487–488.

Bryant, T., Watson, H. and Wendell, P. (1974). Structure of yeast phosphoglycerate kinase. *Nature*, **247** (5435), 14.

Callaway, D. J. (1994). Solvent-induced organization: A physical model of folding myoglobin. *Proteins: Structure, Function, and Bioinformatics*, **20** (2), 124–138.

Chan, H. S. and Dill, K. A. (1993). The protein folding problem. *Physics Today*, **46** (2), 24–32.

Claverie, J.-M. and Notredame, C. (2006). *Bioinformatics for dummies*. John Wiley & Sons.

Dabrowski-Tumanski, P., Jarmolinska, A. and Sulkowska, J. (2015). Prediction of the optimal set of contacts to fold the smallest knotted protein. *Journal of Physics: Condensed Matter*, **27** (35), 354109.

—, Jarmolinska, A. I., Niemyska, W., Rawdon, E. J., Millett, K. C. and Sulkowska, J. I. (2016). Linkprot: A database collecting information about biological links. *Nucleic Acids Research*, **45** (D1), D243–D249.

— and Sulkowska, J. I. (2017). Topological knots and links in proteins. *Proceedings of the National Academy of Sciences*, **114** (13), 3415–3420.

Dayhoff, M., Schwartz, R. and Orcutt, B. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, National Biomedical Research Foundation Silver Spring, pp. 345–352.

Dowker, C. H. and Thistlethwaite, M. B. (1983). Classification of knot projections. *Topology and its Applications*, **16** (1), 19–31.

Freyd, P., Yetter, D., Hoste, J., Lickorish, W. R., Millett, K. and Ocneanu, A. (1990). A new polynomial invariant of knots and links. In *New Developments In The Theory Of Knots*, World Scientific, pp. 12–19.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, **89** (22), 10915–10919.

Hirschberg, D. S. (1975). A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, **18** (6), 341–343.

Jamroz, M., Niemyska, W., Rawdon, E. J., Stasiak, A., Millett, K. C., Sułkowski, P. and Sulkowska, J. I. (2014). Knotprot: a database of proteins with knots and slipknots. *Nucleic Acids Research*, **43** (D1), D306–D314.

Jarmolinska, A. I., Kadlof, M., Dabrowski-Tumanski, P. and Sulkowska, J. I. (2018). Gaprepairer: a server to model a structural gap and validate it using topological analysis. *Bioinformatics*, **34** (19), 3300–3307.

—, Perlinska, A. P., Runkel, R., Trefz, B., Ginn, H. M., Virnau, P. and Sulkowska, J. I. (2019a). Proteins' knotty problems. *Journal of Molecular Biology*, **431** (2), 244–257.

—, Zhou, Q., Sulkowska, J. I. and Morcos, F. (2019b). Dca-mol: A pymol plugin to analyze direct evolutionary couplings. *Journal of Chemical Information and Modeling*, **59** (2), 625–629.

Jones, A. (1985). A polynomial invariant for knots via von neumann algebras. *Bulletin of the American Mathematical Society*, **12** (1), 103.

Kececioglu, J. (1993). The maximum weight trace problem in multiple sequence alignment. In *Combinatorial Pattern Matching*, Springer Berlin Heidelberg, pp. 106–119.

Lamb, J., Jarmolinska, A. I., Michel, M., Menéndez-Hurtado, D., Sulkowska, J. I. and Elofsson, A. (2019). Pconsfam: An interactive database of structure predictions of pfam families. *Journal of Molecular Biology*, **431** (13), 2442–2448.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology*, **104** (1), 59–107.

Lua, R. C. (2012). Pyknot: a pymol tool for the discovery and analysis of knots in proteins. *Bioinformatics*, **28** (15), 2069–2071.

Macgregor, H. and Vlad, M. (1972). Interlocking and knotting of ring nucleoli in amphibian oocytes. *Chromosoma*, **39** (2), 205–214.

Madera, M. and Gough, J. (2002). A comparison of profile hidden markov model procedures for remote homology detection. *Nucleic Acids Research*, **30** (19), 4321–4328.

Mansfield, M. L. (1994). Are there knots in proteins? *Nature Structural Biology*, **1** (4), 213.

Masek, W. J. and Paterson, M. S. (1980). A faster algorithm computing string edit distances. *Journal of Computer and System Sciences*, **20** (1), 18–31.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T. and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, **108** (49), E1293–E1301.

Némethy, G. and Scheraga, H. A. (1977). Protein folding. *Quarterly Reviews of Biophysics*, **10** (3), 239–352.

Przytycki, J. H. and Traczyk, P. (1988). Invariants of links of conway type. *Kobe Journal of Mathematics*, **4**, 115–139.

Reidemeister, K. (1927). Elementare begründung der knotentheorie. In *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, Springer, vol. 5, pp. 24–32.

Sievers, F. and Higgins, D. G. (2014). Clustal omega, accurate alignment of very large numbers of sequences. In *Multiple sequence alignment methods*, Springer, pp. 105–116.

Skolnick, J. and Kolinski, A. (1991). Dynamic monte carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *Journal of Molecular Biology*, **221** (2), 499–531.

Sułkowska, J. I., Rawdon, E. J., Millett, K. C., Onuchic, J. N. and Stasiak, A. (2012). Conservation of complex knotting and slipknotting patterns in proteins. *Proceedings of the National Academy of Sciences*, **109** (26), E1715–E1723.

Sumners, D. W. (1995). Lifting the curtain: using topology to probe the hidden action of enzymes. *Notices of the American Mathematical Society*, **42** (5), 528–537.

Taylor, W. R. (2007). Protein knots and fold complexity: some new twists. *Computational Biology and Chemistry*, **31** (3), 151–162.

Tubiana, L., Polles, G., Orlandini, E. and Micheletti, C. (2018). Kymoknot: A web server and software package to identify and locate knots in trajectories of linear or circular polymers. *The European Physical Journal E*, **41** (6), 72.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. and Hwa, T. (2009). Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, **106** (1), 67–72.