Analiza podobieństwa struktur przestrzennych białek przy użyciu deskryptorów lokalnej struktury

Autoreferat rozprawy doktorskiej

Paweł Daniluk

1 Wstęp

Białka jako niezwykle zróżnicowana klasa biopolimerów pełnią fundamentalną rolę we wszystkich znanych organizmach żywych, biorąc udział w praktycznie każdym procesie życiowym komórki. Szczególnie istotną gałęzią badań szeroko rozumianej biochemii i biofizyki jest poznanie funkcji pełnionych przez konkretne białka i mechanizmów ich realizacji. To z kolei ma praktyczne znaczenie dla rozumienia między innymi procesów chorobotwórczych i przeciwdziałania im.

Biologiczna funkcja danego białka zależy w dużej mierze od jego struktury przestrzennej. Co więcej, można spodziewać się, że białka o podobnej strukturze mogą być spokrewnione ewolucyjnie i pełnić zbliżone funkcje. Ponieważ eksperymentalne określenie struktury białka jest niejednokrotnie łatwiejsze od rozpoznania funkcji, jaką pełni ono w żywym organizmie, odnajdowanie podobnej struktury o znanej funkcji może być użyteczną metodą przewidywania funkcji nowoodkrytego białka. Ponadto szczegółowe określenie podobieństwa struktur ze wskazaniem odpowiadających sobie regionów może pomóc w zidentyfikowaniu tych miejsc, które odpowiadają za realizowanie rozważanej funkcji.

Struktura białka natomiast zależy od jego sekwencji. Jednakże białka o znacząco różnych sekwencjach mogą posiadać bardzo zbliżone struktury przestrzenne. Zjawisko to jest przejawem procesów molekularnej ewolucji, które w procesie selekcji prowadzą do struktur o określonych, funkcjonalnych właściwościach i które muszą być dostatecznie stabilne z punktu widzenia fizyki. W szczególności możliwe jest, że wskutek ewolucji konwergentnej dwa białka niezależnie osiągną tę samą strukturę[4]. Możliwy jest również przypadek bardziej złożony, gdy struktury różnią się kolejnością występowania elementów w odpowiadających im sekwencjach i jednocześnie zachowują tę samą "architekturę" i funkcję[18].

Porównywanie oraz możliwość klasyfikacji struktur przestrzennych białek z wykorzystaniem dobrze zdefiniowanych, zalgorytmizowanych procedur, ma zatem istotne znaczenie z punku widzenia pełniejszego rozumienia mechanizmów funkcjonowania białek, mechanizmów ewolucji molekularnej oraz związków między procesami ewolucji molekularnej

a prawami fizyki, które wyznaczają warunki konieczne dla istnienia obserwowanych w przyrodzie struktur.

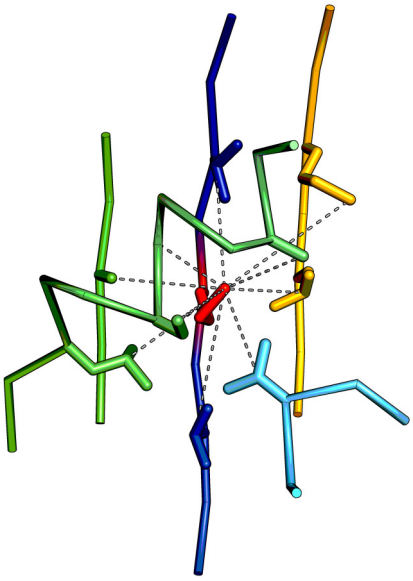
Niejednokrotnie zdarza się, że dwa białka mają podobny kształt, mimo że różnią się kolejnością występowania podobnych regionów w sekwencji[14]. Klasyczne rozumienie uliniowienia nie obejmuje takiego przypadku, a standardowe metody porównywania sekwencji nie pozwalają na wykrycie podobieństwa. Nierzadko jest ono jednak widoczne już przy wizualnej inspekcji rozważanych struktur. Najczęściej występują tzw. permutacje cyrkularne, które mogą powstawać m. in. wskutek duplikacji genów lub zmian zachodzących w łańcuchu białka podczas zwijania[18][1]. Występują również bardziej skomplikowane przestawienia fragmentów w sekwencji, które powodowane są między innymi przez zmianę długości pętli łączących elementy struktury drugorzędowej, co z kolei wymusza ich przestawienia na skutek ograniczeń stereochemicznych[10].

Struktury białek nie powinny być traktowane jak obiekty sztywne. Wiele funkcji, które pełnią, realizowanych jest poprzez celowe zmiany konformacji przestrzennej[9, 7]. Również eksperymentalne procedury pomiaru struktury mogą dawać rozbieżne wyniki spowodowane naturą eksperymentu. Dość istotne jest zatem uwzględnienie potencjalnych odkształceń podczas oceniania podobieństwa.

Można wyróżnić dwie podstawowe metodologie porównywania struktur białek – globalną i lokalną. Pierwsza polega na iteracyjnym ulepszaniu uliniowienia i superpozycji struktur. Wychodząc od pewnego uliniowienia, oblicza się optymalne nałożenie odpowiadających sobie aminokwasów, a następnie w tak uzyskanej superpozycji identyfikuje się pary bliskich przestrzennie aminokwasów, traktując je jako uliniowienie w następnym kroku iteracji. Metody tego typu sprawdzają się, jeżeli w strukturach porównywanych białek nie występują odkształcenia i podobieństwo jest wystarczająco duże, aby proces był zbieżny do globalnego optimum.

Alternatywą dla podejścia globalnego są metody oparte na identyfikowaniu podobieństw lokalnych, z których w kolejnych fazach obliczenia budowane jest globalne uliniowienie. Jest wiele możliwości dekompozycji struktury i co za tym idzie, sposobów obliczania lokalnych podobieństw. Do najważniejszych należy badanie odległości pomiędzy aminokwasami, podobieństwa pojedynczych wycinków łańcucha głównego lub elementów struktury drugorzędowej. Globalne uliniowienie jest następnie obliczane poprzez wybór możliwie licznego podzbioru lokalnych podobieństw, które są ze sobą zgodne. Definicja pojęcia zgodności oraz sposób przeszukiwania przestrzeni rozwiązań zależą od metody. Ze względu na złożoność obliczeniową związaną z kombinatorycznym rozmiarem przeszukiwanej przestrzeni zazwyczaj rezygnuje się z rozważania permutacji cyrkularnych i przestawień segmentów, nawet jeżeli metodologia pozwalałaby na ich znajdowanie. Taka sytuacja ma miejsce w przypadku metody DALI i jej publicznie dostępnej implementacji DaliLite[13, 12]. Niekiedy rozważa się wprowadzanie "zawiasów", aby umożliwić porównywanie struktur, pomiędzy którymi występuje odkształcenie. Pełniejsze przedstawienie aktualnego stanu wiedzy w tej dziedzinie można znaleźć w pracach[15, 11].

[1] Powiemy, że jedno białko jest cyrkularną permutacją drugiego, jeżeli istnieje podział obydwu struktur na dwie podjednostki (odpowiednio $A_1{-}B_1$ i $A_2{-}B_2$) takie, że struktury $A_1{-}B_1$ i $B_2{-}A_2$ są podobne w sensie klasycznego uliniowienia (bez przestawień).



Rysunek 1: Przykładowy deskryptor zbudowany wokół aminokwasu 70 domeny białkowej `d1lg7a_`. Deskryptor `d1lg7a_#70` obejmuje 9 kontaktów (linie przerywane) pomiędzy aminokwasem centralnym (kolor czerwony), a aminokwasami będącymi środkami elementów. Niektóre pięcioaminokwasowe elementy nakładają się tworząc dłuższe segmenty (dwa odcinki harmonijki β i helisa α).

Pojęcie multi-uliniowienia wielu struktur jest uogólnieniem uliniowienia. Można je definiować na dwa sposoby: jako znajdowanie podstruktury występującej we wszystkich porównywanych białkach bądź znajdowanie wszystkich podobieństw z zastrzeżeniem, że zidentyfikowane odpowiedniości pomiędzy aminokwasami muszą być jednoznaczne. Istniejące metody znajdowania multi-uliniowień strukturalnych często są rozszerzeniem algorytmów obliczających uliniowienia par. Na podstawie podobieństwa wszystkich par porównywanych struktur budowane jest wtedy drzewo binarne, którego liściom przypisane są rozważane struktury. Następnie węzłom drzewa przypisuje się uliniowienia struktur bądź multi-uliniowień występujących w potomkach, które są obliczane w sposób analogiczny do uliniawiania par struktur. Stosuje się również strategię analogiczną do klasteryzacji hierarchicznej polegającą na scalaniu w każdym kroku iteracji pary najbardziej podobnych multi-uliniowień. Istnieją również metody rozważające wszystkie struktury jednocześnie. Więcej informacji na ten temat można znaleźć w pracy [3].

Ponieważ nie istnieje uniwersalna funkcja miary podobieństwa struktur, obiektywna ocena jakości uliniowienia jest dość trudna. W przypadku projektowania testów metod porównywania struktur ważny jest również dobór testowego zestawu białek. W tej rozprawie będziemy wykorzystywać zbiory testowe i wzorcowe uliniowienia wykorzystane w pracach [15, 3]. Jakość obliczonego uliniowienia będziemy oceniać porównując je z wzorcowym i zliczając jednakowo uliniowione pary aminokwasów.

Lokalny Deskryptor Struktury jest niewielkim fragmentem struktury białka, który może

być rozumiany jako opis lokalnego otoczenia przestrzennego danego aminokwasu. W zasadzie można go zbudować dla każdego aminokwasu rozważanego białka. Aby to uczynić, należy zidentyfikować aminokwasy, z którymi aminokwas założycielski jest *w kontakcie* (oddziałuje fizycznie). Następnie wokół wybranych w ten sposób aminokwasów budowane są elementy poprzez dołączenie dwóch aminokwasów poprzedzających i następujących na łańcuchu białka. Nakładające się elementy łączone są w segmenty (rys. 1). Promień deskryptora jest tym samym przybliżoną miarą zasięgu oddziaływań między aminokwasami. Natomiast sam deskryptor może być traktowany jako wycinek struktury znajdujący się wewnątrz nieregularnej powierzchni opowiadającej praktycznemu zasięgowi wpływu pojedynczego aminokwasu na resztę struktury. Deskryptor zatem w odróżnieniu od tradycyjnych fragmentów struktury opisuje sąsiedztwo w sensie przestrzennym, a nie sekwencyjnym.

Przez uliniowienie dwóch deskryptorów będziemy rozumieć częściowe odwzorowanie pomiędzy zawartymi w nich aminokwasami, które zachowuje relację bycia w kontakcie z aminokwasem założycielskim oraz z ustaloną dokładnością zachowuje kształt odwzorowywanych fragmentów. Problem znajdowania najliczniejszego uliniowienia zadanych deskryptorów jest NP-zupełny z zastrzeżeniem, że dowód NP-zupełności nie uwzględnia biologicznej wiedzy o strukturze białek i rzeczywistych rozmiarach deskryptorów.

Celem pracy było stworzenie metody konkurencyjnej wobec wiodących metod porównywania i klasyfikacji struktur białek, wykorzystującej lokalne deskryptory struktury oraz uwzględniającej przedstawione powyżej postulaty dotyczące przestawień sekwencyjnych i elastyczności porównywanych struktur. Znaczący fragment pracy stanowią rozważania teoretyczne dotyczące złożoności rozważanych problemów.

2 Porównywanie struktur dwóch białek

Pod pojęciem uliniowienia struktur będziemy rozumieć pewne częściowe odwzorowanie pomiędzy zbiorami ich aminokwasów, które w założeniu ma być "izomorfizmem" w sensie pewnych biologicznych relacji występujących pomiędzy aminokwasami w obrębie rozważanych struktur. Stopień podobieństwa struktur określa miara maksymalnego uliniowienia.

Algorytm porównywania deskryptorów może zostać wykorzystany do wykrywania lokalnych podobieństw pomiędzy dwiema strukturami białek. Niech Φ będzie zbiorem takich podobieństw dla zadanej pary struktur. Powiemy, że uliniowienie ma wsparcie w Φ, jeżeli istnieje pewien podzbiór Φ, którego suma elementów jest uliniowieniem zawierającym rozważane[2]. Będziemy rozważać problem znajdowania maksymalnego uliniowienia struktur o wsparciu w zbiorze uliniowień par podobnych deskryptorów. Taka definicja pozwala uwzględniać zamiany kolejności fragmentów.

Problem znajdowania maksymalnego uliniowienia jest NP-zupełny przy założeniu, że deskryptory składają się z co najmniej dwóch segmentów strukturalnych. Natomiast je-

[2]Uliniowienie deskryptorów bądź struktur jest funkcją częściową ze zbioru aminokwasów jednej struktury w drugi. Teoriomnogościowa suma funkcji częściowych rozumianych jako szczególny przypadek relacji również może, choć nie musi, być funkcją częściową.

żeli ograniczyć przestrzeń rozpatrywanych rozwiązań do uliniowień nie zawierających przestawień sekwencyjnych, problem znajdowania maksymalnego uliniowienia jest NP-zupełny dla deskryptorów trzy-segmentowych oraz można go rozwiązać w czasie wielomianowym dla jednosegmentowych deskryptorów. Wynik ten ma charakter ogólny i jest stosowalny do dowolnego rodzaju fragmentów strukturalnych, których podobieństwa służą jako wsparcie budowanego uliniowienia.

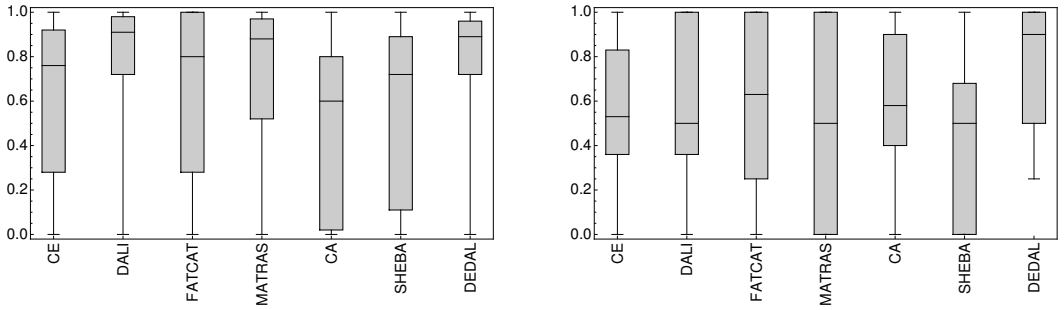
Zaimplementowaliśmy trzy rodzaje algorytmów poszukujących uliniowienia struktur o maksymalnej mierze. Algorytmy dokładne (TS - Tree Search i CTS - Continuous Tree Search) oparte są o analizę drzewa decyzyjnego z wykorzystaniem odcięć, kiedy to możliwe. Mają one wykładniczą pesymistyczną złożoność obliczeniową, ale gwarantują znalezienie uliniowienia maksymalizującego miarę podobieństwa przy założeniu, że jest ona monotoniczna. Algorytm probabilistyczny (REMC) oparty na metodzie Monte-Carlo z wymianą replik może maksymalizować dowolną miarę podobieństwa. Wreszcie zaproponowany został algorytm przybliżony (MS) oparty na twierdzeniu Motzkina-Strausa o związku kliki o maksymalnej liczności z maksimami pewnej formy kwadratowej, co pozwala znajdować najliczniejszą klikę w grafie niesprzeczności z nadzieją, że maksymalizuje ona miarę podobieństwa.

Program i metoda uliniawiania par struktur przy pomocy deskryptorów otrzymały nazwę DEDAL (*DEscriptor Defined ALignment*[6].

Skuteczność metod porównywania struktury jest często oceniania przez wielkość i RMSD obliczonych uliniowień. Takie podejście jest użyteczne w przypadku metod, które optymalizują te parametry, jednak może prowadzić do faworyzowania metod, które dopuszczają błędy w uliniowieniu wynikające z przestrzennej bliskości aminokwasów, zamiast kierować się rzeczywistą rolą, jaką pełnią rozważane aminokwasy oraz "architekturą" cząsteczki białka. To z kolei może prowadzić do błędnych ocen skuteczności, zwłaszcza w przypadkach, gdy podobieństwo strukturalne jest niewielkie i trudne do wykrycia. Dlatego w naszych rozważaniach wykorzystaliśmy zweryfikowaną przez ekspertów bazę danych zawierającą nietrywialne podobieństwa strukturalne i ocenialiśmy stopień podobieństwa obliczonych uliniowień do referencyjnych. Posłużyliśmy się w tym celu liczbą będącą stosunkiem liczby par aminokwasów uliniowionych zgodnie z uliniowieniem referencyjnym do rozmiaru uliniowienia referencyjnego.

Wykorzystane zbiory testowe pochodzą z pracy [15]. Zbiór SISY zawiera 69 nieredundantnych par wybranych z bazy SISYPHUS. Zbiór RIPC zawiera 40 par domen z bazy ASTRAL. Są one podobne strukturalnie, ale trudne do uliniowienia ze względu na występowanie powtórzeń, rozległych insercji lub delecji, permutacji cyrkularnych oraz odkształceń przestrzennych. Dla 23 par autorzy podają referencyjne uliniowienia wynikające z wiedzy o ewolucyjnej lub funkcjonalnej odpowiedniości aminokwasów.

Porównaliśmy wyniki metody DEDAL z wynikami metod CE, DALI, FATCAT, MATRAS, CA i SHEBA obliczonymi w pracy [15] (rys. 2). Wykresy pudełkowe pokazują, że DEDAL jest co najmniej tak samo skuteczny jak DALI i MATRAS (rys. 2a). Średnia dokładność uzyskana na zbiorze SISY wynosi 76% (mediana wynosi 89%). Dla porównania DALI osiąga średnią dokładność 75% (mediana 91%), zaś MATRAS – 67% (mediana 88%). Różnica pomiędzy metodami jest większa w przypadku zbioru RIPC (rys. 2b), gdzie dolny kwartyl jakości uliniowień obliczonych algorytmem DEDAL jest porówny-



Rysunek 2: Jakość, z jaką odtwarzane są uliniowienia ze zbiorów (a) SISY i (b) RIPC. Wykresy pudełkowe prezentują rozkłady jakości uliniowień odtworzonych przez badane metody. Wyniki metod innych niż DEDAL pochodzą z pracy [15].

walny z medianą innych metod. Średnia dokładność wynosi 77% (mediana 90%) podczas, gdy DALI osiąga średnią jakość 60% (mediana 50%).

3 Uliniowienia wielu struktur białek

Omówiony w poprzednim rozdziale problem znajdowania uliniowienia pary struktur można uogólnić. W tym rozdziale zdefiniujemy problem znajdowania uliniowień wielu struktur (multi-uliniowień) i omówimy istotne aspekty tego problemu oraz przedstawimy algorytm ewolucyjny, który może być wykorzystany do jego rozwiązania.

Pojęcie *multi-uliniowienia* jest kalką językową z angielskiego *multi-alignment* i oznacza uliniowienie więcej niż dwóch struktur lub sekwencji. W naszych rozważaniach będziemy rozumieli je szerzej niż jako operację wstawienia spacji w uliniawiane ciągi, aby zmaksymalizować pewną funkcję dopasowania – tak jak w poprzednim rozdziale będziemy dopuszczali przestawienia kolejności.

Dosyć istotnym aspektem jest określenie miary podobieństwa. W przypadku problemu multi-uliniowienia sekwencji stosuje się jedną z trzech strategii:

- suma par (ang. *sum-of-pairs score*, *SP-score*),
- uliniowienie gwiaździste (ang. *star alignment*),
- uliniowienie przy zadanym drzewie filogenetycznym (ang. *tree alignment*).

Każde z tych podejść ma nieco inne właściwości i zastosowania. W szczególności uliniowienie gwiaździste, które polega na znalezieniu sekwencji najbardziej podobnej do uliniawianych będącej niejako ich uśrednieniem, można w kontekście porównywania struktur rozumieć jako poszukiwanie rdzenia wspólnego dla wszystkich porównywanych sekwencji, natomiast maksymalizację sumy podobieństwa wszystkich par jako znajdowanie sumy wszystkich podobieństw. Znajdowania uliniowienia przy zadanym drzewie filogenetycznym jest podejściem pośrednim i ma zastosowanie tylko w sytuacji, gdy na podstawie

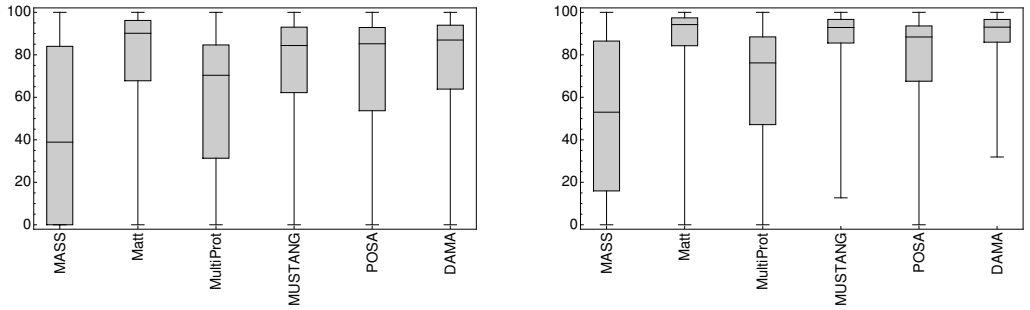
dodatkowej wiedzy można postawić hipotezę o pokrewieństwie filogenetycznym. Należy pamiętać, że niezależnie od przyjętej strategii problem multi-uliniowienia sekwencji dla większości miar podobieństwa symboli jest NP-trudny[8].

W przypadku porównywania struktur przestrzennych białek wykrycie rdzenia wspólnego dla wszystkich porównywanych struktur, o ile takowy istnieje i jest dobrze określony, jest łatwiejsze od znalezienia wszystkich podobieństw. Mimo to w dalszych rozważaniach posługiwaliśmy się strategią sumy par. Przyjmiemy, że multi-uliniowienie można opisać zbiorem uliniowień wszystkich par struktur, które do niego należą. Jednak niestety optymalne multi-uliniowienie nie zawsze odpowiada zbiorowi optymalnych uliniowień par. Uściślając, nie każdy zbiór uliniowień par indukuje multi-uliniowienie struktur. Co więcej problem znajdowania optymalnego multi-uliniowienia (o wsparciu w zadanym zbiorze uliniowień desktyptorowych) jest NP-zupełny nawet w sytuacji, gdyby możliwe było obliczanie optymalnego uliniowienia pary struktur w czasie stałym. Również problem znajdowania optymalnego uliniowienia dwóch multi-uliniowień, mimo że podobny do znajdowania optymalnego uliniowienia pary struktur, jest NP-zupełny przy powyższym założeniu. Niemniej jednak doświadczenie pokazuje, że można go wydajnie rozwiązywać przy użyciu algorytmów wymienionych w poprzednim rozdziale oraz heurystycznego algorytmu zachłannego usuwającego ewentualne niezgodności pomiędzy uliniowieniami wchodzącymi w skład wynikowego multi-uliniowienia.

Opracowaliśmy algorytm ewolucyjny służący do poszukiwania optymalnych multi-uliniowień. Opiera się on na obserwacji, że każdemu multi-uliniowieniu można przypisać pewne drzewo binarne, którego liście są etykietowane uliniawianymi strukturami, a węzły uliniowieniami multi-uliniowień zawartych w potomkach, zaś korzeń etykietowany jest rozważanym multi-uliniowieniem. Takie drzewo o maksymalnej sumie miar multi-uliniowień w węzłach nazwiemy optymalnym drzewem rozpinającym multi-uliniowienie. Operacje mutacji i krzyżowania algorytmu ewolucyjnego polegają na modyfikowaniu multi-uliniowienia w pewnym węźle drzewa rozpinającego lub scalaniu wybranych fragmentów drzew rozpinających. Opisana metoda i program ją implementujący otrzymały nazwę DAMA (*Descriptor Assisted Multiple Alignment*).

Zbiór SISY-multiple[3] jest rozwinięciem zbioru SISY. Zawiera on multi-uliniowienia pochodzące z bazy uliniowień SISYPHUS[2], które zostały oczyszczone przez pominięcie nieaktualnych struktur usuniętych z bazy PDB. Również struktury występujące wielokrotnie w multi-uliniowieniu zostały pominięte, aby uniknąć trudnej do uwzględnienia na etapie oceniania jakości wyników wieloznaczności. Ostatecznie spośród 149 multiuliniowień z bazy SISYPHUS w zbiorze SISY-multiple pozostało 106 liczących co najmniej 3 struktury.

Jakość multi-uliniowień obliczonych programem DAMA ocenialiśmy przez porównywanie z uliniowieniami wzorcowymi. Rozważaliśmy dwie miary podobieństwa pomiędzy uliniowieniem obliczonym a wzorcowym. Miara Q_C jest liczbą pełnych kolumn uliniowionych zgodnie ze wzorcem unormowaną przez liczbę kolumn w uliniowieniu wzorcowym[17]. Mniej restrykcyjna miara Q_P jest proporcjonalna do liczby poprawnie uliniowionych par aminokwasów[16]. Porównaliśmy jakość uliniowień obliczonych programem DAMA z wynikami dla innych metod przedstawionymi w pracy [3] (rys. 3). DAMA daje wyniki o porównywalnej jakości z metodami Matt i MUSTANG, jest nieco lepsza od POSA oraz



Rysunek 3: Jakość z jaką odtwarzane są uliniowienia ze zbioru SISY-multiple. Wykresy pudełkowe prezentują rozkłady jakości uliniowień odtworzonych przez badane metody według miary Q_C (a) i Q_P (b). Wyniki metod innych niż DAMA pochodzą z pracy [3].

istotnie lepsza od metod MASS i MultiProt. Ponadto DAMA ma przewagę nad pozostałymi metodami w przypadku, gdy wymagane jest wykrycie permutacji[3]. Skonstruowane algorytmy realizujące multiuliniowienia i klasyfikacje struktur oraz ich zastosowania do rozwiązywania konkretnych problemów w dziedzinie biologii molekularnej są przedmiotem przygotowywanej do druku pracy [5].

4 Wnioski

Liczba struktur białek zdeponowanych w bazie PDB przyrasta ostatnio o 7-8 tysięcy rocznie. Ogółem znanych jest 73 tysiące struktur pogrupowanych w ok. 1400 foldów. Tak liczny i coraz bardziej kompletny zbiór danych daje możliwości nowych odkryć i w perspektywie nadzieję na znacznie pełniejsze zrozumienie zależności pomiędzy sekwencją, strukturą i funkcją białek. Aby to osiągnąć, konieczne są wszakże coraz dokładniejsze i bardziej wydajne metody analizy tych struktur. Wyniki uzyskane w tej pracy służą temu celowi. W szczególności zrealizowane zostały podstawowe cele sformułowane we Wstępie.

Opisane w pracy metody zostały przetestowane na dostępnych w literaturze zbiorach danych SISYPHUS[2], SISY, RIPC[15], SISY-multiple[3], a ich wyniki skonfrontowane z innymi, popularnymi metodami (CE, DALI, FATCAT, MATRAS, C_α-match, SHEBA, MASS, Matt, MultiProt, MUSTANG, POSA). Dzięki zastosowaniu podziału struktury na relatywnie duże i specyficzne fragmenty, metody obliczania uliniowień wykorzystujące deskryptory lokalnej struktury są skuteczne w tzw. trudnych przypadkach, które obejmują permutacje cyrkularne i inne przestawienia sekwencyjne oraz odkształcenia przestrzenne. Równocześnie w odróżnieniu od metod specjalnie zaprojektowanych z myślą o takowych, bardzo dobrze radzą sobie z przypadkami łatwymi, co wykazało porównanie z metodą DALI.

Projektując testy kierowaliśmy się założeniem, że ocena poprawności uliniowienia struk-

[3]Jedynie MASS wśród pozostałych metod ma możliwość uliniawiania struktur z przestawieniami.

tur nie powinna być wypadkową jego wielkości i jakości rozumianej jako pewna miara związana z superpozycją uliniowionych fragmentów. Istotny jest biologiczny sens uzyskanego dopasowania. W szczególności aminokwasy pełniące odpowiadające sobie funkcje powinny zostać uliniowione. Tylko takie uliniowienie daje użyteczne przesłanki do wyciągania wniosków o pokrewieństwie funkcjonalnym badanych struktur.

Prezentowana metoda opiera się na relatywnie nieskomplikowanej koncepcji identyfikowania zbioru bazowych, strukturalnych klocków (w tym przypadku par podobnych deskryptorów), określenia sposobu dopasowywania klocków do siebie (relacja niesprzeczności uliniowień deskryptorowych) oraz budowania maksymalnych zespołów pasujących do siebie klocków (wyszukiwania klik). Jednak, mimo prostej koncepcji, jakością wyników przewyższa konkurencyjne metody. Jest to argument przemawiający za tezą, że deskryptory lokalnej struktury są dobrym formalizmem opisu struktury białka. Przypuszczalnie ich główną zaletą jest obejmowanie fragmentów struktury, które mimo bliskości przestrzennej mogą być znacznie oddalone w sekwencji.

Zasadniczą trudnością, jaka wiąże się ze składaniem uliniowień z fragmentów, jest kombinatoryczna złożoność tego problemu. W przypadku zbyt małych i niewystarczająco specyficznych par podobnych fragmentów jest zbyt wiele, co prowadzi do zbyt wysokich kosztów obliczeń i wymusza uproszczenia. W metodzie deskryptorowej złożoność kombinatoryczna niejako rozkłada się na dwa poziomy obliczeń: identyfikowanie par podobnych deskryptorów i poszukiwanie optymalnego uliniowienia. Dzięki temu, mimo że formalnie oba problemy obarczone są wykładniczą złożonością obliczeniową, są one rozwiązywalne relatywnie niewielkim kosztem.

Zarówno służący do porównywania par struktur progam DEDAL, jak i obliczający multi-uliniowienia program DAMA zostały udostępnione w ramach stworzonego w tym celu serwisu *Essentia Proteomica* (http://bioexploratorium.pl/EP).

Ponieważ lokalne deskryptory struktury okazały się niezwykle użyteczne do porównywania struktur białek, wyniki zachęcają do dalszego prowadzenia badań w tym kierunku. W szczególności użyteczne wydaje się narzędzie do szybkiego przeszukiwania dużych zbiorów struktur pod kątem podobieństwa do struktury zadanej w zapytaniu. Byłoby to narzędzie analogiczne do stosowanych w dziedzinie sekwencji[1]. Przeszukiwanie rozpoczynałoby się od zidentyfikowania podobieństw pomiędzy deskryptorami występującymi w zadanej sekwencji a klastrami deskryptorów struktur występujących w przeszukiwanej bazie. Na ich podstawie identyfikowane byłyby struktury potencjalnie podobne. Aplikacją dualną do przedstawionej jest narzędzie wykrywające konserwowane motywy sekwencyjne występujące w badanym zbiorze struktur. Następnie takie motywy sekwencyjne mogłyby być powiązane z funkcją lub innymi cechami białek. Ich wykrycie w nowej strukturze wskazywałoby na występowanie powiązanej z nimi cechy. Interesujące wydaje się również obliczenie multi-uliniowień dużych zbiorów struktur i odtwarzanie na ich podstawie ewolucyjnego pokrewieństwa białek.

Z punktu widzenia użytkowników zaprezentowanych i proponowanych narzędzi istotna jest łatwość ich obsługi. Tradycyjna forma serwisu WWW jest wygodna dla osób korzystających okazjonalnie z udostępnianych usług. Niestety natura takich serwisów i chronicznie ograniczone zasoby komputerowe grup badawczych praktycznie wykluczają stworzenie usługi bardziej interaktywnej. Dlatego w przyszłości planujemy opracowanie

narzędzia pozwalającego na w pełni interaktywne obliczanie uliniowień i multi-uliniowień. Szczególnie cennymi funkcjami takiego programu byłaby możliwość wskazania konserwowanych zdaniem użytkownika aminokwasów i wymuszanie ich uliniowienia, wprowadzanie ręcznych poprawek do drzewa rozpinającego multi-uliniowienie, czy też umożliwienie pogłębionego przeszukiwania przestrzeni rozwiązań poprzez wykluczenie uliniowień nie spełniających oczekiwań eksperta.

Literatura

[1] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, Sep 1997.

[2] A. Andreeva, A. Prlic, T. J. Hubbard, and A. G. Murzin. SISYPHUS–structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res*, 35(Database issue):D253–9, 2007.

[3] C. Berbalk, C. S. Schwaiger, and P. Lackner. Accuracy analysis of multiple structure alignments. *Protein Sci*, 18(10):2027–35, 2009.

[4] P. Bork, C. Sander, and A. Valencia. Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Sci*, 2(1):31–40, Jan 1993.

[5] P. Daniluk and B. Lesyng. DAMA – a novel method for multiallingment of protein structures. *In preparation for Bioinformatics.*

[6] P. Daniluk and B. Lesyng. A novel method to compare protein structures using local descriptors. *BMC Bioinformatics, under review.*

[7] S. Dobbins, V. Lesk, and M. Sternberg. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein–protein docking. *Proceedings of the National Academy of Sciences*, 105(30):10390, 2008.

[8] I. Elias. Settling the intractability of multiple alignment. *J Comput Biol*, 13(7):1323–39, Sep 2006.

[9] M. Gerstein and N. Echols. Exploring the range of protein flexibility, from a structural proteomics perspective. *Current opinion in chemical biology*, 8(1):14–19, 2004.

[10] N. V. Grishin. Fold change in evolution of protein structures. *J Struct Biol*, 134(2-3):167–85, 2001.

[11] H. Hasegawa and L. Holm. Advances and pitfalls of protein structural alignment. *Current opinion in structural biology*, 19(3):341–348, 2009.

[12] L. Holm and J. Park. DaliLite workbench for protein structure comparison. *Bioinformatics*, 16(6):566–7, 2000.

[13] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J Mol Biol*, 233(1):123–38, 1993.

[14] Y. Lindqvist and G. Schneider. Circular permutations of natural protein sequences: structural evidence. *Curr Opin Struct Biol*, 7(3):422–7, 1997.

[15] G. Mayr, F. S. Domingues, and P. Lackner. Comparative analysis of protein structure alignments. *BMC Struct Biol*, 7:50, 2007.

[16] J. M. Sauder, J. W. Arthur, and R. L. Dunbrack, Jr. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, 40(1):6–22, Jul 2000.

[17] J. D. Thompson, F. Plewniak, and O. Poch. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*, 27(13):2682–90, Jul 1999.

[18] C. Vogel and V. Morea. Duplication, divergence and formation of novel protein topologies. *Bioessays*, 28(10):973–8, Oct 2006.