Politechnika Poznańska

Wydział Informatyki

Dr inż. Piotr Łukasiak

Autoreferat

Załącznik 2

Poznań 2019

1. **Dane personalne**

Imię i Nazwisko: Piotr Łukasiak
Miejsce pracy: Politechnika Poznańska
Wydział Informatyki
Instytut Informatyki
ul. Piotrowo 3, 60-965 Poznań,

Pracownia Bioinformatyki,
Instytut Chemii Bioorganicznej
Polskiej Akademii Nauk
ul. Noskowskiego 12/14, 61-704 Poznań

2. **Posiadane dyplomy, stopnie naukowe**

31.06.2006 Master of Business Administration w zakresie zarządzania, Carlsson School of Management, University of Minnesota

24.02.2004 stopień doktora inżyniera nauk technicznych w zakresie informatyka, Wydział Informatyki i Zarządzania, Politechnika Poznańska. Tytuł pracy doktorskiej "Algorytmiczne aspekty przewidywania drugorzędowych struktur białek" realizowana w Instytucie Informatyki, pod kierunkiem prof. dr. hab. inż. Jacka Błażewicza.

4.07.2000 stopień magistra inżyniera nauk technicznych na kierunku informatyka w zakresie rozproszonych systemów komputerowych, Wydział Elektryczny, Politechnika Poznańska.

Studia podyplomowe:

z zakresu: Executive Master of Business Administration, Szkoła Główna Handlowa, 2006

3. Informacje o dotychczasowym zatrudnieniu w jednostkach naukowych/ artystycznych.

Asystent: od 01.01.2004 do 29.02.2004, Instytut Informatyki, Wydział Informatyki i Zarządzania, Politechnika Poznańska

Adiunkt: od 01.03.2004 do 01.03.2016, Instytut Informatyki, Wydział Informatyki, Politechnika Poznańska

Starszy wykładowca: od 01.03.2016 do obecnie, Instytut Informatyki, Wydział Informatyki, Politechnika Poznańska

Asystent: od 01.01.2000 do 01.03.2004 Pracownia Bioinformatyki, Instytut Chemii Bioorganicznej Polskiej Akademii Nauk

Adiunkt: od 01.03.2004 do 01.03.2016 Pracownia Bioinformatyki, Instytut Chemii Bioorganicznej Polskiej Akademii Nauk

Starszy specjalista: od 01.03.2016 do obecnie Pracownia Bioinformatyki, Instytut Chemii Bioorganicznej Polskiej Akademii Nauk

4. Osiągnięcie naukowe

4.1. Tytuł osiągnięcia naukowego

Osiągnięciem naukowym zgodnie z art. 16 ust. 2 z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. nr 65, poz. 595 ze zm.) jest monografia pt.:

Wybrane zagadnienia badań operacyjnych w bioinformatyce strukturalnej

4.2. Publikacje prezentujące wyniki badań stanowiące osiągnięcie habilitacyjne

Łukasiak P., „Wybrane zagadnienia badań operacyjnych w bioinformatyce strukturalnej", Poznan Monographs in Computing and Its Applications, Wydawnictwo NAKOM, Poznań 2019

Kopia pracy wchodzącej w skład osiągnięcia naukowego stanowi Załącznik 5.

4.3. Omówienie celu naukowego prac i osiągniętych wyników wraz z omówieniem ich ewentualnego wykorzystania.

Ludzkość obecnie znajduje się w dobie spektakularnego rozwoju biologii molekulamej i genetyki, rzutującego na możliwości kliniczne, terapeutyczne i diagnostyczne w całej medycynie. Nie bez powodu wiek XXI określany jest mianem stulecia medycyny molekularnej oraz medycyny spersonalizowanej. Przełomowe znaczenie w tym zakresie ma stosowanie zaawansowanych technologii o wysokiej przepustowości zarówno w badaniach kwasów nukleinowych, jak i produktów genów. Umożliwiły i przyśpieszyły one poznanie informacji genetycznej zawartej w genomach wielu gatunków roślin, zwierząt i drobnoustrojów (Lesk, 2013). Pozwoliły na porównania znaczenia poszczególnych genów badanych podczas doświadczeń prowadzonych na modelach zwierzęcych z funkcją ich odpowiedników u człowieka, u którego ze względów etycznych różnego typu eksperymenty nie są możliwe. W codziennym języku biologicznym i medycznym pojawiły się pojęcia genarniki (różnych genomik) w kontekście badań genów lub transkryptomiki, proteomiki, matabolomiki, farmakagenarniki i wielu innych "-omik" dotyczących badań ekspresji genów lub analizy ich produktów. Zdobywana w ten

sposób wiedza określana łącznie mianem biologii systemowej, ma coraz szersze zastosowanie zarówno w kontekście chorób dziedzicznych jak i powszechnie występujących chorób nabytych rozpoznawanych w każdej dziedzinie medycyny. Poznanie podstaw molekularnych budowy komórek i uruchamianych w nich szlaków sygnałowych w odpowiedzi na różnego typu czynniki wewnętrzne i zewnętrzne za pomocą maszyn obliczeniowych, są podstawą nowoczesnych terapii dobieranych po kątem potrzeb konkretnego chorego, z uwzględnieniem m. in. indywidualnego metabolizmu leków stosowanych u pacjenta, jego lekooporności lub lekowrażliwości, co wiąże się również z możliwością lepszej oceny i przewidywania niepożądanych działań stosowanych terapii (Ridley, 1999).

Obliczeniowa bioinformatyka wydaje się być oksymoronem dla wielu biologów. Jeszcze nie tak dawno biologia była ukierunkowana na analizę jakościową wiedzy dziedzinowej, podczas gdy zaawansowane metody numeryczne, programowanie, obliczenia i techniki wizualizacji były przedmiotem zainteresowania jedynie inżynierów i fizyków. Częściowo takie podejście jest zrozumiałe, gdyż w ścisłym znaczeniu bioinformatyka odzwierciedla sposób w jaki informacja jest reprezentowana i przesyłana w systemach biologicznych, począwszy od poziomu molekularnego do globalnych procesów rządzących funkcjonowaniem wszystkich organizmów żywych.

Jednak z praktycznego punktu widzenia bioinformatyka jest nauką, która obejmuje zbieranie, manipulowanie, analizowanie, i przesyłanie ogromnych ilości danych przy użyciu maszyn obliczeniowych. Mówiąc precyzyjniej, bioinformatyka (Lesk 2013) jest interdyscyplinarnym obszarem wiedzy łączącym elementy biologii, nauk obliczeniowych i technologii informacyjnej, której zadaniem jest:

- analiza i interpretacja danych biologicznych,
- opracowywanie nowych algorytmów i metod statystycznych dla problemów biologicznych,
- opracowywanie i wdrażanie narzędzi umożliwiających wydajne zarządzanie różnymi rodzajami informacji.

W rozwijającej się bardzo dynamicznie branży biotechnologicznej, każdy naukowiec i przedsiębiorca ma nadzieję opracować rozwiązanie, które zapewni finansowanie jego dalszych badań i wdrożenie ich do rzeczywistości rynkowej. Opracowanie takiego rozwiązania jest w dzisiejszych czasach zależne od dostępnej technologii obliczeniowej przyspieszającej cykl „testowanie-analiza-synteza" leków, który ma bezpośrednie

przełożenie gospodarcze. Na każde 5000 odkryć badawczych rocznie zgłaszanych w USA przez farmaceutyczne laboratoria badawczo-rozwojowe, jedynie 5 przechodzi do etapów testów na ludziach, a tylko jedno z nich jest wprowadzane na rynek, przy czym należy mieć na uwadze, iż. średni czas wprowadzenia takiego leku na rynek wynosi ponad 12 lat. Ponieważ firmy farmaceutyczne otrzymują ograniczony okres wyłączności na proces patentowy, kluczowe jest jak skrócenie wprowadzenia danego leku do sprzedaży.

Większość firm farmaceutycznych postrzega komputeryzację jako środek pozwalający pokonać wspomniane przeszkody. Dostępne metody obliczeniowe pozwalają na skrócenie czasu analizy przydatności danego rozwiązania farmaceutycznego mogąc być zastosowanymi w bardzo szerokim zakresie, w szczególności ma to miejsce przy przewidywaniu skuteczności i skutków ubocznych leków opartych na analizie genomu, wydajnej wizualizacji i analizie struktur białkowych mając na celu szybsze zrozumienie i przewidzenie skuteczności określonych leków.

Zanim jednak metody obliczeniowe mogłyby zostać zastosowane w zagadnieniach biologicznych, muszą być zdefiniowane w opisie analizowanych danych pewne standardy umożliwiające rozpoczęcie ich przetwarzania.

Przykładowo, przeszukując medyczną bazę danych pod kątem wyników badań klinicznych związanych z daną chorobą, musi być dostępny a priori standardowy słownik pojęć do zakodowania tych informacji, tak aby w efektywny sposób móc te dane analizować.

Jednym z głównych wyzwań stojących przed bioinformatyką jest nadążanie za najnowszymi technikami i odkryciami w biologii molekularnej i informatyce. Odkrycia i rozwój przyrastają wykładniczo w obu dziedzinach. Początkowo postęp w obu dziedzinach był niezależny i w większości ze sobą niepowiązany. Jednak z czasem wraz ze wzrostem rozwoju technologicznego w dziedzinie informatyki, postęp w dziedzinie biologii stał się nierozerwalnie uzależniony od dostępności maszyn obliczeniowych i nowych, efektywnych obliczeniowo metod.

Dwa kluczowe wydarzenia pod koniec lat dwudziestych ubiegłego wieku, które spowodowały przełom w obu dyscyplinach to odkrycie penicyliny przez Alexandra Fleminga i mechaniczny komputer analogowy (Vannevar Integraph Bush Product), który mógł rozwiązywać proste równania. W latach trzydziestych Alan Turing, brytyjski matematyk, opracował swój teoretyczny model maszyny obliczeniowej, na którym oparte są wszystkie nowoczesne maszyny obliczeniowe. Model Turinga określa podstawowe

właściwości systemu obliczeniowego: skończony program, duża baza danych i deterministyczny tryb obliczeń krok po kroku. Co więcej, architektura jego hipotetycznej maszyny, która ma skończoną liczbę dyskretnych stanów, używa skończonego alfabetu i jest zasilana przez nieskończenie długą taśmę, jest uderzająco podobna do procesu translacji RNA na białka.

Również analizując teorię informacji, można zauważyć, iż odpowiada ona procesom biologicznym. W modelu teorii informacji Shannona (Shannon, 1948) opisany system komunikacji składa się z pięciu głównych elementów: źródła informacji, nadajnika, medium, odbiornika i odbiorcy. W tym modelu źródłem informacji może być nośnik elektroniczny zawierający informację o sekwencji całego ludzkiego genomu lub cały genom w danym organizmie. Informacja jest przesyłana przez medium. Sygnałem może być sekwencja nukleotydów w cząsteczce DNA lub ładunek elektryczny w nośniku danych. Medium może być macierz wewnątrzkomórkowa, w której znajduje się DNA lub ścieżki, którymi musi zostać przesłany ładunek elektryczny, aby odczytać informację z elektronicznego nośnika. Przy propagacji pożądanego sygnału jest on do pewnego stopnia zniekształcany przez szum informacyjny. W komórce szum może być spowodowany ciepłem, światłem lub promieniowaniem jonizującym, a w przypadku nośnika, szum może być spowodowany np. wibracjami lub innymi polami natury elektromagnetycznej pochodzącymi z otoczenia. Gdy sygnał jest przechwytywany, odbiornik dekoduje wiadomość lub informację dostarczając ją do odbiorcy. W modelu Shannona informacja jest oddzielona od sygnału. Na przykład ciąg bitowy jest sygnałem, który musi zostać przetworzony, aby zinterpretować przesyłaną wiadomość. Analogicznie, nić RNA jest sygnałem, który jest przenoszony od jądra do cytoplazmy, a komunikat jest specyficzną instrukcją syntezy białek.

Centralny dogmat biologii molekularnej (CDBM) został sformułowany przez Francisa Cricka (Watson i Crick, 1953), według którego przepływ informacji genetycznej następuje od DNA poprzez RNA do białka. Teoria informacji Shanona odnosi się w równym stopniu do przepływu informacji podczas procesu transkrypcji (DNA – RNA), jak i procesu translacji.

Traktując CDBM jako proces przepływu informacji, można przyjąć, iż:

- Dane są reprezentowane jako liczby lub inne struktury pochodzące z obserwacji, eksperymentu lub obliczeń,

- Informacje to dane w kontekście - zbiór danych i powiązanych z nimi wyjaśnień, interpretacji, jak również inne materiały dotyczące konkretnego obiektu, zdarzenia lub procesu
- Metadane to dane dotyczące kontekstu, w którym wykorzystywane są informacje, takie jak opisowe podsumowania, kategoryzacja danych czy też informacja na wysokim poziomie.

Powyższe założenia pozwalają przekształcić praktycznie każdy problem natury biologicznej na jego postać matematyczną, co poprzez zastosowanie odpowiednich metod obliczeniowych umożliwia szybsze znalezienie rozwiązania dla analizowanej instancji. Podstawowe zastosowania maszyn obliczeniowych w bioinformatyce dotyczą następujących aspektów (Hein i in, 2013):

1. Kontrola procesu (kontrola pracy urządzeń, robotyka, automatyczne zbieranie danych)
2. Archiwizacja (bazy danych, Infrastruktura IT, słownik pojęć)
3. Przetwarzanie i analiza danych (dopasowywanie wzorca, symulacja, eksploracja danych, wyszukiwanie informacji, analiza statystyczna, wizualizacja)
4. Komunikacja (publikacje lokalne, publikowanie w sieci, internet)

Zastosowania informatyki w kontekście kontroli procesu, wiążą się bezpośrednio z elementami fizycznymi kontrolując np. silniki krokowe, zespoły ramion robota, liczniki zdarzeń i inne urządzenia wejściowe i wyjściowe.

Techniki przetwarzania numerycznego znajdują swoje zastosowanie w wielu miejscach bioinformatyki począwszy od analizy sekwencji, analizy danych z mikromacierzy, do wykrywania genów, przewidywania struktury białek czy też analizy filogenetycznej.

Szybkość przetwarzania informacji ma kluczowe znaczenie dla analizy danych biologicznych, zwłaszcza gdy istnieją ogromne ilości danych związanych z przetwarzaniem, dopasowaniem sekwencji i przewidywaniem sekwencji. Craig Venter z Celera Genomics potwierdził swoje słowa „Szybkość ma znaczenie, odkrycie nie może czekać" dzięki superkomputerowi o wartości 80 milionów dolarów.

Najbardziej rozpowszechnionym działaniem w pracach bioinformatycznych jest przeszukiwanie baz danych i wyszukiwanie wzorca w sekwencji nukleotydów (aminokwasów). Ponieważ w bioinformatyce większość danych ma formę abstrakcyjną, potrzebne są technologie wizualizacji aby użytkownik łatwiej mógł interpretować

otrzymane dane. Ta potrzeba jest najbardziej widoczna w obszarach wizualizacji sekwencji, rozwoju interfejsu użytkownika, wizualizacji struktur białek i analiz statystycznych. Technologie wizualizacji mogą zapewnić intuicyjną reprezentację zależności pomiędzy dużymi grupami obiektów lub punktów danych, ukazując również kontekst problemu.

Wszystkim podjętym pracom towarzyszył warunek podstawowy o praktycznym ich wykorzystaniu. Nadrzędnym wątkiem badawczym przyświecającym również zagadnieniom poruszanym poniżej jest zweryfikowanie prawdziwości hipotezy o praktycznej możliwości stworzenia rozwiązań na bazie metod obliczeniowych i narzędzi informatycznych zastosowanych w zagadnieniach biologicznych w celu podwyższenia jakości naszego życia poprzez medycynę spersonalizowaną. Problem ten jest niestety bardzo złożony, jednak rozwiązując jego podproblemy prawdopodobnie będzie można rozwiązać również problem podstawowy.

Główny wątek badawczy podjęty w niniejszych rozważaniach należy rozpatrywać na dwóch płaszczyznach, ze względu na interdyscyplinarność dziedziny bioinformatyki. Na płaszczyźnie bioinformatycznej dotyczył on różnorodnych aspektów związanych z przewidywaniem i analizą struktur przestrzennych białek wraz z oceną jakości ich modeli przestrzennych. Problem ten był głównie rozważany na przykładzie struktur białkowych, aczkolwiek ze względu na podobieństwo matematycznego modelu struktury białka do modelu cząsteczki RNA, w wybranych aspektach został również rozszerzony na struktury RNA. Na płaszczyźnie informatycznej watek badawczy związany był z opracowaniem efektywnych obliczeniowo metod analitycznych rozwiązujących wspomniany problem na różnych poziomach złożoności wraz z opracowaniem analitycznych metod wizualizacji biologicznych struktur przestrzennych.

W ramach przeprowadzonych badań zrealizowane zostały następujące prace:

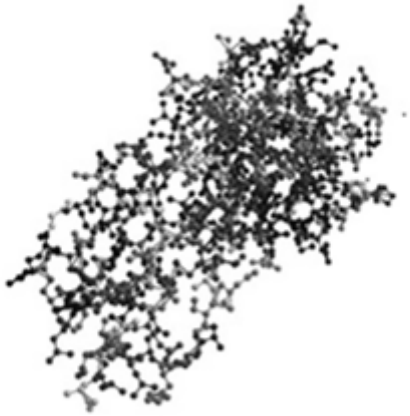
- zaproponowano heurystyczne rozwiązania uproszczonego modelu HP struktur białkowych (Blazewicz i in, 2005)
- zaproponowano kombinatoryczny model podziału białek na domeny (Milostan i Lukasiak, 2016)
- zaproponowano obliczeniowy model składania białka z konserwatywnych podstruktur zwanych deskryptorami w tym:
    - zaproponowano unikatowy model bazy danych podstruktur białkowych zwanych bibliotekami deskryptorów

  - zaproponowano nową definicję struktury deskryptora utrzymującą relację zwrotności przy ich porównywaniu
  - zaproponowano efektywne obliczeniowo algorytmy porównywania deskryptorów (Antczak i in, 2016)
  - zaproponowano strategię grupowania struktur deskryptorów
  - zaproponowano obliczeniową metodę rekonstrukcji łańcucha głównego białka w oparciu o stworzone biblioteki deskryptorów
- opracowano algorytm lokalnego i globalnego porównywania modelu przestrzennego struktury białka ze strukturą referencyjną (Lukasiak i in, 2015)
- opracowano algorytm lokalnego i globalnego porównywania modelu przestrzennego struktury RNA ze strukturą referencyjną (Lukasiak i in, 2013)
- opracowano algorytm oceny jakości modeli białka w oparciu o dane fizykochemiczne powiązane z aminokwasami
- opracowano miarę oceny jakości struktur przestrzennych cząsteczek biologicznych w oparciu o otoczenie sferyczne (Lukasiak, 2012)

Tytułem wstępu do dalszych rozważań warto wspomnieć, iż białka są związkami organicznymi powstałymi z połączenia ze sobą cząsteczek nazywanych aminokwasami. Każde białko jest charakteryzowane unikalnym układem łańcucha aminokwasów zwanym sekwencją (Rysunek 1). W sekwencji może występować 20 różnych aminokwasów, a długość sekwencji może wynosić od kilkunastu do ponad tysiąca aminokwasów. W białku, wzdłuż łańcucha aminokwasów można wyróżnić konserwatywne fragmenty struktury przestrzennej (mające podobny kształt przestrzenny różniąc się jednocześnie sekwencją) zwane strukturami drugorzędowymi. Ułożenie łańcucha aminokwasów w przestrzeni zwane jest strukturą przestrzenną lub trójwymiarową (Lesk, 2013). W celu uproszczenia rozważań można przyjąć, iż zazwyczaj łańcuch białkowy ma tendencję do osiągania tzw., stabilnej struktury natywnej charakteryzującej się optymalnym wypełnieniem wewnętrznych przestrzeni białka minimalizując powierzchnię kontaktu z otoczeniem (Rysunek 2). Białko może zawierać kilka tzw. centrów gęstości aminokwasów, niezależnych funkcjonalnie, zwanych domenami. W odniesieniu do RNA, zamiast zbioru 20 aminokwasów, sekwencja jest tworzona przez 4 typy elementów zwanych nukleotydami (Rysunek 3).

MSFIEKMIGSLNDKREWKAMEARAKALPKEYHHAYKAIQKYMWTSGGPTDWQDTKRIFGG

Rysunek 1 Fragment struktury pierwszorzędowej białka (każda litera odpowiada pojedynczemu aminokwasowi).



Rysunek 2 Struktura trzeciorzędowa białka.

Rysunek 3 Struktura trzeciorzędowa RNA.

Dla wspomnianych powyżej cząsteczek można w prosty sposób zdefiniować ich model matematyczny, w którym są one reprezentowane przez zbiór $A=\{a_i,...a_n\}$, gdzie i=1...n, a_i odpowiada aminokwasowi (nukleotydowi) znajdującemu się na i-tej pozycji w sekwencji, zaś n reprezentuje liczbę aminokwasów (nukleotydów) w sekwencji. Dodatkowo, z każdym elementem zbioru A powiązany jest 3-elementowy wektor W, gdzie poszczególne pozycje wektora odpowiadają wartości poszczególnych współrzędnych x,y,z w przestrzeni kartezjańskiej. Mając tak zdefiniowany model można postawić pytanie związane z przewidywaniem struktury przestrzennej cząsteczki tzn. czy możliwe jest na podstawie samego wektora A stworzyć wektor W, czyli czy na podstawie samej

sekwencji aminokwasów możemy wywieść strukturę przestrzenną białka (RNA). Problem ten jest do chwili obecnej nierozstrzygnięty, natomiast warto zwrócić uwagę, iż decyzyjna wersja tego problemu jest łatwa w praktyce, gdyż na pytanie czy istnieje struktura przestrzenna odpowiadająca zadanej sekwencji odpowiedź brzmi „TAK", ponieważ sekwencja została uzyskana z istniejącego organizmu. Oczywiście problem zmienia swoją trudność, jeśli dopuszczamy jako instancje wejściowe sekwencje wygenerowane sztucznie. Ewentualnie można postawić pytania pomocnicze, jaka minimalna informacja dodatkowa jest potrzebna, aby problem można było rozwiązać.

Powyższy problem można zdekomponować na podproblemy inspirowane np. hierarchiczną strukturą białek. Biorąc jednak pod uwagę wzrost możliwości maszyn obliczeniowych, jak również dużą liczbę metod stosowanych do przewidywania struktur przestrzennych, pojawia się kolejny, bardzo istotny problem do rozwiązania. Ogromna liczba potencjalnych modeli przestrzennych dla tej samej struktury powoduje, iż bardzo ważnym zagadnieniem staje się ocena jakości wygenerowanych modeli. Ocena ta może być realizowana przy założeniu, iż jeżeli dobry algorytm potrafi wygenerować struktury o akceptowalnej jakości w sposób powtarzalny, to możemy przyjąć założenie, iż wstępna weryfikacja jest wystarczająca, aby uwiarygodnić otrzymane wyniki. Realizacja takiego założenia była przyczyną stworzenia konkursu CASP (Moult i in, 2013), polegającego na udostępnianiu wybranych sekwencji białek całej społeczności naukowej (nieopublikowane struktury przestrzenne na czas konkursu są jedynie do wiadomości organizatorów), a następnie porównanie nadesłanych modeli z rzeczywistym rozwiązaniem, co pozwoli określić, które metody są najefektywniejsze. Tak postawiony problem może zostać przedstawiony jako wyliczenie w oparciu o zdefiniowaną miarę jak bardzo różni się struktura białka A ze zdefiniowanym wektorem współrzędnych W_A od struktury rzeczywistej (referencyjnej) B ze skojarzonym wektorem współrzędnych W_B.

Analogiczne pytanie można postawić w odniesieniu do struktur RNA, na który to problem stara się odpowiedzieć inicjatywa zwana RNA-Puzzles będąca odpowiednikiem CASP'u dla cząsteczek RNA. Liczba zdefiniowanych miar jest duża zarówno dla białek, jak i RNA, jednak pomimo tego do tej pory nie ma jednoznacznej jednej miary, która byłaby zaakceptowana jako najlepiej odzwierciedlająca różnice pomiędzy strukturami. Trudniejszym problemem jest ocena jakości modelu strukturalnego bez znajomości struktury referencyjnej. W tym przypadku porównanie na bazie geometrii nie jest możliwe (ze względu na brak struktury referencyjnej), w związku z czym niezbędne jest pozyskanie dodatkowej informacji opartej o homologię sekwencyjną i strukturalną do

struktur istniejących lub przeanalizowanie numerycznych parametrów fizykochemicznych powiązanych z „cegiełkami" sekwencyjnymi (aminokwasami lub nukleotydami) tworząc w oparciu o nie odpowiednie reguły klasyfikujące.

Ponieważ problem przewidywania struktur przestrzennych białek w przestrzeni kartezjańskiej jest problemem złożonym, zaproponowane zostały uproszczone modele przestrzeni, dyskretyzując potencjalne wartości współrzędnych, które mogą przyjmować poszczególne atomy. Tak zdefiniowana przestrzeń powoduje, że uzyskane modele już z założenia będą obarczone błędem w porównaniu do struktury referencyjnej, jednak stanowią one bardzo często etap pośredni, za pomocą którego identyfikowane są podstawowe czynniki procesu zwijania, a jednocześnie struktura wynikowa uzyskiwana jest w skończonym czasie. Do dalszych badań wybrano model HP.

Celem tego modelu jest zbadanie, w jaki sposób oddziaływania hydrofobowe wpływają na zwijanie się białka. Model HP zaproponowano na podstawie obserwacji, potwierdzających że większość hydrofobowych aminokwasów jest zlokalizowana wewnątrz molekuły białka (Dill, 1999).

Uproszczenie reprezentacji białka w modelu HP polega na tym, iż każdy aminokwas jest modelowany jako pojedynczy „koralik" lub „punkt" skończonego zbioru typów (najczęściej są tylko dwa), które umieszczane są w punktach wierzchołkowych zdyskretyzowanej przestrzeni, którą jest krata. Aby zagwarantować ciągłość łańcucha białkowego, sąsiadujące ze sobą w sekwencji reszty aminokwasowe muszą znajdować się w sąsiednich wierzchołkach sieci. Ograniczenia przestrzenne są wyrażane przez założenie, że nie można umieścić więcej niż jednej reszty aminokwasowej w tym samym wierzchołku sieci.

Reprezentacje białek w modelu HP mogą być przedstawiane w różnych rodzajach siatek (zazwyczaj kwadratowych lub trójkątnych), w dwóch lub trzech wymiarach.

Białka reprezentowane w ten sposób przypominają prawdziwe białka poprzez wprowadzenie funkcji energetycznej i zestawu warunków, które określają energię oddziaływania między aminokwasami zajmującymi sąsiadujące miejsca sieci. Funkcja energetyczna naśladuje interakcje pomiędzy aminokwasami w prawdziwych białkach, które obejmują efekty wiązań sferycznych, hydrofobowych i wodorowych. Aminokwas jest reprezentowany jako kula (przyjęto, iż czarna kula reprezentuje aminokwas hydrofobowy, biała kula reprezentuje aminokwas polarny), a wiązania je łączące są przedstawione jako linie łączące odpowiednie kule. Kule są podzielone na typy, a funkcja

energetyczna określa interakcje w zależności od rodzaju kulki, tak jak różne rodzaje aminokwasów oddziałują inaczej. Jeden z najpopularniejszych modeli kratowych, model hydrofobowo-polarny (model HP), ma tylko dwa typy kulek - hydrofobowy (H) i polarny (P) - naśladując efekt hydrofobowy i określając interakcję między kulami typu H.

Dla każdej sekwencji przedstawionej w dowolnej siatce, funkcja energetyczna może być szybko obliczona z funkcji celu. Dla prostego modelu HP jest to wyliczenie wszystkich kontaktów między resztami H sąsiadującymi w strukturze, ale nie w łańcuchu. Białkową sekwencję uważa się za najbardziej zbliżoną do rzeczywistego białka tylko wtedy, gdy posiada pojedynczą strukturę o stanie energetycznym niższym niż to samo białko przedstawione w jakiejkolwiek innej strukturze. Jest to energetyczny stan podstawowy lub stan macierzysty. Względne pozycje kul w stanie natywnym stanowią trzeciorzędową strukturę białka kratowego.

NP-trudność złożoności obliczeniowej modelu HP została zaprezentowana przez Ungera i Moulta (Unger i Moult 1993). W kolejnych latach wykazano również, że decyzyjna wersja tego problemu dla modelu HP zarówno w przestrzeni dwuwymiarowej, jak i trójwymiarowej jest NP-zupełna (Crescenzi i in. 1998).

Ze względu na wspomnianą złożoność tego modelu nastąpiła próba jego rozwiązania za pomocą algorytmów aproksymacyjnych przeszukujących przestrzeń rozwiązań problemu bez gwarancji znalezienia rozwiązania optymalnego (Blazewicz i in 2004, 2005). Większość stosowanych metod opiera się na podejściach wyszukiwania, takich jak dynamika molekularna czy też metoda Monte Carlo (Beutler i Dill 1996). Unger i Moult (1993) zaproponowali implementację algorytmu genetycznego, która jest znacznie szybsza niż tradycyjna metoda Monte Carlo w testach kratowych. O'Toole i Panagitopolus dostosowali niektóre warianty tej metody dla większych cząsteczek (O'Toole i Panagiotopoulos 1992). Metoda oparta na prawie tej samej idei została zaproponowana przez Beutlera i Dilla (Beutler i Dill 1996). Toma i Toma pokazali metodę interakcji kontaktowych (Toma i Toma 1996). Algorytm aproksymacji, który działa w czasie liniowym i gwarantuje, że energia znalezionej konformacji jest co najmniej równa 3/8 wartości optymalnej, została zaproponowana przez Harta i Istraila (Hart i Istrail 1996). Yue i Dill zaproponowali przynajmniej jeden dokładny algorytm (Yue i Dill 1993), który przeszukuje całą przestrzeń konformacji danej sekwencji za pomocą metody podziału i ograniczeń. Ograniczenia i odcięcia zastosowane w zaproponowanych metodach

ograniczają złożoność przestrzeni poszukiwań od 5n (w przypadku modelu 3D i długości sekwencji n) do około 1,125n (Yue i in. 1995, Yue i Dill 1993).

Rzeczywista (natywna) struktura danego białka charakteryzuje się minimalną energią swobody w porównaniu do innych struktur utworzonych z tej samej sekwencji. W związku z powyższym problem znalezienia takiej struktury może zostać zdefiniowany jako problem minimalizacji funkcji energii. W przypadku modelu HP może on mieć następującą postać:

$$min_{\underline{a}} E(\underline{s},\underline{a})$$

$$E(\underline{s},\underline{a}) = \xi * HH_c(\underline{s},\underline{a})$$

gdzie:

– $\underline{s}$ – sekwencja zawierająca n aminokwasów; $s_i = 1$, jeżeli aminokwas na pozycji i-tej jest hydrofobowy; $s_i = 0$, , jeżeli aminokwas na pozycji i-tej jest polarny.

– $\underline{a}$ - (n–2) elementowy wektor kątów zdefiniowany jako trójki kolejnych trzech aminokwasów w sekwencji;

– HH_c – funkcja zliczająca sąsiadujące ze sobą w przestrzeni kontakty pomiędzy aminokwasami typu H, które nie są sąsiadami w sekwencji (są sąsiadami topologicznymi).

– ξ – stała mniejsza od zera odzwierciedlająca wpływ kontaktów hydrofobowych na wartość funkcji energii. W większości przypadków przyjmuje się wartość równą -1 (ξ = –1).

Definicja problemu dla modelu trójwymiarowego jest podobna do definicji, które zostały już podane dla modeli 2D. Główną różnicą jest sposób obliczania pozycji każdego aminokwasu w przestrzeni i liczbie stopni swobody.

Skonstruowany algorytm opiera się na strategii Tabu Search (TS), która została zaproponowana przez Freda Glovera (Glover 1989; Glover i Laguna 1997). Aby dostosować ją do rozważanego problemu należy zdefiniować podstawowe parametry takie jak sąsiedztwo, ruch, warunek zatrzymania, lista tabu, poziomy aspiracji.

Ruch r przekształca każde rozwiązanie x ze zbioru wszystkich rozwiązań Ω w inne rozwiązanie $x' \in \Omega: x \rightarrow^{r} x'$.

Następujące możliwości ruchu są dopuszczalne:

- zmiana pojedynczego elementu (kąta) z wektora $\underline{\alpha}$ w konformacji sekwencji białka $\underline{s}$ na jedną z sąsiednich wartości.
- zmiana pojedynczego elementu z wektora $\underline{\alpha}$ w konformacji sekwencji białka s na jedną z dopuszczalnych wartości.
- zmiana jednego lub dwóch kolejnych elementów z wektora $\underline{\alpha}$, opisującego konformację białka, na dopuszczalne wartości.
- zmiana jednego, dwóch lub trzech kolejnych elementów z wektora $\underline{\alpha}$ na dopuszczalne wartości.

Proponowany zestaw ruchów gwarantuje szybkie przechodzenie pomiędzy różnymi konformacjami białek, natomiast zmiany kątów występują tylko na kolejnych pozycjach..

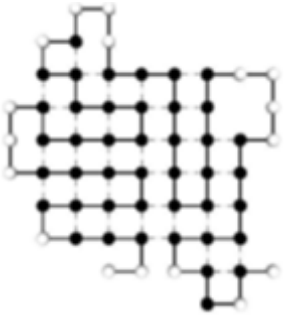
Zbiór rozwiązań $N(x)$ nazywany jest sąsiedztwem x, jeżeli dla każdego $x' \in N(x)$ istnieje ruch r taki, że $x \rightarrow^{r} x'$.

Lista tabu to pamięć krótkotrwała (lista cykliczna), która zawiera ograniczoną liczbę zakazanych ruchów. Jeśli lista jest pełna, dodanie nowego ruchu usuwa ruch z końca listy.

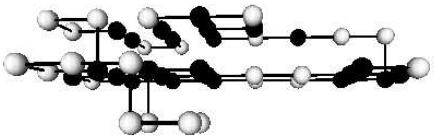
Procedura wyszukiwania może przejść od rozwiązania x do x', wykonując zakazany ruch (z listy tabu), jeśli jednocześnie spełnione są dwa warunki:

- wartość funkcji energii E dla rozwiązania x' jest mniejsza niż energia E_{min} najlepszego znalezionego rozwiązania

nie ma rozwiązania $x \in N(x)$, takiego że $r \notin R_z$, $x \rightarrow x''$ i $E_{x''} \leq E$.



Rysunek 4 Przykładowe rozwiązanie optymalne dla problemu 2D.



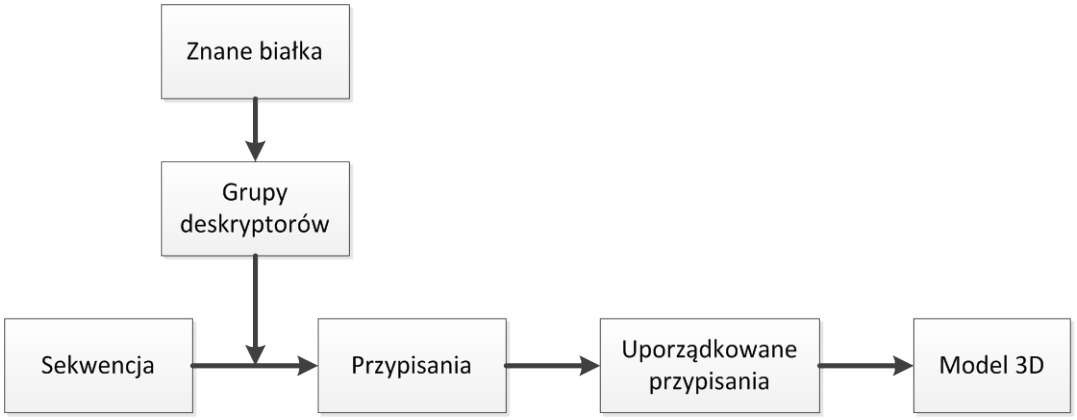
Rysunek 5 Przykładowe rozwiązanie suboptymalne dla problemu 3D.

Zaproponowana metoda została przetestowana na różnych zbiorach testowych znajdując dla modelu 2D optymalne konformacje w ponad 90% przypadków dla sekwencji zawierających do 100 aminokwasów (Rysunek 4). Warunek, że natywna konformacja powinna być stabilna, nie jest jedynym, który musi być brany pod uwagę podczas prowadzonych analiz. Algorytm musi w znaleźć natywną konformację w krótkim czasie, zaczynając od stanu zdenaturowanego charakteryzowanego przez losową populację. W przypadku zbioru testowego 3D jakość rozwiązań uzyskanych przez prezentowaną metodę była niższa, niż w przypadku modelu dwuwymiarowego ze względu na fakt zwiększenia złożoności problemu (Rysunek 5). Duża liczba stopni swobody uniemożliwiała uzyskanie idealnego rozwiązania dla liczby iteracji mniejszej niż 1000. Suboptymalne rozwiązania były na poziomie 60-70% optymalnej wartości funkcji energetycznej. Uzyskane konformacje mogą posłużyć do wstępnej analizy rzeczywistych struktur, jak również posłużyć jako punkt wyjścia do dalszych badań poprzez zastosowanie metod wieloetapowych, gdzie kolejnym poziomem będzie zastosowanie algorytmu genetycznego lub algorytmu Monte Carlo. Wykazano, że za pomocą strategii Tabu Search istnieje możliwość znalezienia konformacji strukturalnych zbliżonych do struktur rzeczywistych

Christian Anfinsen (Anfinsen 1973) przedstawił hipotezę, że struktura przestrzenna każdego białka jest ściśle określona i wynika ze specyficznego ułożenia łańcucha polipeptydowego, który zależny jest od zestawu i kolejności aminokwasów budujących białko. Analiza struktur przestrzennych molekuł biologicznych, jak również rozwiązanie problemu modelowania struktur przestrzennych bezpośrednio z sekwencji pierwszorzędowej (sekwencji aminokwasów lub sekwencji nukleotydów) jest trudna ze względu na bardzo dużą liczbę zmiennych wynikających z fizykochemicznych zależności mogących wpłynąć na proces predykcji. Metody tego typu zazwyczaj modelują struktury obarczone dużym błędem, i wymagają bardzo intensywnego wsparcia ze strony ekspertów dziedzinowych. W celu uproszczenia tego problemu przyjęto założenie, iż podobieństwo sekwencyjne determinuje podobieństwo strukturalne, w związku z czym jeżeli znaleziona zostanie w bazie danych znanych struktur wystarczająco

podobna sekwencja do sekwencji szukanej (powyżej 60% podobieństwa), struktura przestrzenna przez nią tworzona powinna być podobna do struktury szukanej. Przy założeniu pewnego poziomu ogólności rozważań, można stwierdzić, iż strategie działania tych metod polegają na przeszukiwaniu bazy znanych struktur, w celu znalezienia najbardziej zgodnych sekwencji, których struktury mają posłużyć do budowania modelu przestrzennego. Ocena danej struktury białkowej odbywa się przy użyciu specjalnie zaprojektowanej funkcji celu, wybierającej z bazy struktur takie konformacje, które najlepiej odpowiadają analizowanej sekwencji. Może ona odbywać się przy pomocy potencjałów fizyko-chemicznych jak i innych funkcji oceniających zgodność struktur i sekwencji, np. pokrywania się elementów struktury drugorzędowej. Przeprowadzone analizy wskazują jednak, iż podobieństwo sekwencyjne powinno być na poziomie przekraczającym 60%, aby to założenie było możliwe do praktycznego wykorzystania.

Próbując rozwiązać ten problem zaproponowano dekompozycję struktury przestrzennej na podstruktury, jednak w odróżnieniu do dotychczasowych rozwiązań skupiających się na krótkich konserwatywnych odcinkach łańcucha głównego (struktury drugorzędowe), zaproponowano podstruktury mogące zawierać jednocześnie kilka odcinków łańcuchów głównego, stanowiąc jako całość jeden element strukturalny zwany deskryptorem. Paradygmat lokalnych deskryptorów struktur białkowych został opisany (Hvidsten i in. 2003) i jest wykorzystywany w wielu badaniach (Strombergsson i in. 2008, Bjorkholm i in. 2009). W ten sposób można tworzyć i odtwarzać struktury przestrzenne opierając się na lokalnej homologii.



Rysunek 6 Proces rekonstrukcji struktury przestrzennej

Zaproponowany proces rekonstrukcji struktury przestrzennej został przedstawiony na rysunku (Rysunek 6).Poprzez pojęcie deskryptora można rozumieć lokalne,

przestrzenne otoczenie wokół centralnego aminokwasu (lun nukleotydu w odniesieniu do RNA) pełniącego rolę centrum deskryptora. Wszystkie analizy poniżej w celu uproszczenia przedstawione są dla białka. W odniesieniu do RNA proces postępowania jest analogiczny, przy czym zamiast sekwencji aminokwasów analizowana jest sekwencja nukleotydów oraz zamiast atomu C_α analizowany jest odpowiedni, ustalony odgórnie atom łańcucha głównego sekwencji RNA. W przypadku białka otoczenie zawiera te elementy łańcucha polipeptydowego, które zlokalizowane są w przestrzennym sąsiedztwie centralnego aminokwasu. Poprzez taką definicję, deskryptor obejmuje zarówno bliskie jak i odległe sekwencyjnie interakcje pomiędzy poszczególnymi atomami. Liczba i długość fragmentów sekwencji obejmowanych przez deskryptor zależy od kształtu łańcucha głównego i stopnia upakowania łańcuchów bocznych w analizowanym sąsiedztwie.

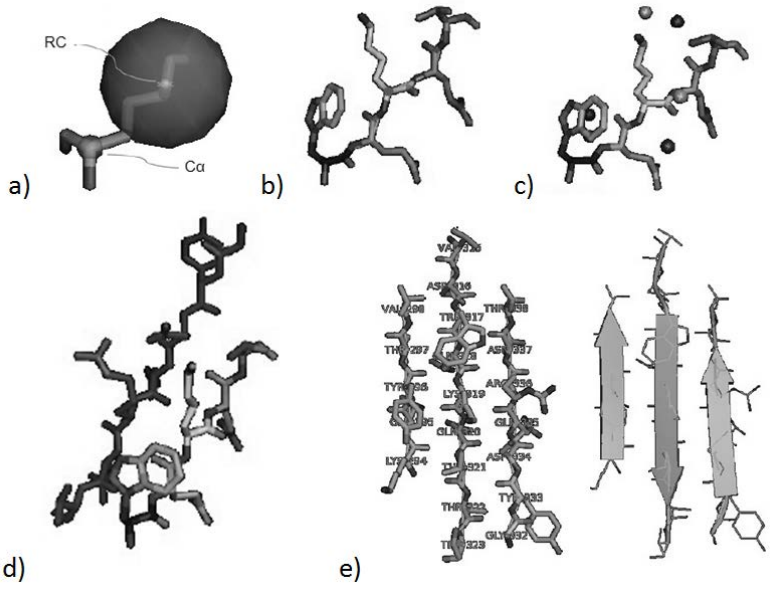
Proces budowy deskryptora składa się z następujących kroków (Rysunek 7):

- Jako aminokwas centralny rozumiany jest aminokwas będący podstawą danego deskryptora. Następnie dla każdego aminokwasu w analizowanej strukturze (z wyjątkiem aminokwasu centralnego), obliczane są dwie odległości euklidesowe pomiędzy analizowanym aminokwasem, a aminokwasem centralnym w odniesieniu do atomu C_α (wyznaczona odległość jest oznaczana jako d_α) oraz „atomu wirtualnego" reprezentującego geometryczny środek łańcucha bocznego RC (wyznaczona odległość jest oznaczana jako d_{RC}). Atom C_α pełni rolę reprezentanta łańcucha głównego, natomiast atom wirtualny RC pełni rolę atomu reprezentującego łańcuch boczny. Dwa aminokwasy są rozważane jako będące w kontakcie, jeżeli jest spełniona poniższa zależność.

  $$(d_{RC} \leq 6.5\ \text{Å}) \vee ((d_{RC} \leq 8.0\ \text{Å}) \wedge (d_{RC} \leq d_\alpha - 0.75\ \text{Å})$$

  Wartości powyższych parametrów ustalono na podstawie przeprowadzonych eksperymentów, z zachowaniem balansu pomiędzy uniwersalnością motywów strukturalnych, a ich przestrzenną odrębnością.
- Dla każdego aminokwasu będącego w kontakcie z aminokwasem centralnym oraz aminokwasu centralnego budowany jest ciągły fragment łańcucha peptydowego (zwany *elementem*) składający się z rozpatrywanego aminokwasu i czterech jego najbliższych sekwencyjnie sąsiadów (po dwóch z każdej strony). Element, który zbudowany został wokół aminokwasu centralnego deskryptora określany jest jako *element centralny*.

- Wszystkie elementy nakładające się strukturalnie tworzą dłuższe fragmenty zwane segmentami. Segment, który zawiera element centralny określany jest jako segment główny.



Rysunek 7 Schemat budowy deskryptorów (a) - atom centralny; (b) – element centralny; (c) – element centralny z atomami będącymi w kontakcie; (d) – segment; (e) – deskryptor.

Deskryptory jako podstruktury zawierające informacje o lokalnym otoczeniu sekwencyjnym, jak i przestrzennym zostały użyte do prób rekonstrukcji struktur przestrzennych. Aby jednak mieć możliwość rozwiązania tego problemu, najpierw zaplanowano stworzenie bazy danych deskryptorów pogrupowanych w biblioteki o podobnym kształcie, następnie opracowanie i zaimplementowanie metody obliczeniowej przypisującej deskryptory do sekwencji i składającej z nich strukturę przestrzenną.

Proces budowy biblioteki grup deskryptorowych składa się z następujących kroków:

- Wygenerowanie zbioru deskryptorów w oparciu o reprezentatywny zbiór znanych struktur białkowych.
- Pogrupowanie zbioru deskryptorów w celu uzyskania grup deskryptorowych przechowujących strukturalnie podobne motywy przestrzenne.
- Odfiltrowywanie grup, których rozmiar, reprezentowany przez liczbę przechowywanych deskryptorów, jest mniejszy od zadanego progu.

Deskryptory pozwalają na utrzymanie zależności wywodzących się z rzeczywistej struktury. Baza danych deskryptorów, została utworzona na podstawie

niehomologicznych domen białkowych zidentyfikowanych w bazie danych SCOP (Murzin i in. 1995) i przechowywanych w bazie ASTRAL 1.75A (Brenner i in. 2000). Jako reprezentatywny zbiór struktur przestrzennych został wybrany zbiór 10281 domen białkowych. Dla każdego aminokwasu każdej domeny został zbudowany deskryptor. W celu zapewnienia maksymalnej prawidłowości danych odfiltrowane zostały wszystkie deskryptory, w których zidentyfikowano przynajmniej jeden niekompletny aminokwas uzyskując 1,663,333 deskryptorów składających się z przynajmniej jednego ciągłego fragmentu łańcucha głównego.

W celu przeprowadzenia operacji grupowania odzwierciedlającej różnice w kształtach geometrycznych odpowiednich deskryptorów, dokonano porównania deskryptorów między sobą. Niestety, ze względu na to, iż deskryptory mogą różnić się od siebie liczbą segmentów, liczbą elementów, jak również liczbą aminokwasów (nukleotydów w przypadku RNA) nie istniała metodologia porównywania deskryptorów. Aby zapewnić wewnętrzne podobieństwo grup zaproponowano dodatkowe ograniczenia.

Strukturalne dopasowanie deskryptorów realizowane jest na poziomie elementów. Wymaga ono na wstępie weryfikacji przestrzennego dopasowania wszystkich kombinacji par elementów zidentyfikowanych w porównywanych deskryptorach za wyjątkiem elementu centralnego. Jeżeli parę elementów deskryptora zawierającą element centralny i dowolny inny element należący do tego deskryptora nazwiemy dupleksem, to obecność centralnego elementu w dupleksie pomaga ustabilizować taką strukturę w przestrzeni 3D podczas procesu dopasowywania. Zgodnie z definicją zaproponowaną przez Hvidstena, para deskryptorów jest klasyfikowana jako strukturalnie podobna, jeśli element centralny charakteryzuje się dobrą jakością (wartość RMSD nie jest większa niż 1,2 Å), a wynikowe dopasowanie jest strukturalnie podobne (jego wartość RMSD jest mniejsza niż 3,5 Å), a współczynnik proporcji elementów i reszt aminokwasowych w porównywanych deskryptorach wynosi odpowiednio nie mniej niż 4/5 i 2/3.

Niech $A=\{a_*,a_1,...,a_n\}$ i $B=\{b_*,b_1,...,b_m\}$, gdzie a_* i b_* są centralnymi elementami poszczególnych deskryptorów. Liczba elementów w deskryptorach musi spełniać następujące nierówności:

$$0{,}8 * |B| \leq |A| \leq 1{,}2 * |B|$$

Dystans pomiędzy deskryptorami jest wyznaczany za pomocą miary *RMSD*. Miara ta może zostać wyliczona jeżeli porównywane cząsteczki zawierają taką samą liczbę

atomów. Niech *RMSD(u,v)* będzie funkcją zwracającą wartość *RMSD* wyliczoną dla cząsteczek *u* i *v*. Dla każdego strukturalnego dopasowania $RMSD(A_*,B_*) \leq 1{,}2$ Å .

Poza elementem centralnym, dopasowanie dotyczy również dupleksów. Zbiór dupleksów może zostać zdefiniowany jako $D_A=\{d_{ai}=(a_*,a_i) : i=1,...,n\}$, $D_B=\{d_bj=(b_*,b_j) : j=1,...,m\}$, zaś pary elementów tych zbiorów są mierzone za pomocą funkcji RMSD.

Element centralny musi znajdować się w każdym dopasowaniu strukturalnym, a całkowita liczba par *N* w dopasowaniu musi spełniać nierówność $N \geq 0{,}8^* |A|$ i $N \geq 0{,}8^*|B|$. Dopasowanie musi zawierać co najmniej 2/3 aminokwasów *A* i *B*.

Wartość *RMSD* wszystkich par dupleksów w dopasowaniu nie może być większa niż *3,5 Å*, tak samo jak globalna wartość *RMSD* wyliczona dla wszystkich dopasowanych podstruktur w porównywanych deskryptorach.

Wprowadzenie powyższych założeń pozwoliło zaproponować algorytm obliczeniowy porównywania deskryptorów między sobą o złożoności wielomianowej (Antczak i in, 2016).

Problem największego strukturalnego dopasowania może zostać zamodelowany jako problem optymalizacji kombinatorycznej, problem największego dopasowania (ND) w następujący sposób:

$$\text{Max} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij},$$

$$\text{przy ogr.} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij} \leq L,$$

$$\sum_{i=1}^{n} x_{ij} \leq 1, \qquad \forall_{j=1,\ldots,m}$$

$$\sum_{j=1}^{m} x_{ij} \leq 1, \qquad \forall_{i=1,\ldots,n}$$

$$x_{ij} \in \{0,1\}, \qquad \forall_{i=1,\ldots,n,\ j=1,\ldots,m}$$

gdzie:

c_{ij} jest kosztem dopasowania elementów i i j

L – ograniczenie całkowitego rozwiązania

x_{ij} – zmienna decyzyjna

W przypadku rozważanego problemu c_{ij}=RMSD(d_{Ai}, d_{Bj}) dla i=1..n, j=1..m, a L jest stałą definiowaną przez użytkownika. Im mniejsza wartość L, tym bardziej dokładne jest otrzymane dopasowanie. W rozważanym problemie liczba elementów wybranych do rozwiązania powinna być podobna do liczności elementów w deskryptorze, w związku z czym można przyjąć, że L zależy od najmniejszej wartości z pary (n,m). Długość dopasowania N jest równa wartości funkcji celu powiększonej o 1 (o element centralny).

Algorytm 1. Problem największego dopasowania jest rozwiązywany w pętli przy malejącej wartości K, pierwsze możliwe rozwiązanie problemu jest optymalne i traktowane jako odpowiedź w tym etapie. Następnie obliczana jest globalna wartość RMSD dla dopasowanych podstruktur białkowych złożonych z par dupleksów uzupełnionych parą elementów centralnych deskryptorów. Gdy globalna wartość RMSD nie jest większa niż 3,5, rozwiązanie jest zwracane jako wyjście algorytmu, w przeciwnym razie rozwiązani nie zostało znalezione.

Algorytm 2. Problem największego dopasowania jest rozwiązywany dla wszystkich wartości K z podanego zakresu, wszystkie uzyskane rozwiązania problemu są zapamiętywane na liście. Następnie, dla każdego rozwiązania na liście obliczane jest globalne RMSD i wśród tych rozwiązań, które spełniają ograniczenie (RMSD ≤ 3,5 Å) jako rozwiązanie wybierane jest to, który składa się z największej liczby dupleksów (podstawowe kryterium optymalizacji) i największej liczbie reszt aminokwasowych (drugorzędne kryterium optymalizacji). Jeśli żadne rozwiązanie nie spełnia ograniczenia globalnego RMSD, rozwiązanie nie istnieje.

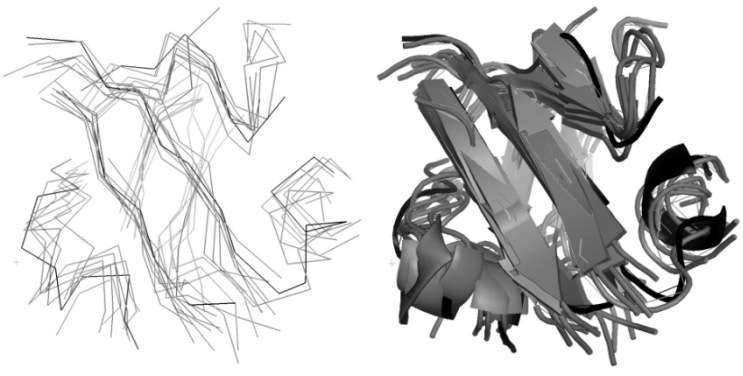
Algorytm 3. Problem największego dopasowania jest rozwiązywany dla wszystkich wartości K z podanego zakresu. Dodatkowo, na liście razem z wszystkimi rozwiązaniami przechowywane są wszystkie możliwe rozwiązania częściowe. Częściowe rozwiązanie opiera się na niepełnym rozwiązaniu ostatecznego rozwiązania uzyskanego dla danego K, gdzie to ostatnie ma postać listy przypisanych par dupleksów posortowanych według ich kosztów w niezmniejszającej się kolejności. Uwzględnione zostają wszystkie wystarczająco długie prefiksy. Rozwiązanie częściowe jest tworzone przez dodawanie w kolejności tych par, które łącznie spełniają warunek równej liczby reszt aminokwasowych w obu dopasowanych podstrukturach deskryptorów. Jeśli dodanie kolejnych par powoduje iż utworzone rozwiązanie częściowe jest niedopuszczalne (z powodu różnej liczby reszt aminokwasowych), są one pomijane. Akceptowalne

częściowe rozwiązanie jest zapamiętywane do dalszego rozważenia, jeśli spełnia ono ograniczenia związane z L i K.

Następnie dla każdej pozycji z listy rozwiązań i rozwiązań częściowych weryfikowana jest spełnialność ograniczenia związanego z globalnym RMSD. Spośród elementów spełniających to ograniczenie, jako odpowiedź wybierane jest to rozwiązanie, które składa się z największej liczby dupleksów, a następnie największej liczby aminokwasów.

W powyższym przypadku problem został zredukowany do modelu kombinatorycznego w oparciu o problem największego dopasowania i rozwiązany algorytmami wielomianowymi w trzech wersjach. Co więcej, zaproponowano algorytm dokładny w czasie wykładniczym. Zaproponowane algorytmy zapewniają symetrię w procesie strukturalnego dopasowania deskryptorów (tj. dopasowanie deskryptora A do deskryptora B daje taki sam wynik jak dopasowanie B do A). Dowiedziono, że zaproponowany model może zostać rozwiązany przez algorytm wielomianowy metodą węgierską w $O(n^3)$, a problem optymalne rozwiązanie porównania dwóch deskryptorów (największe strukturalne dopasowanie) w $O(n^4)$.

Mając zdefiniowaną strategię porównywania deskryptorów, zastosowane zostały schematy klastrowania do stworzenia ich bibliotek (Rysunek 8). Biblioteka grup deskryptorowych została zbudowana z deskryptorów co najmniej trójsegmentowych (własności przestrzenne są obserwowane głównie w jądrze białka) co pozwoliło uzyskać 847,416 deskryptorów. Cały zbiór deskryptorów został podzielony na podzbiory deskryptorów charakteryzujących się równą liczbą elementów (15 podzbiorów deskryptorów od 3 do 17). W wyniku przeprowadzanych eksperymentów uzyskana została biblioteka lokalnych, powtarzających się motywów przestrzennych, które reprezentują stosunkowo różne konformacje geometryczne (klasy podobieństwa sekwencyjnego) (Lukasiak, 2013b).

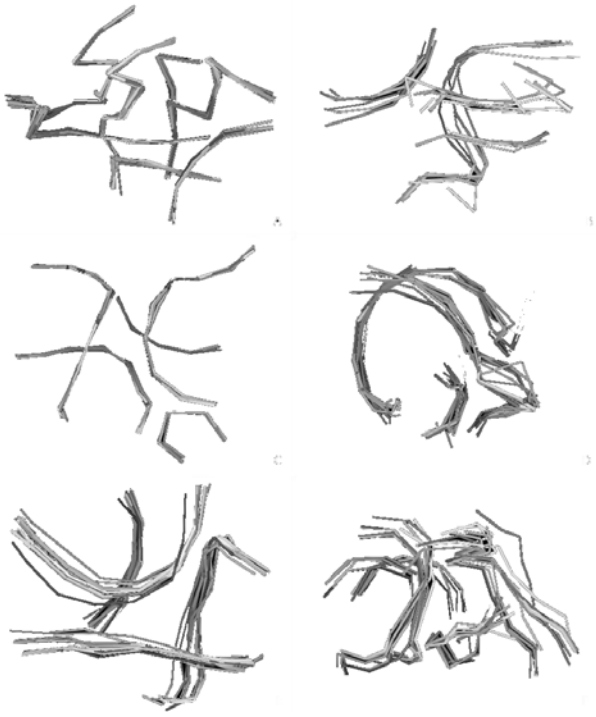


Rysunek 8 Przykładowa biblioteka deskryptorów dla białek.

Każda grupa deskryptorów reprezentuje inny geometryczny kształt uzyskany na podstawie reprezentatywnego zbioru struktur białkowych.

Podstawową miarą oceniającą jakość biblioteki grup deskryptorowych jest pokrycie sekwencyjne wyznaczane poprzez liczbę aminokwasów pokrytych przez deskryptory w stosunku do liczby aminokwasów wszystkich domen biorących udział w konstruowaniu biblioteki. Jak zostało zauważone, wraz ze wzrostem współczynnika minimalnej liczby deskryptorów w grupie, liczba grup przechowywanych w bibliotece znacząco maleje. Podobną zależność można również zaobserwować pomiędzy minimalną liczbą deskryptorów w grupie a całkowitą liczbą deskryptorów przechowywanych w bibliotece. Pokrycie sekwencyjne wyznaczone zarówno w oparciu o jedynie deskryptory założycielskie jak i wszystkie deskryptory w bibliotece wskazuje, że potencjalna użyteczność bibliotek podczas procesu ewaluacji prawidłowości strukturalnej białek jest odwrotnie proporcjonalna do wzrostu współczynnika minimalnej liczby deskryptorów w grupie.

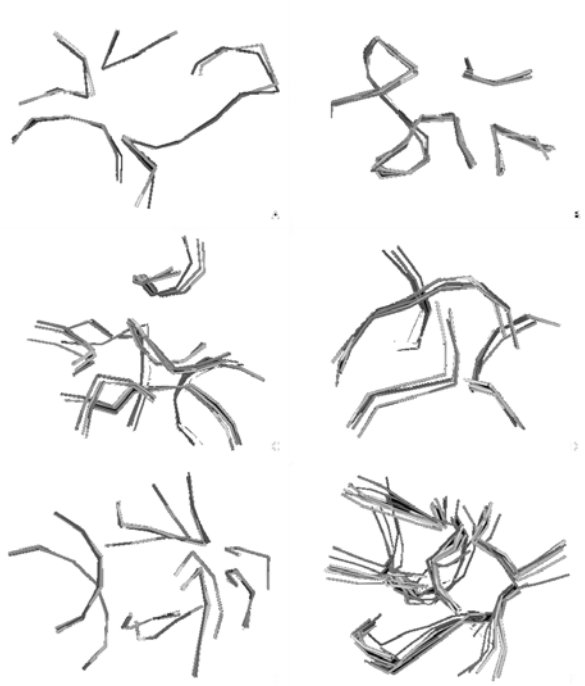
W celu zbudowania biblioteki motywów przestrzennych RNA z puli wcześniej wygenerowanych deskryptorów dla 2100 struktur RNA wybrano ponad dwadzieścia tysięcy deskryptorów spełniających kryteria analogiczne, jak w przypadku deskryptorów białkowych (Rysunek 9).



Rysunek 9 Wizualizacja wybranych bibliotek deskryptorów RNA

Przedstawione biblioteki cechowały się wysokim podobieństwem strukturalnym pomiędzy poszczególnymi deskryptorami wchodzącymi w ich skład (Lukasiak, 2017).

Ponadto przeprowadzono klastrowanie dla zestawu deskryptorów pochodzących z różnych struktur zawierających pseudowęzły (Rysunek 10).



Rysunek 10 Wizualizacja wybranych bibliotek deskryptorów RNA dla struktur z pseudowęzłami.

Wygenerowanie bibliotek deskryptorów pozwoliło zaproponować metodologię rekonstrukcji struktury przestrzennej (Lukasiak, 2012)).

Rekonstrukcja struktur na podstawie fragmentów przypisań (deskryptorów) została zaimplementowana w postaci algorytmu heurystycznego. Algorytm rozpoczyna działanie od wczytania sekwencji składanego białka oraz zbioru przypisań. Przypisania zawierają jedynie zakresy odpowiadających sobie aminokwasów białka oraz przypisanego deskryptora, dlatego konieczne jest jeszcze pobranie struktury deskryptora z bazy danych deskryptorów. Po wczytaniu wszystkich danych następuje proces składania przypisań w coraz większe struktury według jednej z poniżej opisanych metod. Składanie trwa tak długo jak pozwala na to konkretny algorytm, generując zbiór potencjalnych rozwiązań. Parametr konfiguracyjny ogranicza maksymalną liczbę zwracanych modeli.

Generacja modelu na podstawie przypisań wymaga połączenia podzbioru przypisań w jedną strukturę. Zastosowana została prosta metoda łączenia przypisań poprzez

uśrednianie współrzędnych atomów łączonych przypisań. Po określeniu, które przypisania należy ze sobą połączyć, struktury nakładane są na siebie w oparciu o część wspólną wynikającą z pokrywanej przez przypisania sekwencji. Atomy części niezależnych pozostają bez zmian. Dla każdego atomu należącego do części wspólnej wyznaczony zostaje nowy zestaw współrzędnych poprzez uśrednienie współrzędnych pary łączonych atomów.

Każdy wygenerowany model weryfikowany jest pod względem poprawności oraz sortowany według wartości funkcji oceny. Weryfikacja polega na sprawdzeniu szeregu własności złożonego modelu, nie wymagających znajomości składanego białka.

Weryfikowane są kolejno trzy parametry:

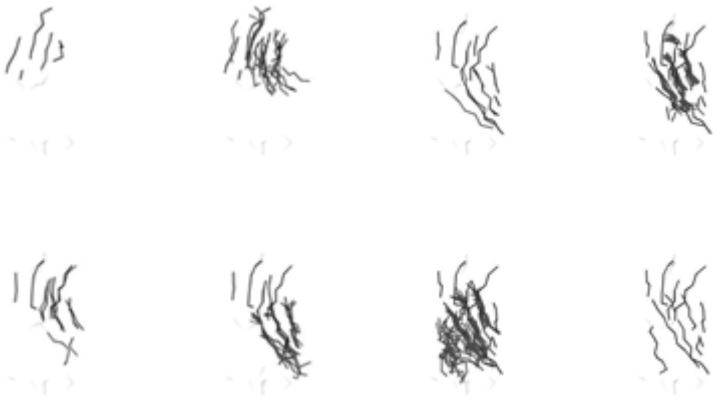
- odległość między kolejnymi atomami C_α ciągłych fragmentów złożonej struktury. Wartość wykraczająca poza zakres zdefiniowany przez parametry konfiguracyjne powoduje odrzucenie modelu.
- Kolizje łańcucha głównego - odrzucane zostają modele, w których niesąsiadujące ze sobą sekwencyjnie atomy C_α kolidują w przestrzeni. Sprawdzenie kolizji odbywa się przy użyciu potencjału Scwrl'a, opartego na energii van der Waals'a.
- Stopień uśrednienia współrzędnych atomów - nadmierne uśrednianie prowadzi do pogorszenia jakości modelu, więc modele w których odległości współrzędnych atomów od uśrednionych współrzędnych przekraczają konfigurowalny próg zostają odrzucone.

Można wyróżnić dwa etapy oceny modeli:

- podczas składania, nie wymaga znajomości składanego białka, służy do wybrania najlepszych modeli, które stanowić będą rozwiązanie. Preferowane są modele o większym pokryciu, złożone z możliwie małej liczby przypisań.
- po zakończeniu procesu składania - polega na porównaniu złożonych modeli z faktyczną strukturą białka.

Proces składania przebiega w dwóch etapach. W pierwszym, przypisania grupowane są w zbiory przypisań posiadających część wspólną, o określonej jakości. Pozwala to szybko odrzucić przypisania, które nie mogą zostać połączone z żadnym innym, przez co nie są przydatne w budowaniu modelu. Z każdej grupy budowana jest jedna struktura, co eliminuje redundancję i przyspiesza dalsze obliczenia zmniejszając liczbę dopasowywanych struktur. Drugi etap procesu składania polega na wyborze i połączeniu

podzbioru struktur utworzonych w pierwszym etapie w jeden model, pokrywający możliwie dużą część białka, przy zachowaniu określonych reguł łączenia struktur.

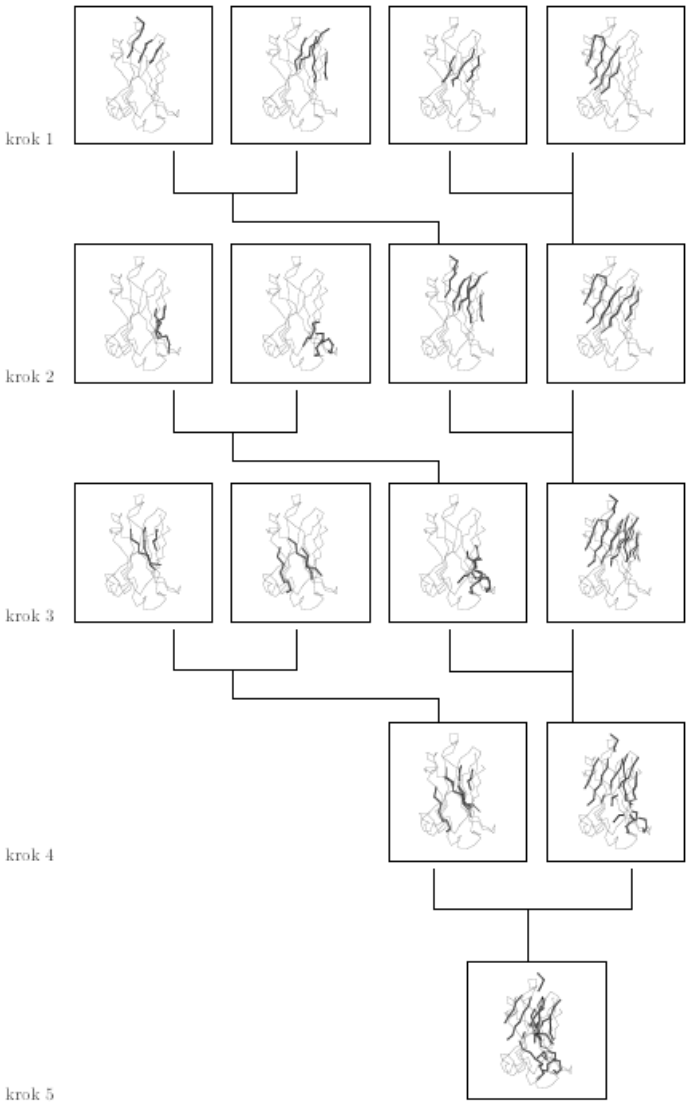


Rysunek 11 Kolejne etapy składania, metoda rekurencyjna.

Zaproponowano dwie metody składania:

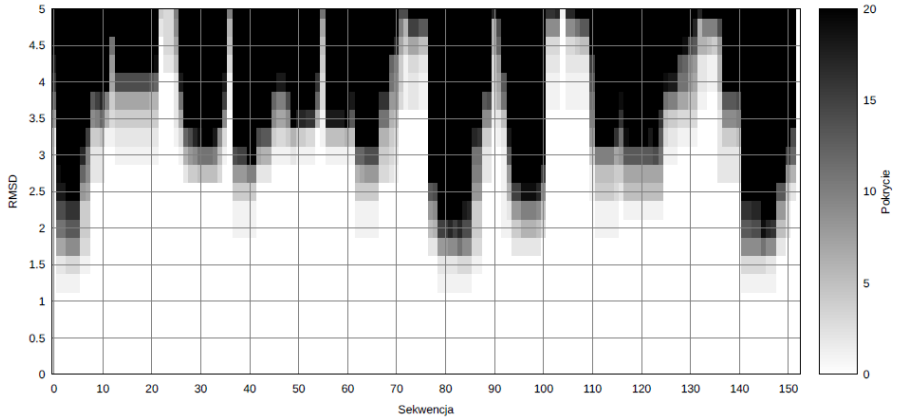
Pierwsza przetestowana metoda (Rysunek 11) polega na rekurencyjnej rozbudowie jednego modelu. Proces rozpoczyna się od wyboru przypisania startowego. Ze względu na duży wpływ tego wyboru i trudność w ocenie pojedynczego przypisania, algorytm wykonywany jest wielokrotnie, rozpoczynając od każdego przypisania.

Druga z metod rozbudowuje wiele modeli niezależnie (Rysunek 12). Dla zbioru struktur otrzymanych w pierwszym etapie, generowana jest lista wszystkich par struktur, jakie można połączyć zgodnie z parametrami konfiguracyjnymi. Każde połączenie jest wykonywane, tworząc nową strukturę. Jeżeli powstała struktura spełnia wymagania zapisane w parametrach konfiguracyjnych, jest ona dodawana do zbioru wejściowego. Dla dodanej struktury generowane są możliwe połączenia i dopisywane do listy możliwych połączeń. Proces powtarza się tak długo, jak długo możliwe jest wykonanie połączenia.



Rysunek 12 Kolejne etapy składania, metoda równoległa.

W kolejnej części eksperymentu badano wpływ jakości przypisań, wyrażonej w RMSD. W tym celu ograniczano zbiór przypisań do podzbioru przypisań o określonej jakości. Pokrycie białka przypisaniami przedstawiono na rysunku (Rysunek 13).

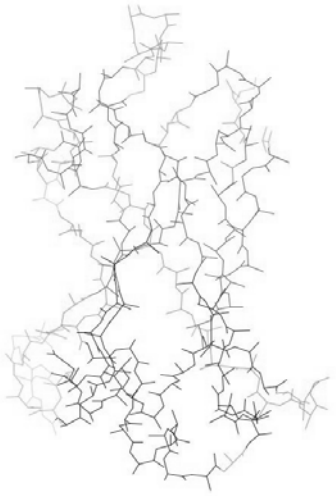


Rysunek 13 Pokrycie białka przypisaniami. Odcień szarości reprezentuje krotność pokrycia aminokwasu na osi odciętych przypisaniami o RMSD mniejszym od wartości na osi odciętych.

Dla zbiorów testowych o niskim RMSD, jakość modelu bliska jest jakości przypisań. Wraz ze spadkiem jakości przypisań znacząco spada jakość modelu. Wynikać to może z kumulujących się błędów dołączanych przypisań. Dołączenie jednego błędnego przypisania może pozwolić na dołączenie kolejnego lub uniemożliwić przyłączenie poprawnego przypisania.

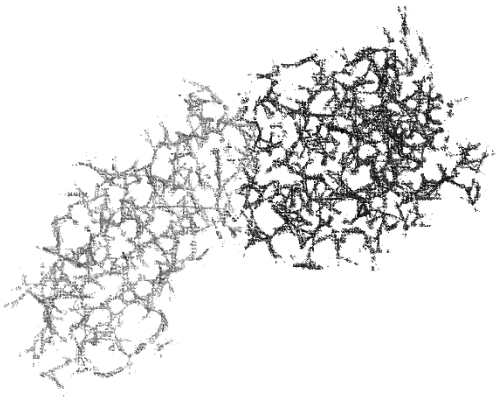
Jak można zauważyć, „wąskim gardłem" całej analizy jest wygenerowanie przypisań o odpowiedniej jakości. Zaproponowana metoda ze względu na efektywną strategię przeszukiwania całej przestrzeni jest w stanie zaproponować akceptowalne rozwiązanie nawet, gdy jakość składanych przypisań jest niższa od 3,5 Å (Rysunek 14).

Dla rekurencyjnej metody rozbudowy składania zwiększenie wartości parametru liczby zgodnych przypisań powoduje spadek pokrycia, co związane jest z mniejszą liczbą używanych przypisań. Jakość modelu wyrażona w RMSD przeważnie rośnie, jednak liczba generowanych modeli jest znacznie mniejsza.



Rysunek 14 Złożony model białka dla przypisań natywnych

Mając strukturę przestrzenną białka, istotnym problemem staje się identyfikacja (o ile taka sytuacja ma miejsce w danej strukturze) niezależnych, ale połączonych ze sobą podstruktur białka zwanych domenami (Rysunek 15).



Rysunek 15 Przykład białka dwudomenowego z wyszczególnieniem poszczególnych domen.

Idealny algorytm powinien dać odpowiedź na pytanie: jak podzielić daną strukturę białka na domeny, która to wiedza nie jest dostępna a priori. Na wejście metody rozpoznawania domen podawana jest struktura trzeciorzędowa, natomiast wyjściem jest jednoznaczne przypisanie poszczególnych aminokwasów do odpowiednich domen (Milostan i Lukasiak, 2016).

Zdefiniowanie domeny jako formalnej podstruktury jest trudne, aczkolwiek możliwe jest określenie pewnych podstawowych cech, które musi spełniać prawidłowa domena. Zgodnie z literaturą ((Holm i Sander 1994b) oraz (Xu i in. 2000)) domena powinna mieć następujące właściwości:

- powinna zawierać co najmniej 40 reszt (aminokwasów).
- domena musi być wystarczająco mała, aby spełnić następujący warunek ((Holm & Sander 1994a)):

$$\frac{\sum_{i,j} p_{i,j}}{n_a} \geq g_m$$

gdzie:

- *i* oraz *j* są dowolnymi dwoma atomami należącymi do aminokwasów rozdzielonych co najmniej trzema resztami w sekwencji, $p_{i,j} = 1$, jeśli odległość między *i* i *j* wynosi 4,0 Å lub mniej, w przeciwnym razie $p_{i,j} = 0$;
- n_a jest liczbą atomów w domenie;
- g_m jest pewnym progiem określonym na podstawie znanych domen (np. $g_m = 0{,}54$).

Interfejs między domenami musi być wystarczająco mały, aby liczba kontaktów wewnątrz domeny lub interakcji między aminokwasami tworzącymi domenę była znacznie większa niż liczba kontaktów między aminokwasami należącymi do różnych domen.

Liczba segmentów w domenie *D*, nie powinna być zbyt duża. Formalnie definiując, następujący warunek musi zostać spełniony:

$$\frac{r(D)}{s(D)} \geq l_s$$

gdzie $r(D)$ i $s(D)$ oznaczają odpowiednio liczbę aminokwasów i segmentów w domenie *D*; l_s jest pewnym progiem określonym na podstawie znanych domen (np. $l_s = 35$).

Algorytm rozpoznający domeny można zaproponować przyjmując następującą definicję kontaktu pomiędzy dwoma aminokwasami (Fidelis i in., 2004):

- każdy aminokwas jest reprezentowany jako dwa punkty w trójwymiarowej przestrzeni: jeden punkt ma współrzędne atomu C_α, a drugi jest geometrycznym środkiem łańcucha bocznego aminokwasu. Dla danej pozycji *i* w łańcuchu białkowym oznaczmy odpowiednio te punkty jako C_α^i i S^i
- aminokwasy na pozycjach *i* oraz *j* w danym łańcuchu białkowym są w kontakcie, jeśli którykolwiek z warunków jest spełniony:

$$|S^i - S^j| \leq r \text{ Å},$$

$$r\text{Å} < |S^i - S^j| < r + 1.5 \text{ Å i } |S^i - S^j| < |C_\alpha^i - C_\alpha^j| - 0{,}75 \text{ Å},$$

gdzie r jest parametrem metody – próg odcięcia (np. r=4,5 Å).

W celu rozwiązania tego problemu, zaproponowana została metoda iteracyjnego „kolorowania" struktury białka, a poprzez to identyfikacji domen, jako tej części białka, które została pokryta tym samym kolorem. Kolory powinny wskazywać kluczowe elementy domen białkowych i podawać szacunkową liczbę klastrów.. Ponadto możliwe jest włączenie informacji o strukturach drugorzędowych do strategii grupowania. Takie informacje mogą być generowane przy użyciu np. algorytmu DSSP (Kabsch i Sander 1983) czy LAD (Blazewicz i in, 2004).

Klastry D_i i D_j są uważane za sąsiadujące w grafie G, jeśli istnieje krawędź $e=\{v_i, v_j\}$ łącząca dwa wierzchołki $v_i \in D_i$ i $v_j \in D_j$.

Połączenie $links(D_i, D_j)$ oznaczmy jako sumę wag krawędzi w między węzłami w podgrafie D_i i podgrafie D_j. Innymi słowy:

$$links(D_i, D_j) = \sum_{i \in D_i, j \in D_j} w\{i,j\}$$

gdzie $w\{i, j\}$ – jest wagą krawędzi $e=\{i, j\}$.

Na koniec załóżmy, że w rozważanym grafie G nie ma żadnych pętli. Warunek ten jest wymuszany przez procedurę konstruowania grafu połączeń.

Otrzymane rdzenie muszą zostać połączone w większe obszary, aby skomponować domeny. Procedura przypisywania początkowych kolorów jest podatna na błędy i może przypisywać kolor myląco. Dlatego też wprowadzona została miara jakości klastra:

$$q_i = \frac{\sum_{j \neq i} links(D_i, D_j)}{links(D_i, D_i)}$$

Jeśli q_i jest większe od progu niezgodności to klaster i zostaje usunięty i wszystkie jego wierzchołki ponownie są „szare". W procedurze udoskonalania takie wierzchołki można przypisać ponownie do odpowiednich klastrów.

Dwa klastry D_i i D_j w grafie G, które w rzeczywistości są potencjalnymi fragmentami domen, można scalić, gdy spełnione są następujące warunki:

- D_i i D_j , $(i \neq j)$ są sąsiadującymi klastrami w grafie G,
- wyróżnik m zdefiniowany następująco:

$$m_{i.j} = \frac{links(D_i, D_j)}{links(D_i, D_i)}$$

jest większy niż pewien próg *t*, który jest parametrem metody.

Wartości wyróżnika $m_{i,j}$ dla całego grafu kontaktu można przedstawić jako macierz *M* wielkości $n \times n$, gdzie *n* jest liczbą skupień. Jeśli dla danego klastra istnieje więcej niż jeden klaster, który spełnia powyższe warunki, to klaster *i* jest łączony w klaster *j*, taki, że $m_{i,j} = max_j(m_{i,j})$. Po scaleniu klastrów zgodnie z macierzą *M* można przeliczyć współczynniki *m* dla mniejszej liczby klastrów. Ten etap można powtórzyć dostatecznie wiele razy dopóki wszystkie współczynniki *m* w macierzy *M* nie będą niższe niż próg podobieństwa.

Rysunek 16 Przykładowe przypisanie domen.

Dla większości sekwencji wyniki proponowanej metody są porównywalne z bazą danych domen białkowych SCOP pozwalając otrzymać rozwiązanie w skończonym i akceptowalnym czasie (Rysunek 16).

Algorytm zaproponowany powyżej ma złożoność wielomianową. Generowanie wykresów kontaktowych wymaga czasu $O(n^2)$, gdzie *n* jest długością sekwencji aminokwasów. Sortowanie wierzchołków ma złożoność $O(nlog(n))$, generowanie początkowych kolorów zajmuje $O(n^2)$ w najgorszym przypadku. Złożoność procedury scalania to $O(n^3)$ (dokładniej mówiąc, jest to $O(n^2m)$, gdzie *m* jest liczbą kolorów.) Liczba *m* nie jest większa niż *n/2*. Obliczenie wartości połączeń (D_i, D_j) zajmuje $O(n^2)$ dla wszystkich *i, j*. Obliczanie miar jakości q_i dla wszystkich *i* wymaga dodatkowego $O(m)$ czasu przy założeniu, że wartości połączeń (D_i, D_j) zostały obliczone. Współczynniki $m_{i,j}$ można obliczyć w czasie $O(m^2)$. Procedura scalania jest powtarzana nie więcej niż $(n/2)-1$ razy, dając ogólną złożoność $O(n^3)$.

Przewidywanie i modelowanie trzeciorzędowej struktury białek i RNA jest bardzo ważnym aspektem z punktu widzenia pełnionej przez nich funkcji. W związku z powyższym liczba metod obliczeniowych próbujących rozwiązać ten problem wzrasta w

bardzo szybkim tempie ze względu na postęp w rozwoju technologii maszyn obliczeniowych. Mając to na względzie, istotnym zagadnieniem stał się problem walidacji otrzymanych modeli. Liczba otrzymanych modeli przestrzennych wygenerowanych metodami obliczeniowymi wykazuje po weryfikacji znaczne odchylenia od struktury referencyjnej (struktury rzeczywistej), co powoduje, iż ewaluacja poprawności strukturalnej otrzymanego modelu staje się kluczowym aspektem pozwalającym na jego praktycznego wykorzystanie do projektowania leków (Kihara i in. 2009).

Zagadnienie ewaluacji struktur przestrzennych cząsteczek biologicznych jest rozpatrywane na dwóch poziomach w zależności od dostępnej informacji:

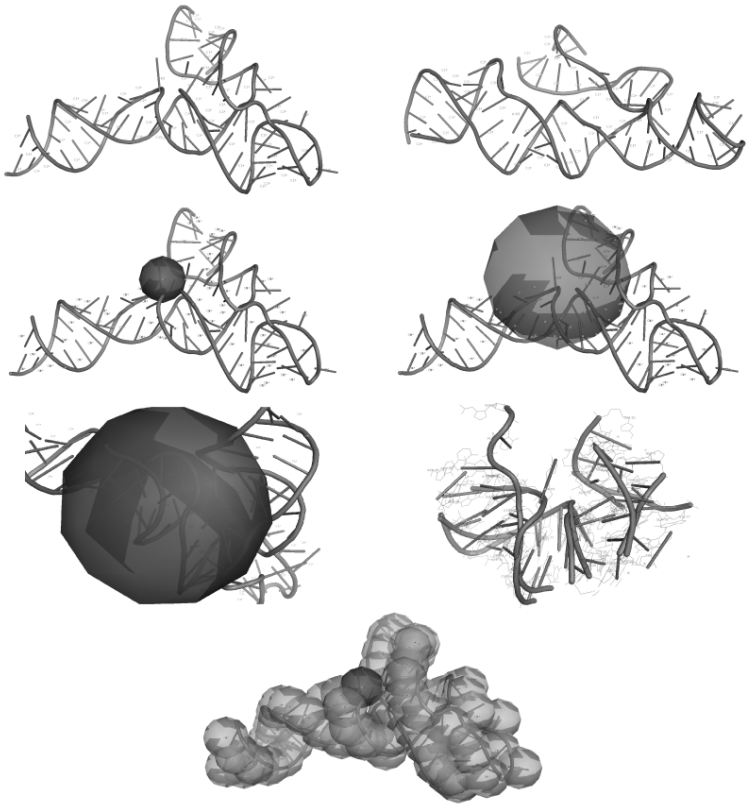
- ewaluacja struktur przestrzennych cząsteczek biologicznych w oparciu o strukturę referencyjną (analizowana jest różnica pomiędzy dopasowaniem strukturalnym modelu do struktury referencyjnej
- ewaluacja struktur przestrzennych cząsteczek biologicznych a priori, (struktura referencyjna jest nieznana).

Poza wspomnianym podziałem ze względu na dostępną informację, istnieje również podział ze względu na poziom szczegółowości:

- analiza globalna - wyznaczenie jednej mierzalnej wartości charakteryzującej poprawność strukturalną całego modelu przestrzennego,
- analiza lokalna – charakteryzująca poprawność strukturalną lokalnego otoczenia poszczególnych elementów łańcucha głównego.

Ocena przydatności poszczególnych modeli do praktycznego ich wykorzystania jest nadal otwartym problemem. Wykorzystanie danego modelu jest możliwe jeżeli jest on poprawny pod kątem stereochemicznym geometrycznym i topologicznym w odniesieniu do struktury uzyskanej eksperymentalnie.

Zaproponowana metoda oceny, bierze pod uwagę wiele poziomów szczegółowości definiowanych przez użytkownika. System dla białek o nazwie SphereGrinder (Lukasiak i in. 2015), zrealizowany przy współpracy z Protein Structure Prediction Center z Uniwersytetu Kalifornijskiego w Davis, został po raz pierwszy wykorzystany podczas CASP9. Został również stworzony system dedykowany strukturom RNA, o nazwie RNAlyzer (Lukasiak i in. 2013, Lukasiak i in, 2015).



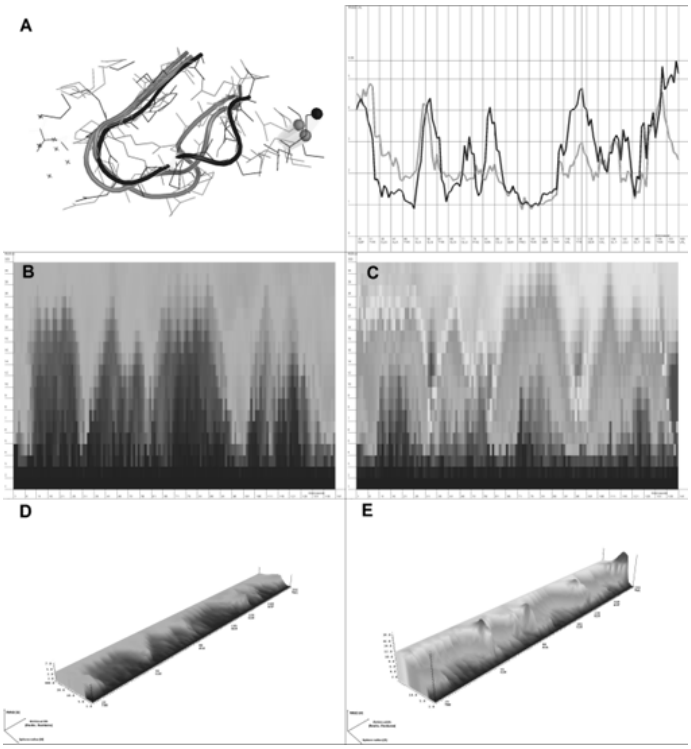
Rysunek 17 Proces porównywanie dwóch struktur - koncepcja sfer

Koncepcją znajdującą się u podstaw zaproponowanej metody jest idea analizy punktów przestrzeni za pomocą rosnących/malejących sfer.

Proponowane rozwiązanie uwzględnia wyznaczanie funkcji oceny w oparciu o wszystkie atomy zidentyfikowane w sferze (Rysunek 17).

Dla każdego aminokwasu/nukleotydu wzdłuż łańcucha głównego struktury referencyjnej oraz dla każdego zadanego promienia sfery budowana jest sfera. Dla każdej zbudowanej sfery określany jest zbiór atomów struktury referencyjnej, który został w niej zlokalizowany. W kolejnym kroku struktura modelu jest przeszukiwana w celu identyfikacji zbioru atomów w modelu, które odpowiadają jednoznacznie poszczególnym atomom danej sfery w strukturze referencyjnej. Odpowiadające sobie zbiory atomów skojarzone z daną sferą ze struktury referencyjnej oraz modelu, których liczność jest spójna, są optymalnie przestrzennie nakładane na siebie z wykorzystaniem technik Wolfganga Kabscha i Andrew McLachlana (McLachlan 1972; Kabsch 1976). Uzyskane w taki sposób dopasowanie strukturalne jest oceniane wybraną funkcją oceny.

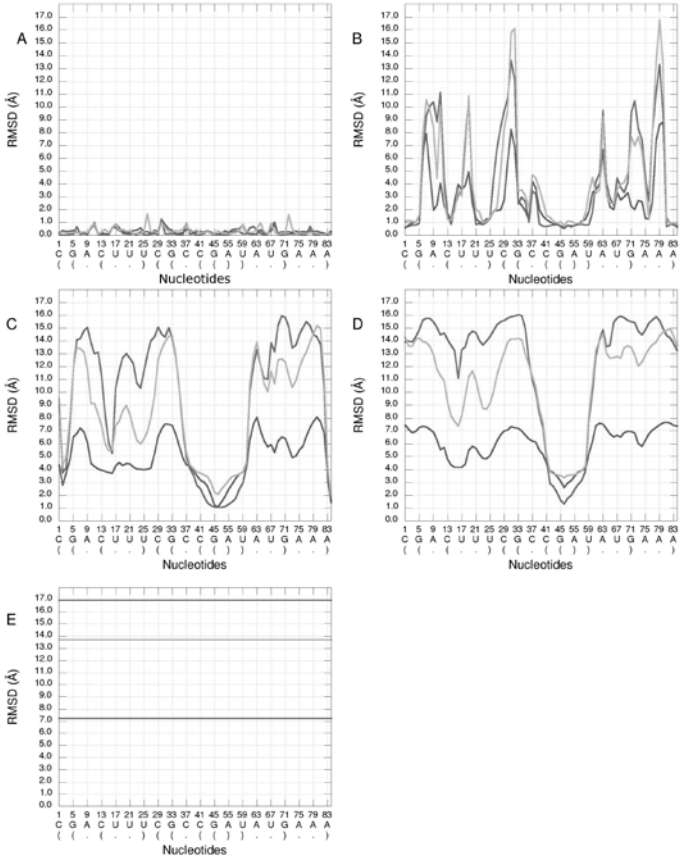
Zaproponowane rozwiązanie zostało dedykowane zarówno białkom (Rysunek 18), jak i RNA (Rysunek 19).



Rysunek 18 Przykładowa wizualizacja dla białek A – porównywane struktury (lewa strona), liniowy wykres odległości modeli od struktury referencyjnej wzdłuż całej sekwencji dla zadanego promienia sfery, B, C – odpowiednio mapa 2D i 3 D dla dwóch modeli (oś rzędnych - sekwencja aminokwasów, oś odciętych – rosnąca od dołu do góry wartość promienia sfery , im jaśniejszy odcień, tym większa wartość RMSD

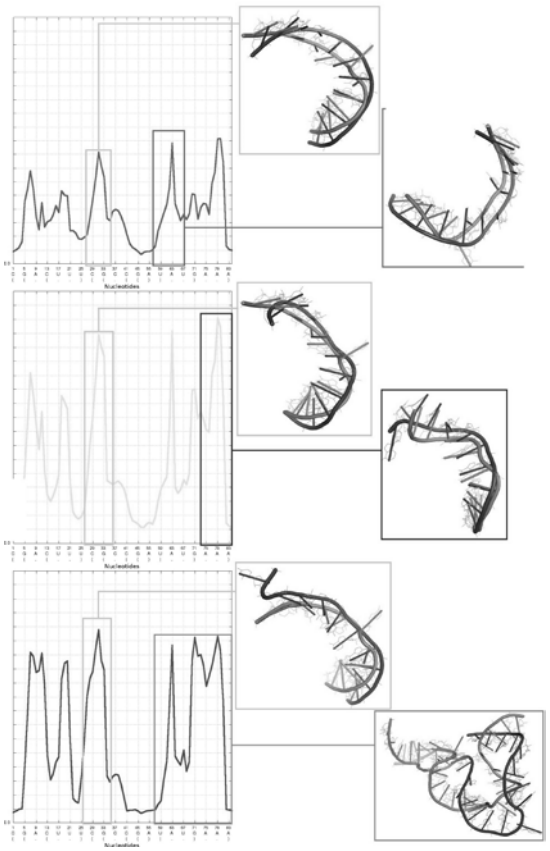
Proponowane analizy graficzne obejmowały:

Liniowy wykres integrujący wyniki wielu analizowanych modeli, gdzie każda krzywa opisuje wartości funkcji oceny skojarzone z dokładnie jednym analizowanym modelem (Rysunek 19). Wartość funkcji oceny na osi Y (obecnie RMSD) jest obliczana dla sfery o określonym promieniu zbudowanej wokół każdej reszty wzdłuż łańcucha głównego cząsteczki (oś X).



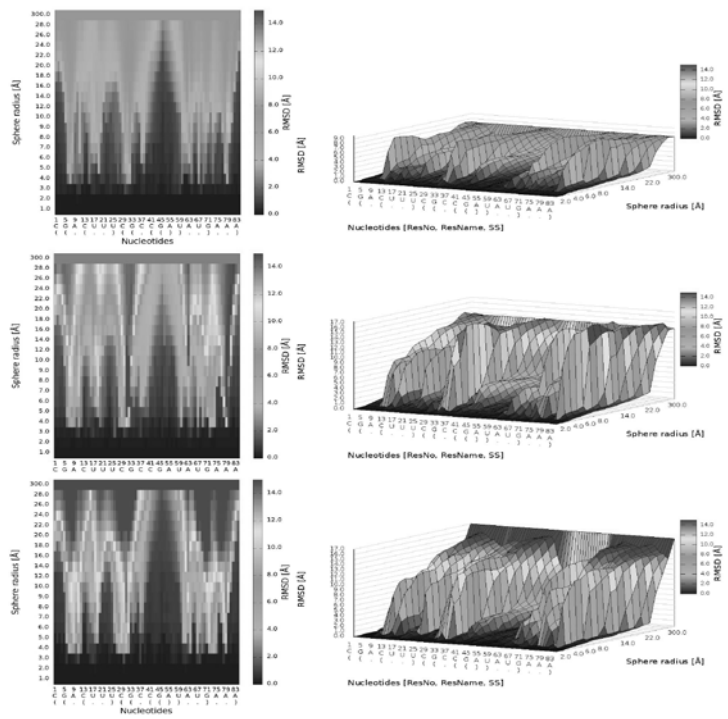
Rysunek 19 Przykładowa wizualizacja dla trzech modeli RNA. Każdy wykres (A-E) reprezentuje wyniki uzyskane dla innej wartości pro- mienia sfery (3, 8, 20, 38, and 300 Å). Oś X reprezentuje sekwencję nukleotydów, oś Y reprezentuje wartości RMSD.

Kolejny zestaw wykresów (Rysunek 20) dotyczy dokładnie tych samych modeli, przy czym na każdym wykresie prezentowany jest jeden model. Dla każdego modelu zostały zdefiniowane dwa przykładowe motywy strukturalne, które reprezentują znaczące nieprawidłowości.



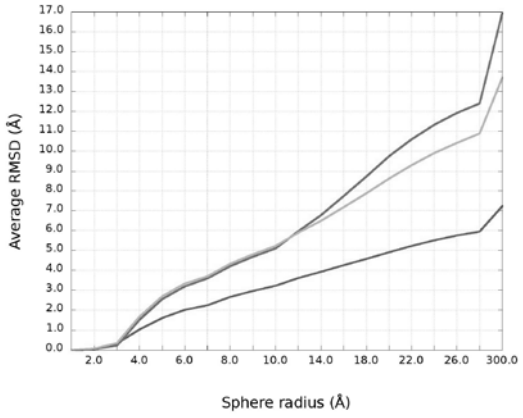
Rysunek 20 Przykładowa wizualizacja nieprawidłowych fragmentów strukturalnych dla trzech wybranych modeli RNA-Puzzles. Lewa strona – wizualizacja 2D, prawa strona – zniekształcone struktury.

Kolejne wykresy są wizualizacjami pozwalającymi na identyfikację nieprawidłowości strukturalnych odkrytych w pojedynczym modelu dla pełnego zestawu poziomów szczegółowości zdefiniowanych przez użytkownika (Rysunek 21). W przypadku obu wykresów oś X reprezentuje sekwencję reszt wzdłuż łańcucha głównego analizowanej cząsteczki, natomiast oś Y w obu wizualizacjach dotyczy wektora poziomów szczegółowości zdefiniowanych przez użytkownika.



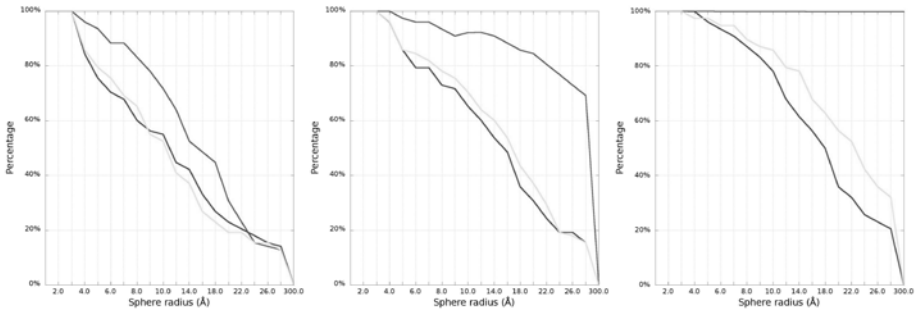
Rysunek 21 Przykładowa wizualizacja 2D i 3D dla trzech modeli RNA-Puzzles. Oś X - sekwencja nukleotydów, oś Y - promień sfery.

Kolejny wykres (Rysunek 22) prezentuje uśredniony współczynnik jakości wielu analizowanych modeli - każda krzywa opisuje uśrednioną wartość funkcji oceny dla dokładnie jednego modelu. Oś Y reprezentuje średnią miarę jakości spośród wszystkich reszt wzdłuż łańcucha głównego cząsteczki dla sfery o określonym promieniu.



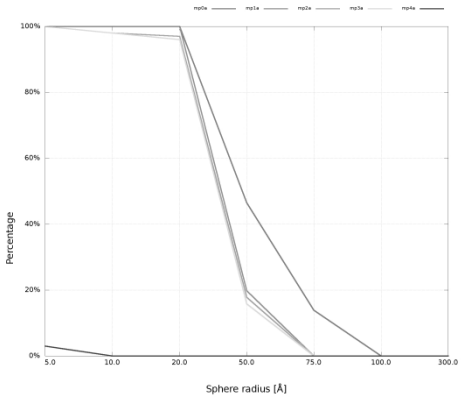
Rysunek 22 Przykładowa wizualizacja wykresu integrującego wyniki dla trzech wybranych modeli RNA-Puzzles, oś X - promienie sfer, oś Y uśrednione wartości RMSD dla sfer o stałym promieniu.

Następny wykres liniowy (Rysunek 23) prezentuje procent reszt rozważanej cząsteczki charakteryzujących się strukturalnie prawidłowym lokalnym otoczeniem przy założonym odcięciu.



Rysunek 23 Wizualizacja wykresu integrującego wyniki dla trzech modeli z RNA-Puzzles. Wykres przedstawia procent zbiorów atomów w każdym z modeli, bliższych strukturze referencyjnej niż zadany próg odcięcia (4, 7, 10 Å) dla sfer o stałym promieniu 24 Å.

Prezentowana metoda analizy pozwala rozpoznawać fragmenty w potencjalnie prawidłowych w modelach, które globalnie zostały zaklasyfikowane jako zupełnie nieprawidłowe.



Rysunek 24 Przykładowe wyniki miary SG dla wybranych pięciu modeli przy odcięciu na poziomie 2 Å

Razem z graficznymi metodami wizualizacji zaproponowana została metoda oceny struktur przestrzennych w odniesieniu do struktury referencyjnej jako procent sfer o zadanym promieniu (6 Å dla białka, 20 Å dla RNA) utworzonych na każdym aminokwasie (nukleotydzie) sekwencji, w których liczba atomów odpowiada liczbie atomów w odpowiadających im sferach struktury referencyjnej. Innymi słowy, zakładając, że mamy dany zbiór sfer o zadanym promieniu utworzonych na każdym elemencie sekwencji struktury referencyjnej $R=\{R_1, ..., R_N\}$, gdzie N jest długością sekwencji, oraz zbiór sfer o zadanym promieniu utworzonych na każdym elemencie sekwencji modelu $M=\{M_1, ..., M_N\}$, gdzie N jest długością sekwencji i skojarzone z nimi zbiory $L_R=\{r_a,...,r_N\}$, gdzie r_N jest liczbą

atomów w sferze R_N oraz $L_M=\{m_a,...,m_N\}$, gdzie m_N jest liczbą atomów w sferze M_N, wartość miary SG wynosi

$$SG = \frac{\sum_{i=1}^{N} x_i}{N} * 100\% \quad \text{gdzie } x_i=1 \text{ jeżeli } (m_i/r_i)=1,\ x_i=0 \text{ jeżeli } (m_i/r_i)\neq 1$$

Miara ta została włączona w zbiór miar stosowanych do ewaluacji struktur przestrzennych białek w CASP.

Przykładowe wyniki tej miary dla odcięcia na poziomie 2 Å (Rysunek 24).

Ewaluacja prawidłowości strukturalnej białek bez znajomości struktury referencyjnej, czyli wzorca umożliwiającego porównanie jest zagadnieniem trudniejszym od problemu oceny jakości ze strukturą referencyjną. Do rozwiązania tego problemu zaproponowano dwie strategie:

- ocenę jakości/prawidłowości w oparciu o pojedynczy model
- ocena oparta na konsensusie.

Zaprezentowana nowa metoda lokalnej/globalnej ewaluacji struktur białkowych bez potrzeby posiadania dodatkowych informacji pozwala rozpoznać nieprawidłowości strukturalne wynikające zarówno z niedoskonałości w łańcuchu głównym jak i łańcuchach bocznych w oparciu o podejście deskryptorowe.

Proponowana metoda ewaluacji wiarygodności jest następująca:

- dla każdego aminokwasu wzdłuż łańcucha polipeptydowego ewaluowanego modelu przestrzennego budowany jest deskryptor, który jest pod dawany ocenie z wykorzystaniem molekularnej funkcji oceny wybranej przez użytkownika (np. DFIRE).
- dla każdego deskryptora zbudowanego w analizowanym modelu przeszukiwana jest biblioteka grup deskryptorowych w celu znalezienia najbardziej zbliżonej strukturalnie grupy.
- dla wszystkich deskryptorów z modelu, dla których udało się dopasować grupę deskryptorową jakość oceny jest wyliczana na podstawie wskaźnika odchylenia jego wartości oceny wyznaczonej dla aktualnie wykorzystywanej molekularnej funkcji oceny (np. DFIRE) względem standardowego rozkładu normalnego.
- Wiarygodność lokalnego otoczenia strukturalnego aminokwasów w modelu jest obliczana jako średnia współczynników wiarygodności wszystkich deskryptorów, w których dany aminokwas został zidentyfikowany.

Globalna miara wiarygodności modelu jest obliczana jako średnia ocen poszczególnych aminokwasów, dla których udało się uzyskać lokalną miarę wiarygodności skalowaną do rozmiaru analizowanej cząsteczki białkowej (rozumianej jako liczba aminokwasów).

Miary, które najlepiej poradziły sobie z odtwarzaniem miary wzorcowej GDT_TS to liczba dopasowanych deskryptorów w modelu, pokrycie sekwencyjne dopasowanych deskryptorów w modelu oraz metoda ewaluacji bazująca jedynie na dopasowaniu sekwencyjnym poszczególnych aminokwasów w ramach strukturalnie podobnej grupy i zbiorze cech strukturalnych białek. W przypadku metody wykorzystującej wybrane molekularne funkcje oceny, a w szczególności potencjał elektrostatyczny (ECSTQA), Lennarda-Jonesa (LJQA) oraz statystyczny (DFIREQA) można zauważyć, że najlepiej sprawdza się miara wiarygodności bazująca na potencjale statystycznym.

Podsumowanie

Informatyka jest dziedziną nauki, która ze względu na swój potencjał zaczyna być obecna we wszystkich dziedzinach naszego życia. Połączenie mocy obliczeniowych wraz z efektywnymi metodami badań operacyjnych pozwala przeprowadzić procesy analizy danych na niespotykaną do tej pory skalę. Wzrost ilości generowanych danych powoduje wręcz niemożność wyekstrahowania najbardziej istotnych danych. Jednakże sama analiza musi zostać przedstawiona w formie akceptowalnej i zrozumiałej przez człowieka co wymusza wprowadzania razem z metodami obliczeniowymi również opracowania schematów analizy na poziomie graficznym. Wspomniany rozwój narzędzi i maszyn obliczeniowych wpłynął bezpośrednio również na inne nauki przyrodnicze takie jak np. biologia. Rozwój eksperymentalnych technologii spowodował wzrost ilości dostępnych danych o życiu, o zasadach funkcjonowania różnorodnych organizmów nas otaczających, jednakże nie był równoznaczny z proporcjonalnym wzrostem ilości dostępnej wiedzy. Podstawowym problemem, który się pojawił była możliwość przetworzenia tych danych w sposób umożliwiający wyodrębnienie istotnych informacji z powstałego szumu informacyjnego.

W celu rozwiązania powyższych zagadnień zrealizowano następujące wątki badawcze:

- Zaproponowana została metoda oparta na heurystyce Tabu Search, która przy przyjętych założeniach potrafi znaleźć rozwiązanie struktury białka w zdyskretyzowanej kwadratowej i trójkątnej przestrzeni 2D i 3D. Otrzymane rozwiązanie może być potraktowane jako początkowe rozwiązanie dla modelowania w rzeczywistej przestrzeni.

- Stosując metodę dekompozycji białka, w oparciu o paradygmat deskryptorów, zaproponowano metodę stworzenie nowego alfabetu struktur białkowych o struktur RNA. Alfabet złożony jest z podstruktur zawierających informację nie tylko o kolejności aminokwasów czy też nukleotydów w sekwencji, ale również o zależnościach przestrzennych tych atomów w strukturze, które są w sąsiedztwie przestrzennym względem siebie nie będąc sąsiadami w sekwencji.
- Zaproponowane zostały nowe rozwiązania algorytmiczne umożliwiające porównywanie deskryptorów ze sobą w sposób detereministyczny wraz z metodą ich ustandaryzowania pomimo początkowej różnicy w ich wielkości (liczności atomów wchodzących w skład danego deskryptora. Umożliwiło to zaproponowanie dwóch baz danych bibliotek deskryptorowych dla sekwencji białkowych i nukleotydowych. Na podstawie stworzonych bibliotek zaproponowana została strategia rekonstrukcji łańcucha głównego białka jako wzorca mogącego posłużyć do modelowania przestrzennego kształtu danego białka.
- Zaproponowany został algorytm identyfikacji domen oparty o metodę gęstościową w oparciu o nadrzędną zasadę, iż aminokwasy należące do danej domeny są zlokalizowane bliżej siebie niż aminokwasy należące do domeny drugiej (założenie to jest zbieżne z podstawową zasadą grupowania przy algorytmach uczenia bez nauczyciela).
- Przedstawiona została nowa metoda pozwalająca na ewaluację struktur przestrzennych cząsteczek biologicznych w oparciu o strukturę referencyjną oraz wektor poziomów szczegółowości analizy definiowany przez eksperta. Opracowana metoda, udostępniając zestaw dwu/trójwymiarowych wizualizacji, pozwala zidentyfikować z jednej strony nieprawidłowe motywy strukturalne, które powinny zostać poddane szerszej analizie i procesowi udoskonalania, z drugiej strony prawidłowe motywy strukturalne nawet w modelach, które globalnie zostały zakwalifikowane jako bardzo odległe strukturalnie od struktury referencyjnej.
- Przedstawiona została również nowa metoda obliczeniowa pozwalająca na ewaluację prawidłowości strukturalnej modeli przestrzennych białek bez potrzeby posiadania struktury referencyjnej, na dwóch poziomach szczegółowości analizy (lokalnym i globalnym).

W wyniku przeprowadzonych badań i dokonanych analiz wykazano, iż rozważany zbiór zagadnień może być rozwiązany w sposób efektywne obliczeniowo pozwalając wyodrębnić istotne informacje dla analizy związanej ze strukturami przestrzennymi cząsteczek biologicznych. Biorąc pod uwagę praktyczny cel zdefiniowany na wstępie można uznać, iż zaproponowane rozwiązania obliczeniowe mogą być z powodzeniem zastosowane w praktyce będąc kolejnym krokiem ku stworzeniu medycyny spersonalizowanej.

Anfinsen C.B., Principles that govern the folding of protein chains, Science, 1973, 4096, 181, 223–230

Antczak M., Kasprzak M., Lukasiak P., Blazewicz J., Structural alignment of protein descriptors - a combinatorial model, BMC Bioinformatics, 2016, 17, 383

Beutler, T. & Dill, K., A fast conformational search strategy for finding low energy structures of model proteins, Protein Science, 1996, 5, 2037–2043.

Bjorkholm P., Daniluk P., Kryshtafovych A., Fidelis K., Andersson R., Hvidsten T.R., Using multi-data hidden Markov models trained on local neighborhoodsof protein structure to predict residue-residue contacts. Bioinformatics, 2009, 25, 1264–1270.

Brenner S. E., Koehl P., Levitt M., The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res, 2000, 28, 254–256.

Blazewicz J., Dill K., Lukasiak P., Milostan M., A tabu search strategy for finding low energy structures of proteins in HP-model Computational Methods in Science and Technology, 2004, 10, 7-19.

Blazewicz J., Lukasiak P., Milostan M., Application of tabu search strategy for finding low energy structure of protein, Artificial Intelligence in Medicine, 2005, 35, 135-145.

Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A., Yannakakis, M., On the complexity of protein folding, Proc. 1998 STOC, and J. of Computational Biology, 1998

Dill K., Polymer principles and protein folding', Protein Sci, 1999, 8, 1166–1180.

Glover F., Laguna M., Tabu Search, Kluwer Academic Publishers, 1997, Boston, USA.

Hart W., Istrail S., Fast protein folding in the hydrophobic-hydrophilic model within three-eights of optimal, Journal of Computational Biology, 1996, 3(1), 53–96.

Hein MY, Sharma K, Cox J, Mann M, Handbook of Systems Biology, 2013, San Diego: Academic Press

Hvidsten T.R., Kryshtafovych A., Komorowski J., Fidelis K., A novel approach to fold recognition using sequence-derived properties- from sets of structurally similar local fragments of proteins. Bioinformatics, 2003, 19 S2, ii81–ii91.

Lesk A., Introduction to bioinformatics. Oxford University Press, Oxford New York, 2013.

Lukasiak P., The novel approach for building models of protein structure, European Chapter on Combinatorial Optimization (ECCO) XXII, April 26-28, 2012 Antalya, Turcja

Lukasiak P., Quality assessment methodologies in analysis of structural models, 25th European Conference on Operational Research, EURO XXV 8-11 July 2012, Wilno, Litwa, 80

Lukasiak P., Antczak M., Ratajczak T., Blazewicz J., Quality evaluation of biomolecules, EURO XXVI 1-4 July 2013b, Rzym, Włochy

Lukasiak P., Antczak M., Ratajczak T., Blazewicz J., SphereGrinder - reference structure-based tool for quality assessment of protein structural models, (Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Washington D.C., USA, 9.11-12.11.2015), 2015, 665-668

Lukasiak P., Antczak M., Ratajczak T., Szachniuk M., Popenda M., Adamiak R.W., Blazewicz J., RNAssess - a web server for quality assessment of RNA 3D structures, Nucleic Acids Research, 2015, 43(W1), W502-506

Lukasiak P., Computational methods for 3D structure analysis, ECCO 2017: 30th Conference of the European Chapter on Combinatorial Optimization, Koper, Slovenia 03-06.05.2017

Milostan M., Lukasiak P., Domgen-Graph based method for protein domain delineation, RAIRO - Operations Research, 2016, 50, 2, 363-374

Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A., Critical assessment of methods of protein structure prediction (CASP)--round x. Proteins, 2013; 82 S2(0 2):1–6

Murzin A. G., Brenner S. E., Hubbard T. and Chothia C., SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol, 1995, 247, 536–540.

O'Toole, E. & Panagiotopoulos, A., Monte carlo simulation of folding transitions of simple model proteins using a chain growth algorithm, J. Chem. Phys., 1992, 97, 8644–8652.

Ridley M., Genome: The Autobiography of a Species in 23 Chapters, 1999, New York: Harper Collins Publishers

Shannon C.E., A Mathematical Theory of Communication, Bell System Technical Journal, 1948, 27, 379–423 & 623–656, July & October

Strombergsson H., Daniluk P., Kryshtafovych A., Fidelis K., Wikberg JE., Kleywegt GJ., Hvidsten T.R., Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space. J Chem Inf Model, 2008, 48, 2278–2288.

Toma, L., Toma, S., Contact interactions method: A new algorithm for protein folding simulations, Protein Sci., 1996 5, 147–153.

Unger R., Moult J., 'Genetic algorithms for protein folding simulations, Journal of Molecular Biology, 1993, 231, 75–81.

Yue K., Dill, K., Sequence-structure relationships in proteins and copolymers, Physical Review, 1993, 48(3), 2267–2278.

Yue K., Fiebig K., Thomas P., Chan H., Shakhnovich E., Dill K., A test of lattice protein folding algorithms, Proc. Natl. Acad. Sci. USA, 1995, 92, 325–329.

Watson J. D. and Crick F. H., Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature, 1953 171, 4356.

5. Omówienie pozostałych osiągnięć naukowo - badawczych

Elementem związanym z inną działalnością naukowo badawczą są działania w zakresie stworzenia i funkcjonowania Europejskiego Centrum Bioinformatyki i Genomiki. Europejskie Centrum Bioinformatyki i Genomiki (ECBiG) to unikatowy, interdyscyplinarny ośrodek badawczy. Łącząc nowatorskie zaawansowane techniki eksperymentalne z analizami obliczeniowymi i modelowaniem, ECBiG umożliwia integrację danych z różnych poziomów informacji genetycznej, ekspresji i układów biologicznych o różnych poziomach złożoności. Celem ECBiG jest prowadzenie badań w obszarze bioinformatyki i genomiki funkcjonalnej i strukturalnej, jak również opracowanie i wdrożenie innowacyjnych metod badawczych, narzędzi obliczeniowych i baz danych. W działalności tego centrum uczestniczyłem od początku biorąc aktywny i znaczący udział w procesie jego tworzenia. Jestem członkiem Komitetu Sterującego tego Centrum oraz Kierownikiem Laboratorium Bioinformatyki Strukturalnej i Systemowej. Koordynowałem i zarządzałem grantem „Rozwój ECBiG" (kierownik prof. dr hab. inż. Jacek Błażewicz), jak również jestem współautorem wniosku wprowadzającego ECBiG na

Polską Mapę Drogową Infrastruktury Badawczej. Obecnie w ramach Centrum zajmuję się kordynacją prac związanych z realizacją grantu z PMDIB mającego na celu opracowaniem Genomicznej Mapy Polski oraz wdrożeniem komercyjnym rezultatów tego projektu.

Zamieszczony powyżej watek badawczy był realizowany równolegle z innymi działaniami naukowo-badawczymi, z których najważniejsze są wyszególnione poniżej:

1. Wątkiem badawczym realizowanym obecnie jest stworzenie Genomicznej Mapy Polski, związanym z weryfikacją hipotezy o stopniu zróżnicowania osób zamieszkujących obszar Polski. Projekt (POIR.04.02.00-30-A004/16) jest realizowany w ramach Programu Operacyjnego Inteligentny Rozwój 2014–2020 Priorytet IV: Zwiększenie Potencjału Naukowo-Badawczego, Działanie 4.2: Rozwój Nowoczesnej Infrastruktury Badawczej Sektora Nauki. Genomiczna Mapa Polski (GMP) oparta wstępnie na reprezentatywnej grupie 5000 genomów stanie się bazą referencyjną dla wszystkich projektów biologicznych i bioinformatycznych realizowanych w dziedzinie analiz genetycznych i genomicznych. Surowe dane genetyczne pełnych sekwencji genomu, a w szczególności interpretacje powiązań genetycznych w jednostkach chorobowych, uzyskane poprzez Genomiczną Mapę Polski (GMP), stanowią źródło do formułowania wielu kolejnych problemów badawczych takich jak metody przetwarzania i analizy GMP czy też tworzenie algorytmicznych rozwiązań do strukturalnej i funkcjonalnej analizy genomów są nietrywialnym wyzwaniem. Głównym rezultatem algorytmicznym będzie prototyp przyjaznego użytkownikowi narzędzia służącego analizie i agregacji zgromadzonej informacji. Integralną jego częścią będzie dwu- i trójwymiarowa wizualizacja uzyskanych wyników obliczeń. W badaniach wykorzystane zostaną m.in. algorytmy klastrowania, które z kolei bazować będą na uzyskanej wcześniej informacji charakterystycznej (np. motywy). Narzędzie integrować będzie algorytmy rozwiązujące wybrane problemy, dostosowane do przetwarzania danych w skali genomowej i efektywne w sensie jakości i czasu.

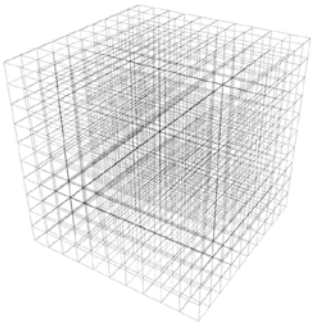
2. W latach 2005-2009 podczas prac w grancie COMPUVAC finansowanym ze środków 6-go Programu ramowego Komisji Europejskiej jako członek Komitetu Sterującego całego konsorcjum zajmowałem się prowadzeniem badań w zakresie stworzenia systemu wspomagania decyzji wytwarzania szczepionek genetycznych. Celem badań było zweryfikowanie hipotezy o możliwości zaproponowania rozwiązań

algorytmicznych umożliwiających weryfikację jakości szczepionki w oparciu dane eksperymentalne przedstawione w postaci rozproszonej, w której eksperymenty były realizowane w różnych warunkach i w różnych miejscach z użyciem różnych odczynników. W wyniku wielu staży naukowych niezbędnych do pozyskania wiedzy dziedzinowej i przeprowadzonych eksperymentów obliczeniowych został opracowany prototyp takiego systemu standaryzujący cały proces wytwarzania szczepionek tego typu o strukturze modułowej, gdzie każdy moduł stanowi istotną część analizy odpowiedzi immunologicznej układu odpornościowego. Zaproponowana została nowa hierarchia wektorów umożliwiająca szybkie przeszukiwanie bazy danych. Kolejnym elementem było zaproponowanie procedury standaryzacji protokołu immunizacji użytego do badań. Wyekstrahowano również zbiór kilku tysięcy reguł umożliwiający weryfikację wyników eksperymentów dotyczących odporności komórkowej związanej z aktywnością limfocytów T pod kątem wiarygodności i staranności. Zaproponowano procedury efektywnego przetwarzania i analizy odpowiedzi humoralnej, jak również sygnatury molekularnej. Opracowany został proces walidacji wszystkich danych wraz z zaproponowaną procedurą oceny jakości analizowanych szczepionek. Istotnym elementem była również opracowana procedura porównywania eksperymentów realizowanych w różnych laboratoriach opartą o ideę złotego standardu referencyjnego eksperymentu odniesienia. Wykonane badania potwierdziły prawdziwość wstępnych założeń.

Blazewicz J., Borowski M., Chaara W., Kedziora P., Klatzmann D., **Lukasiak P.**, Six A., Wojciechowski P., GeVaDSs - decision support system for novel Genetic Vaccine development process, 2012, BMC Bioinformatics, 13:91
Blazewicz J., Lukasiak P., System wspomagania wytwarzania i analizy szczepionek genetycznych, Kosmos, 2009, 58 (1-2), 113-126

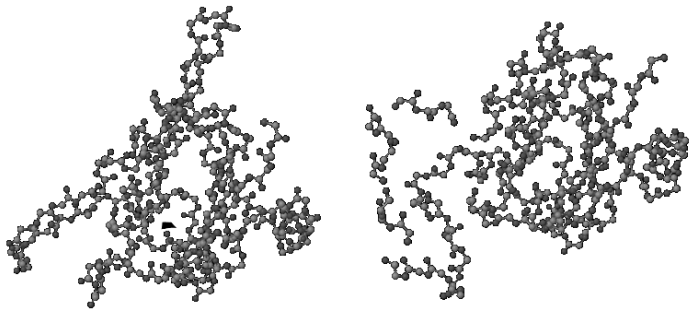
3. Kolejnym wątkiem badawczym było umożliwienie wizualnej analizy struktur białek poprzez analizę gęstości występowania atomów w grupach struktur i różnic w gęstości między grupami w przestrzeni dyskretnej. Podczas badań weryfikowano hipotezę czy gęstość rozmieszczenia atomów w strukturze przestrzennej jest skorelowana z miarą GDT_TS będącą miarą podobieństwa cieszącą się wśród bioinformatyków najwyższym stopniem akceptacji. W celu analizy gęstości

występowania atomów, przestrzeń, w którą wpisane są wybrane struktury, jest dzielona na przylegające do siebie sześciany (Rysunek 25). Zbiór sześcianów dzielących przestrzeń tworzy bryłę o kształcie prostopadłościanu, która w dalszej części pracy nazywana będzie siatką gęstości (ang. *density grid*, w skrócie DG).

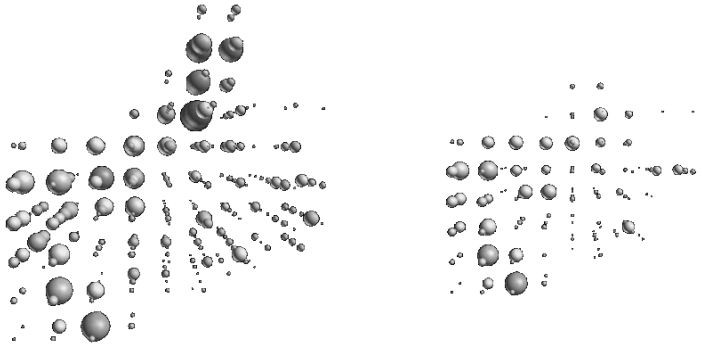


Rysunek 25 Przykład przestrzeni podzielonej przez siatkę gęstości

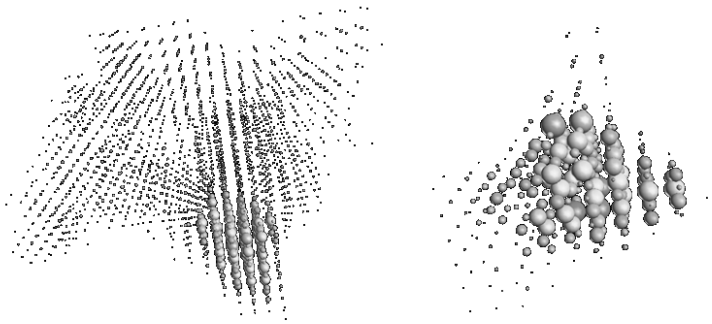
Rozkład gęstości występowania atomów tworzy się przez nałożenie wszystkich atomów wybranych struktur (Rysunek 26) na DG. Współczynnik wynikowy gęstości w każdym z sześcianów jest równy proporcji liczby atomów, których jądro przecinało się z danym sześcianem do liczby struktur użytych do stworzenia siatki gęstości (Rysunek 27).



Rysunek 26 Struktury referencyjna (lewa strona) i struktura porównywana (prawa strona)



Rysunek 27 Wynik całkowity porównania struktur (lewa stona) i z wyizolowanymi obszarami o większej gęstości struktury referencyjnej (prawa strona)



Rysunek 28 Wynik porównania dwóch struktur referencyjnych z CASP (lewa strona brak homologii, prawa strona struktura homologiczna

W celu ułatwienia wizualnej analizy modeli białek zaproponowano algorytm, umożliwiający analizę gęstości występowania atomów w strukturach oraz grupach struktur.

Różnice między modelami są silnie skorelowane z wartościami GDT_TS porównywanych struktur. Największą zaletą porównywania modeli za pomocą siatek gęstości, jest możliwość wyizolowania i wizualnego zbadania interesujących obszarów porównywanych modeli. Płynne określanie odcięcia wartości współczynnika gęstości w wyświetlanych na ekranie sferach, pozwala na precyzyjne wyodrębnienie interesujących badacza obszarów struktur referencyjnej i porównywanej.

Uzyskane wyniki, wskazują, że siatki budowane z modeli dobrej jakości posiadają wyróżniający się obszar „rdzenia" siatki. Wartości współczynników gęstości w takich DG są wyższe niż w siatkach gęstości tworzonych z modeli słabej jakości.

Spowodowane jest to zdecydowanie większym podobieństwem modeli dobrej jakości, przez co ich atomy nie są „rozprowadzane" na dużym obszarze siatki gęstości, a koncentrują się w centrum siatki (Rysunek 28). Uzyskane wyniki potwierdziły wstępne założenia, przez co zaproponowane rozwiązanie może być stosowane zamiennie do wspomnianej miary.

Lukasiak P., Data analysis in bioinformatics, XXIX EURO: 29th European Conference on Operational Research, Valencia, Spain 08-11.07.2018

4. Prace nad rankingiem modeli przestrzennych struktur białek spowodowały opracowanie metody 3D-Judge. Podczas badań weryfikowano hipotezę o możliwości oceny jakości modeli struktur przestrzennych za pomocą danych historycznych. Algorytm 3D-Judge działa na zbiorze n serwerów. Od każdego serwera otrzymuje na wejście dokładnie po jednym modelu uznanym przez dany serwer za najlepszy. Spośród otrzymanych na wejście modeli algorytm wybiera jeden najlepszy (jest zatem selektorem). Wybór najlepszego rozwiązania dokonywany jest na podstawie następujących informacji:
   - sekwencja s dla której struktura przestrzenna jest nieznana
   - dane historyczne H

   W celu formalnego przedstawienia algorytmu wprowadzone zostaną następujące oznaczenia:

   - P : zbiór serwerów
   - $M_p(s)$: najlepszy model wygenerowany przez serwer p dla sekwencji s
   - $sim(M_{p_1}(s), M_{p_2}(s))$: ocena podobieństwa modeli $M_{p_1}(s)$ i $M_{p_2}(s)$
   - $M_{real}(s)$: model ze znaną strukturą dla sekwencji s
   - $sim(M_p(s), M_{real}(s))$: ocena modelu wygenerowanego przez serwer p
   - H jest jest zbiorem sekwencji s dla których model przestrzenny jest znany ($M_{real}(s)$) i dla których wszystkie serwery należące do P ($M_p(s)$) wyznaczyły najlepsze według ich wewnętrznych ocen modele

   3D-Judge używa sztucznej sieci neuronowej(SSN) w celu wybrania najlepszego modelu spośród modeli otrzymanych z serwerów. SSN posiada $(|P|^2/2)-|P|$ neuronów

w warstwie wejściowej. Każdy neuron jest przypisany do odpowiedniej miary podobieństwa $sim(M_{p_1}(s), M_{p_2}(s))$. Wynikiem 3D-Judge jest model, dla którego odpowiadający mu neuron warstwy wyjściowej posiada najwyższą wartość. Istnieje też ukryta warstwa neuronów, której rozmiar jest parametrem metody.

W celu wybrania najlepszego modelu z dostępnych, SSN uczona jest na podstawie danych historycznych *H*. Dla każdej sekwencji *s* wszystkie wyniki porównań podobieństwa pomiędzy modelami oraz wyniki porównań podobieństwa do rzeczywistej struktury są znane. Wszystkim neuronom warstwy wyjściowej przypisuje się wyniki porównania podobieństwa badanych modeli do rzeczywistej struktury. Oznacza to, że SSN jest uczona jak oceniać modele na podstawie podobieństw między nimi. Analizy mogą być prowadzone tylko dla serwerów dla których istnieją dane historyczne. Otrzymane wyniki potwierdziły wstępne założenia.

Jaskowski W., Blazewicz J., **Lukasiak P.**, Milostan M., Krasnogor N., 3D-Judge - A metaserver approach to protein structure prediction, Foundations of Computing and Decision Sciences 2007, 32, 3-14

5. Celem badań było wyznaczenie funkcjonalnych różnic pomiędzy różnymi białkami typu Dicer bazując na trójwymiarowych modelach struktur. Dane wejściowe dla procesu badawczego to istniejąca znana struktura białka typu Dicer z organizmu *Giardia Intestinalls* oraz sekwencje nieznanych białek typu Dicer z organizmu *Arabidopsis Thaliana* i innych organizmów, bazy podziałów domenowych białek, zbiór wzorców strukturalnych dla zidentyfikowanych domen konserwatywnych. Zaproponowane podejście bazuje na paradygmacie modelowania homologicznego. Po uzyskaniu modeli strukturalnych białek przeprowadzony został eksperyment dokowania rzeczywistych cząsteczek RNA w celu potwierdzenia, że uzyskane modele na poziomie strukturalnym spełniają postawione im cele funkcjonalne. Ostatecznym efektem badań jest interesujący model struktury kompleksu białko-RNA. Podsumowując przeprowadzony eksperyment stał się dużym krokiem naprzód w celu zrozumienia w jaki sposób białka typu Dicer funkcjonują w roślinach.

Mickiewicz A., Sarzynska J., Milostan M., Kurzynska-Kokorniak A., Rybarczyk A., **Lukasiak P.**, Kulinski T., Figlerowicz M., Blazewicz J., Modeling of the catalytic core of

Arabidopsis thaliana Dicer-like 4 protein and its complex with double-stranded RNA, Computational Biology and Chemistry 2017, 66 44-56

Sarzynska J., Mickiewicz A., Milostan M., **Lukasiak P.**, Blazewicz J., Figlerowicz M., Kulinski T., Flexibility of Dicer studied by implicit solvent molecular dynamics simulations Computational Methods in Science and Technology, 2010, 16(1), 97-104

6. W tym przypadku wątkiem badawczym było przewidywanie identyfikacji podziałów domenowych dla nieznanych białek. Zaprojektowane i zaimplementowane zostały podejścia algorytmiczne, których celem jest uproszczenie procesu analizy struktur białkowych. Podczas badań weryfikowano hipotezę o możliwości identyfikacji domen w białkach na podstawie ich sekwencji. Podstawową cechą algorytmu jest fakt, że przewiduje on podziały domenowe nieznanego białka z wykorzystaniem jedynie struktury pierwszorzędowej (sekwencja aminokwasów). Zaproponowany algorytm przeszukuje zaprojektowaną specjalnie dla tego problemu, bazę wiedzy w celu zidentyfikowania podziałów domenowych białek w nieznanym białku. Rozpatrywana baza wiedzy zawiera wzorce sekwencyjne o ustalonej długości utworzone na granicach pomiędzy domenami i/lub fragmentami, które zostały wygenerowane na podstawie istniejących baz danych podziałów domenowych białek takich jak *CATH, DALI, SCOP, Conserved Domains Database, InterPro, Uniprot*. Baza wiedzy zawiera również wzorce uzyskane z bazy danych *Pfam*, która zawiera podziały domenowe białek uzyskane na podstawie łańcuchów Markowa. Zaproponowane podejście algorytmiczne wykorzystuje bazę wiedzy w celu dopasowania zbioru wzorców sekwencyjnych reprezentujących granice podziałów domenowych do nieznanej sekwencji białka. Na podstawie dopasowanego zbioru wzorców algorytm dokonuje inteligentnego podziału domenowego analizowanego białka. Otrzymane wyniki nie potwierdziły w pełni postawionej hipotezy badawczej. Ponieważ zaproponowany algorytm ma podstawy opierające się na homologii, wyniki były bardzo dobre (80-90% poprawności) dla struktur homologicznych, natomiast dla struktur o niskim stopniu podobieństwa do znanych struktur wyniki były bardzo słabe, co oznacza, że informacja o strukturze pierwszorzędowej może być wykorzystywana jedynie jako pomocnicza przy rozwiązywaniu tego problemu.

Antczak M., Blazewicz J., **Lukasiak P.**, Milostan M., Krasnogor N., Palik G., DomAns-Pattern based method for protein domain boundaries prediction and analysis, Foundations of Computing and Decision Sciences, 2011, 36(2), 99-119

7. Poniższy wątek badawczy był związany z zaprojektowaniem metod przewidywania funkcji białek. Za podstawę przewidywania funkcji przyjęto wyniki wyszukiwania białek o sekwencjach podobnych do sekwencji badanego białka w bazie danych. Dlatego też na wstępie każdej analizy przeprowadzono eksperyment polegający na traktowaniu kolejno każdego białka ze zbioru danych jako białka o nieznanej funkcji i wyszukiwaniu dla niego podobieństw w bazie danych utworzonej przez pozostałe sekwencje z tego samego zbioru danych. W eksperymencie tym, przeprowadzonym niezależnie dla każdego zbioru danych, sprawdzone zostało zastosowanie różnych macierzy substytucji. Eksperyment prowadzono w dwóch wariantach: bez użycia rejonów sekwencji o niskiej złożoności oraz z wykorzystywania tej procedury. Główna metoda przewidywania funkcji, zastosowana w tej pracy, stanowiąca wariant algorytmu uczenia maszynowego k – najbliższych sąsiadów, opierała się na grupie kryteriów rozstrzygających na podstawie wyników wyszukiwania podobieństw w bazie danych. Wprowadzono grupę czterech kryteriów podstawowych, a także dalsze jedenaście kryteriów złożonych, utworzonych z połączeń tych pierwszych. Bazując na wynikach eksperymentów wyszukiwania podobieństw w bazach sekwencji przeprowadzony został szereg analiz, których celem było poznanie skuteczności zaproponowanego podejścia. Dokonano porównań różnych zbiorów danych (opisanych różnymi pojęciami GO i o zróżnicowanych liczbach tworzących je białek), zależności wyniku od macierzy substytucji i od stosowania filtra rejonów o niskiej złożoności, a także od przyjętego progu wartości E wyznaczającej istotność znajdywanych podobieństw. Rezultaty wykazały wysoką skuteczność tej metody, a w przypadku dużych zbiorów danych – bardzo wysoką (z trafnością decyzji na poziomie około 90% ze wszystkich sekwencji tworzących zbiór). Jednakże, z malejącym podobieństwem sekwencji znajdowanych w bazie danych, maleje również skuteczność tego rodzaju metody. Zależność tę przebadano, aby pokazać jakiej jakości predykcji można oczekiwać w zależności od stopnia podobieństwa sekwencji znalezionych w bazie. Wprowadzono także drugą metodą przewidywania funkcji, również bazującą na wynikach wyszukiwania podobieństw w bazie danych, lecz wykorzystującą do rozstrzygania o pełnionej przez białko funkcji algorytm

drzew decyzyjnych. Testy tego podejścia zostały przeprowadzone dla najmniejszego zbioru danych, w którym otrzymano najniższą skuteczność pierwszej metody. Wykorzystanie bardziej złożonego sposobu wnioskowania o funkcji białka zaowocowało zauważalnym wzrostem ilości białek ze zbioru, których funkcja została przewidziana prawidłowo. Przeprowadzone analizy potwierdziły przyjętą hipotezę.

Blazewicz J., **Lukasiak P.**, Milostan M., Krasnogor N., Jackowiak P., Sequence similarity based methods for protein function prediction, Foundations of Computing and Decision Sciences, 2009, 34, 173-192

8. Innym wątkiem badawczym były prace nad metodą uczenia maszynowego zwaną Logiczną Analizą Danych. W sensie praktycznym nastąpiła próba zweryfikowania hipotezy badawczej o możliwości przewidywania struktur drugorzędowych białek na podstawie znajomości ich sekwencji. W wyniku przeprowadzonych prac zaproponowany został model wspomnianego algorytmu Logicznej Analizy Danych, gdzie danymi wejściowymi były krótkie wzorce sekwencyjne wraz ze skojarzoną z nimi strukturą drugorzędową. Eksperymenty obliczeniowe na wzorcach o różnej długości pozwoliły wybrać optymalną długość wzorca sekwencyjnego i nauczyć metodę opartą o ten algorytm (wygenerować zbiór reguł), które były w stanie potwierdzić wskazaną na wstępie hipotezę na poziomie 60-70%.

Blazewicz J, Hammer P.L., **Lukasiak P.**, Predicting secondary structures of proteins IEEE Engineering in Medicine and Biology Magazine, 2005, Vol. 24(3) s 88-94

Podsumowując, mój dorobek naukowy składa się z:

- 18 artykułów w czasopismach wymienionych w bazie Journal Citation Reports,
- 7 artykułów wymienionych w części B wykazu MNiSW,
- 48 referatów wygłoszonych na krajowych konferencjach tematycznych,
- 15 staży naukowych (wszystkie po doktoracie),
- 63 recenzji projektów lub artykułów o zasięgu krajowym i międzynarodowym,
- Udziału w 12 projektach badawczych jako koordynator/wykonawca,
- 2 promotorstw pomocniczych prac doktorskich

Podsumowanie scjentometryczne moich osiągnięć naukowych przedstawia się następująco:

- Punkty MNiSW 541 (w tym 532 po doktoracie)
- Sumaryczny IF=55,788 (całość po doktoracie)
- Liczba cytowań według bazy Web of Science wynosi 327
- Indeks Hirsha według bazy Web of Science wynosi 8

Poznań, dn. 28.04.2014