UNIWERSYTET WARSZAWSKI | CeNT CENTRUM NOWYCH TECHNOLOGII

3 września, 2019

Prof. dr hab. Joanna Trylska
e-mail: joanna@cent.uw.edu.pl
telefon (22) 55 43 683
https://bionano.cent.uw.edu.pl

Rada Naukowa
Międzyuczelnianego Wydziału Biotechnologii UG i GUMed
ul. Antoniego Abrahama 58
80-307 Gdańsk

Recenzja rozprawy doktorskiej mgra Mateusza Pikory pt. "*Zastosowanie modelu Markova do badania ścieżek zwijania białek*"

Przedstawiona mi do recenzji rozprawa doktorska mgra Mateusza Pikory zatytułowana *"Zastosowanie modelu Markova do badania ścieżek zwijania białek"* została wykonana pod kierunkiem dra hab. Rajmunda Kaźmierkiewicza, prof. UG oraz promotora pomocniczego dra Artura Giełdonia.

Temat badawczy mgra Mateusza Pikory dotyczył struktury białek i ścieżek zwijania białek do struktur natywnych. Białka pełnią kluczową rolę we wszystkich organizmach, między innymi rolę enzymatyczną. Niestety nie znamy struktur trójwymiarowych wszystkich możliwych białek, a wyznaczenie takich struktur nie zawsze jest możliwe. Albo krystalizacja niektórych białek nie jest możliwa albo uzyskane kryształy są słabej jakości, żeby wyznaczyć ich strukturę. Z kolei badania spektroskopii magnetycznego rezonansu jądrowego są zwykle przeprowadzane dla małych białek. Nowa metoda kriomikroskopii elektronowej pozwala już na wyznaczenie wielu struktur ale nie jest jeszcze rutynowo stosowana. Wszystkie metody są drogie, mają swoje ograniczenia, m.in., nie dostarczają informacji o stanach pośrednich dochodzenia do danej struktury trójwymiarowej. Ważnym jest w jaki sposób białka przyjmują struktury natywne lub inne struktury związane z ich funkcją biologiczną. Nieprawidłowo zwinięte białka np. z powodu mutacji są przyczyną wielu chorób. Opracowanie metody analizy ścieżek zwijania białek oraz ich przechodzenia przez różne konformacyjne i energetyczne stany przejściowe może więc pomóc w zrozumieniu dlaczego niektóre białka nie przyjmują właściwej struktury funkcjonalnej.

Celem pracy mgra Pikory było wykorzystanie analizy skupień i modelu Markova do badania jak zwijają się białka. Dane do tych analiz pochodziły z symulacji dynamiki molekularnej kilku białek, które przeprowadził autor rozprawy. Symulacje były realizowane w dwóch modelach oraz dla dwóch zestawów parametrów tzw. pól siłowych. Jeden model białka był pełnoatomowy i symulacje wykonywane były w polu siłowym i programem Amber. Drugi model białka był gruboziarnisty i symulacje były wykonywane w polu siłowym i programem UNRES. Koncepcja pola siłowego UNRES powstała w grupie prof. Harolda Scheragi. Następnie pole siłowe i oprogramowanie zostało

rozwinięte i jest nadal rozwijane na Uniwersytecie Gdańskim w grupie prof. Adama Liwo. Dodatkowym elementem symulacji dynamiki molekularnej było wzmocnienie próbkowania przestrzeni konformacyjnej poprzez użycie różnych temperatur układu. Standardowe symulacje dynamiki przeprowadzane w temperaturze pokojowej są zwykle za krótkie by zaobserwować stany pośrednie zwijania się nawet małych białek. Wobec tego, aby takie zmiany zobaczyć autor rozprawy zastosował metodę dynamiki molekularnej z wymianą replik, tzw. *replica-exchange molecular dynamics*. Ilość danych otrzymywanych z takich symulacji jest ogromna (kilkaset tysięcy czy miliony struktur) i zwykle potrzebne są duże moce obliczeniowe, żeby takie dane nie tylko wytworzyć ale też zanalizować.

Integralną część rozprawy stanowi program napisany przez mgra Pikorę służący do analizy wielu konformacji: grupowania tych konformacji pod względem strukturalnym, budowy modelu Markova i uzyskania diagramów przejść między pogrupowanymi konformacjami. Program został napisany w języku C i jest zrównoleglony poprzez wykorzystanie biblioteki OpenMP. Pozwala na analizę setek tysięcy struktur w rozsądnym czasie.

Rozprawa doktorska mgra Mateusza Pikory została napisana w języku polskim, poza podręcznikiem użytkownika, który jest napisany w języku angielskim. Rozprawa składa się ze standardowych rozdziałów: Wstęp, Materiały i metody, Wyniki, Dyskusja. Wstęp poprzedza wykaz rysunków i skrótów. Bibliografia odwołuje czytelnika do ponad 100 pozycji. Odwołania są do publikacji naukowych indeksowanych w bazie *Journal Citation Reports* oraz do oprogramowania i repozytoriów. Integralną część rozprawy stanowi załącznik w formie kodu źródłowego programu opisanego przez autora w rozprawie, a także podręcznik użytkownika.

Pierwszy rozdział został podzielony na kilka podrozdziałów. W podrozdziale zatytułowanym "Białka" mgr Pikora opisuje budowę i funkcję białek, proces i modele służące do opisu zwijania białek oraz metody doświadczalne i teoretyczne wyznaczania struktury białek. Opis budowy białek jest bardzo szczegółowy, wraz z rysunkiem struktur naturalnych aminokwasów, omówieniem wiązania peptydowego, sekwencji, struktur drugo-, trzecio- i czwartorzędowych białek. Następnie mgr Pikora opisuje metody modelowania molekularnego, koncepcję wyznaczania energii potencjalnej cząsteczek (w tym pola siłowe), metodę dynamiki molekularnej, sposoby grupowania konformacji cząsteczek pod względem podobieństwa strukturalnego, modele Markova i zastosowanie łańcuchów Markova w kontekście symulacji biomolekularnych, między innymi do znajdowania ścieżek przejść pomiędzy konformacjami.

W krótkim rozdziale II autor przedstawia cel pracy i hipotezę badawczą.

Kolejny III rozdział to Materiały i metody. W tej części rozprawy opisana jest zasada działania programu *pdbclust* napisanego przez autora i stanowiącego integralną część rozprawy. Następnie, mgr Pikora opisuje struktury i funkcje białek, dla których przeprowadzał symulacje dynamiki molekularnej i pakiety oprogramowania, które wykorzystał do tych symulacji wraz z protokołami postępowania. Autor używał dwóch pól siłowych i programów - Amber i UNRES.

W rozdziale Wyniki opisana jest najpierw standardowa analiza przeprowadzonych symulacji; średnie odchylenie standardowe od struktury początkowej, współczynnik żyracji i energia, w funkcji czasu

symulacji. Analiza tych wartości w funkcji kroków czasowych potwierdziła prawidłowo wykonane symulacje dynamiki molekularnej. W następnych częściach rozdziału Wyniki przedstawione są analiza skupień oraz modelu Markova. Modele Markova zostały przygotowane dla struktur pochodzących z trajektorii w temperaturze 310 K. W przypadku pola siłowego UNRES odbudowano struktury pełnoatomowe dla zredukowanych konformacji białek. W tej części autor testował podział konformacji na określoną liczbę grup i sprawdzał właściwe budowanie modelu Markova, np. symetryczność macierzy przejścia. Analiza skupień została przedstawiona dla trzech badanych białek, najpierw dla pola siłowego UNRES, a następnie dla pola siłowego Amber. Autor stwierdził, że nie zawsze struktura natywna, rozumiana jako struktura krystalograficzna lub NMR z bazy danych Protein Data Bank, zajmuje największą liczbę stanów. Wyniki tej analizy pokazują różne możliwe ścieżki dochodzenia do struktury natywnej, a także fakt, że po osiągnięciu stanu natywnego białko może się częściowo rozwijać i przyjmować też inne struktury pośrednie. Stany pośrednie zostały w każdym przypadku przeanalizowane pod względem strukturalnym. Na podstawie analiz autor mógł zaproponować różne możliwe ścieżki zwijania badanych białek. Wyjaśnił też dlaczego obserwuje takie a nie inne wyniki np. stwierdzając nadmierną stabilizację helis czy faworyzowanie ich utworzenia w polu siłowym UNRES. W jednym przypadku dla struktury białka o kodzie 2MQ8 pole siłowe UNRES nie zwinęło struktury do konformacji natywnej.
W przypadku pola siłowego Amber dla struktury o kodzie PDB 1BDD nie uzyskano ścieżek przejścia do struktury natywnej biorąc pod uwagę konformacje w temperaturze 310 K. Wobec tego autor zdecydował się przedstawić diagram przejść dla klatek trajektorii pochodzących z symulacji w różnych temperaturach.
Z kolei dla pełnoatomowych symulacji w polu siłowym Amber dla białka o kodzie 2MQ8 nie udało się uzyskać dobrych diagramów, nawet biorąc pod uwagę wiele temperatur. Autor stwierdził, że symulacje były za krótkie i układ nie osiągnął ani stanu struktury natywnej, ani innych przejściowych stabilnych energetycznie stanów, gdyż grafy nie były spójne. Podsumowując, na podstawie otrzymanych trajektorii autor rozprawy mógł przewidzieć krótkotrwałe i bardziej stabilne energetycznie stany przejściowe białek, a także stany, które wchodziły w tzw. ślepy zaułek.

W rozdziale Wyniki przedstawiona została także analiza wydajności programu autora *pdbclust*. Program ten został zrównoleglony i testy wydajności pokazały, że czas symulacji rośnie liniowo aż do około 400-500 tysięcy analizowanych konformacji dla 2, 8 i 32 wątków. Wąskim gardłem jest wczytywanie danych co przy tak dużej liczbie ma znaczenie i nie da się tego ominąć czy przyspieszyć.

Kolejny rozdział jest zatytułowany Dyskusja. W nim autor rozprawy podkreśla, że nie istnieje jedna ścieżka zwijania prowadząca do danej struktury białka, a może istnieć wiele stanów przejściowych. Jest to koncepcja znana ale trudna do potwierdzenia w symulacjach, gdyż przy zwijaniu białek bardzo często należy stosować metody wzmocnionego próbkowania przestrzeni konformacyjnej lub innego rodzaju metody, aby białko zwinęło się do struktury natywnej. Aby pogrupować struktury mgr Pikora stosował algorytm najbliższych sąsiadów ale jako centra grup przyjmował struktury o najniższych energiach oraz strukturę natywną. W klasycznym podejściu tego algorytmu centrum grupy wybierane jest na podstawie liczby sąsiadów. Takie założenie grupowania na podstawie energii układu powoduje, że wybierane są stabilne struktury pośrednie. W Dyskusji autor też podkreśla jak należy dobierać parametry modeli, żeby analiza była rozsądna z punktu widzenia biologicznego.

Ważna część pracy doktorskiej mgra Pikory dotyczy nie tylko przygotowania i przeprowadzenia symulacji wzmocnionego próbkowania dynamiki molekularnej ale także opracowania algorytmu postępowania i zaimplementowania pomysłu analizy danych w postaci oprogramowania *pdbclust*. Oprogramowanie to jest dostępne w ramach tzw. otwartego dostępu dla innych badaczy. Mgr Pikora ma na swoim koncie modyfikacje też innego oprogramowania - RASMOL. Dodanie nowych narzędzi w tym oprogramowaniu zostało opisanych w 2015 roku w publikacji (M. Pikora, A. Giełdoń, *Acta Biochimica Polonica*, RASMOL AB - new functionalities in the program for structure analysis, 2015;62(3):629-31). Czy planowana jest także publikacja metody i wyników badań przeprowadzonych oprogramowaniem *pdbclust*?

Po przeczytaniu rozprawy mam kilka pytania i zagadnień, które mnie zastanowiły.

W obydwu polach siłowych wykorzystano modele ciągłe rozpuszczalnika (wody). O ile w przypadku pola siłowego UNRES jest to zrozumiałe, gdyż to pole siłowe zostało sparametryzowane w taki sposób, żeby efekt rozpuszczalnika był uwzględniony w parametrach, o tyle w pełnoatomowym polu siłowym Amber już tak nie jest. Czy model ciągły w symulacjach pełnoatomowych był zastosowany do celów porównania z UNRES, czy może były przeprowadzane jakieś symulacje w modelach jawnych (*explicite*) cząsteczek wody?

Czy jest jakieś ograniczenie w stosowaniu programu dotyczące wielkości białek bądź liczby konformacji poza czasem oczekiwania na wyniki? Autor badał małe białka, w praktyce przeprowadzamy symulacje dla większych białek. Wykonywanie programu zachowuje się liniowo do ok. 400 000 struktur ale dla małych białek.

Czy sprawdzono, czy w przypadku pełnoatomowej dynamiki molekularnej z wymianą replik rozkłady konformacji/energii otrzymane w różnych temperaturach odpowiednio na siebie nachodziły, czy może w przypadku multiplexed REMD nie było to potrzebne? W przypadku pola siłowego Amber nie jest możliwym stwierdzenie tego faktu z przedstawionych wykresów.

Istotna uwaga dotyczy rysunków 8.1, 12.1 i dalszych, czyli wszystkich wykresów pokazujących zależności różnych wartości od czasu symulacji. Przy tej ilości linii i zmienności wartości na osi Y (zwłaszcza dla pola siłowego Amber) wykresy te są nieczytelne i linie pokrywają się. Zwykle stosuje się uśrednione wartości po pewnych okresach czasowych tzw. "*running average*". Takie przedstawienie by wystarczyło. Zmienność można było pokazać na jednym wykresie i na tym jednym nanieść na dane uśrednioną linię. Można było też pokazać rozkłady tych wartości.

Dodatkowo, na wykresach oś Y zawiera jedynie początkową i końcową wartość bez żadnych wartości pośrednich (czy nawet kresek). Ciężko więc wywnioskować po jakim czasie linie na wykresach się stabilizują. Na niektórych wykresach w załączniku nie widać podpisów na osi X albo jedna liczba jest umieszczona gdzieś na górze wykresu. Autor rozprawy powinien inaczej podejść do prezentacji tego typu danych. Czasem na osi X brakuje także 0. Wykresy te są przedstawione niestarannie i mają słabą jakość na wydruku oraz za małe oznaczenia osi oraz legend.

Schemat analizy i prezentacji wyników jest taki sam dla wszystkich białek i pól siłowych. Wiele informacji dotyczących sposobu analizy i dla danego białka i dla danego pola siłowego powtarza się i

pewnie lepiej, by takie stwierdzenia znalazły się w Materiałach i metodach lub na początku rozdziału Wyniki.

Zauważyłam jeszcze trochę drobnych przeoczeń językowych, część z nich wymieniam poniżej.
- tytuł podrozdziału "Determinowanie struktur białek" nie brzmi po polsku. Lepiej by było Wyznaczanie lub Określanie struktur białek.
- w zdaniach, gdzie występują D i L peptydy, D- i L- powinno być pisane kapitalikiem.
brakuje przecinków w zdaniach przed "który", "które" lub "w której" np. strona 29, 30, 34, 53, 55, 61, 105, 112, 113, 114, 119, 121, 132, 148, 150, 151.
- strona 52, powinno być "Dla dwóch optymalnie..."
- strona 60, powinno być "Dla niewielkich układów..."
- strona 61, nie rozumiem określenia "przy użyciu zachłannego algorytmu z nawrotami"
- strona 68, powinno być "tę decyzję..."
- strona 94 i 95 brakuje spacji przez jednostką Angstrom
- strona 97, powinno być "Dla każdej..."
- strona 99, w podpisie pod rysunkiem 9.1 powinny być podane jednostki wartości RMSD oraz promienia żyracji podane na diagramie.
- strona 121 powinno być "Hipotezę tę..."
- strona 144 powinno być "Dla pełnego..."
- strona 148, brakuje kodu PDB "o kodzie ID...."

Ogólnie jednak rozprawa doktorska mgra Pikory jest napisana bardzo klarownie, poprawną polszczyzną i dobrze się ją czyta. Wstęp jest obszerny, napisany w sposób dydaktyczny i będzie mógł posłużyć nowym członkom zespołu do zapoznania się z budową i modelami zwijania białek do struktury trzeciorzędowej.
Trajektorie pochodzące z symulacji dynamiki molekularnej zawierają miliony konformacji symulowanej cząsteczki, więc szybka analiza takiej ilości danych jest niezwykle potrzebna. Wyciąganie właściwych informacji z trajektorii staje się coraz trudniejszym zadaniem i algorytm oraz oprogramowanie przedstawione przez mgra Pikorę wnosi istotny wkład w metody analizy dużych danych. Analiza skupień połączona z budowaniem modeli Markova może być stosowana nie tylko do białek ale także każdego innego zestawu danych zawierającego konformacje cząsteczek. Stosowanie modeli Markova do analiz trajektorii jest podejściem dosyć nowym i wymaga testów na wielu przykładach, a także rozwoju i dostosowania metod. Należy podkreślić, że istotnym elementem pracy mgra Pikory była analiza ogromnej ilości danych i umiejętność pogrupowania tych danych.

Podsumowując, stwierdzam, że rozprawa doktorska mgra Mateusza Pikory stanowi oryginalne rozwiązanie problemu naukowego, potwierdza jego ogólną wiedzę teoretyczną w dyscyplinie nauk biologicznych oraz umiejętność samodzielnego prowadzenia pracy naukowej. Rozprawa doktorska mgra Pikory spełnia więc warunki ustawy o tytule naukowym i stopniach naukowych. W związku z tym wnoszę o dopuszczenie mgra Pikory do dalszych etapów przewodu doktorskiego.

J. Trylska