dr hab. Adam Jarmuła
Pracownia Bioinformatyki
Instytut Biologii Doświadczalnej
im. M.Nenckiego PAN
Ludwika Pasteura 3
02-093 Warszawa

Warszawa, 28 sierpnia 2019 r.

Recenzja rozprawy doktorskiej p. mgr Mateusza Pikory zatytułowanej „Zastosowanie modelu Markova do badania ścieżek zwijania białek”

Przedstawiona mi do oceny rozprawa doktorska na stopień doktora nauk biologicznych w zakresie biochemii została wykonana na Międzyuczelnianym Wydziale Biotechnologii Uniwersytetu Gdańskiego oraz Gdańskiego Uniwersytetu Medycznego (MWB UG-GUMed). Promotorem rozprawy p. mgr Mateusza Pikory jest p. dr hab., prof. UG, Rajmund Kaźmierkiewicz, kierujący Pracownią Symulacji Układów Biomolekularnych MWB UG-GUMed, zaś promotorem pomocniczym p. dr Artur Giełdoń z Pracowni Symulacji Polimerów w Zakładzie Modelowania Molekularnego na Wydziale Chemii UG.

Rozprawa doktorska p. mgr Mateusza Pikory w części zasadniczej liczy 152 strony i jest podzielona na 5 głównych części. Rozprawę rozpoczyna Wstęp obejmujący 47 stron, po którym następują kolejno ujęty zwięźle Cel Pracy, Materiały i metody (23 strony), Wyniki (57 stron) oraz Dyskusja (5 stron). Dokumentację części zasadniczej kompletują 53 rysunki oraz 6 tabel. Rozprawę uzupełniają Spis treści, Spis rysunków, Wykaz skrótów, Streszczenie w językach polskim i angielskim, część dodatkowa (Dodatki) podzielona na napisaną w j. angielskim Instrukcję programu *pdbclust* (14 stron) oraz szczegółowe wykresy parametrów uzyskanych na podstawie trajektorii przeprowadzonych symulacji dynamiki molekularnej z wymianą replik (19 stron), i wreszcie licząca 108 pozycji literaturowych Bibliografia (9 stron). Na płycie CD dołączonej do rozprawy znajduje się wersja instalacyjna programu *pdbclust* wraz z prostym pakietem testowym.

Modelowanie komputerowe zwijania (fałdowania) białek w struktury natywne ma duże znaczenie w sytuacji bogactwa białek o znaczeniu biologicznym, których struktury trójwymiarowe nie zostały dotąd wyznaczone metodami eksperymentalnymi takimi jak krystalografia białka, spektroskopia magnetycznego rezonansu jądrowego (NMR) czy ostatnio kriomikroskopia elektronowa 3D. Dotyczy to np. w dużym stopniu białek błonowych (membranowych), o często dużych rozmiarach, dla których eksperymentalna determinacja struktury trójwymiarowej bywa szczególnie trudna. Zasadniczym problemem w modelowaniu przez dynamikę molekularną jest brak dostatecznych mocy komputerowych do przeprowadzenia długiej symulacji zwijania białka w sytuacji gdy te „zdarzenia” molekularne realizują się w czasach rzędu mikro- lub częściej mili-sekund, a nawet dłuższych. Alternatywą dla szczególnie wymagającej, gdy idzie o zasoby komputerowe, pełno-atomowej dynamiki molekularnej w środowisku wodnym są symulacje gruboziarniste, gdzie dzięki znaczącej redukcji rozdzielczości układu, a zatem ilości jego stopni swobody, udaje się wykonywać o rzędy wielkości dłuższe symulacje dynamiki molekularnej. Problemem w modelowaniu homologicznym

bywa z kolei brak dostatecznej liczby homologów modelowanych białek i konieczność posłużenia się bardziej zawodnymi metodami *ab initio*. W ostatnim czasie na znaczeniu zyskują w szybkim tempie metody uczenia maszynowego i uczenia głębokiego (ang. *machine learning* i *deep learning*), które zdają się mieć szansę odegrać istotną rolę w modelowaniu zwijania białek.

Rozprawa p. mgr Mateusza Pikory wpisuje się w nurt modelowych badań zwijania białka, dotycząc w szczególności mechanizmów i ścieżek tych procesów. Doktorant używa kombinacji multipleksowej dynamiki molekularnej z wymianą replik (MREMD), analizy skupień (ang. *clustering*) przy pomocy zmodyfikowanego Algorytmu Najbliższego Sąsiada oraz implementacji Łańcucha Markova do zdefiniowania alternatywnych ścieżek fałdowania 3 przykładowych białek z bazy PDB o kodach 1bdd (pojedynczy łańcuch polipeptydowy o długości 60 reszt aminokwasowych), 1l2y (mini-białko z pojedynczym łańcuchem o długości 20 reszt aminokwasowych) i 2mq8 (pojedynczy łańcuch o długości 112 reszt aminokwasowych). Dwa ostatnie zadania, tj. analiza skupień oraz budowa Łańcucha Markova, wykonywane są z pomocą programu komputerowego *pdbclust*, zaprojektowanego i napisanego przez doktoranta w języku C z wykorzystaniem bibliotek do zrównoleglenia obliczeń. Hipotezą badawczą postawioną przez doktoranta jest, że białka nie posiadają jednej dominującej ścieżki fałdowania rozumianej jako jednoznaczna i powtarzalna seria zdarzeń prowadząca od struktury rozwiniętej do konformacji natywnej.

Wstęp w rozprawie doktorskiej p. mgr Mateusza Pikuły podzielony jest na 5 rozdziałów. Pierwszy z nich obejmuje budowę białek (z charakterystykami m. in. wiązania peptydowego oraz struktur od pierwszo- do czwarto-rzędowej), proces zwijania białek oraz różne jego modele, rolę biologiczną białek oraz metody eksperymentalnej i modelowej determinacji struktur białkowych. Rozdział drugi uwzględnia szczegółową charakterystykę modelowania molekularnego (dyskutowane są parametryzacja pól siłowych, optymalizacja energii potencjalnej struktury, charakterystyka podstawowej dynamiki molekularnej oraz dynamika molekularna i multipleksowa dynamika molekularna z wymianą replik). Rozdział trzeci omawia metodę analizy skupień oraz jej podstawowe algorytmy. Rozdział czwarty skupia się na charakterystyce modeli Markova, dostarczając informacji na temat tworzenia i zastosowania łańcuchów Markova. Końcowy rozdział piąty Wstępu jest poświęcony zastosowaniu analizy skupień i łańcuchów Markova w mechanice molekularnej. Łącznie, Wstęp jest napisany kompetentnie i zawiera odpowiednią dawkę informacji potrzebną w dalszym czytaniu rozprawy.

Rozdział Materiały i Metody składa się z 2 dużych części. Część pierwsza omawia fachowo i szczegółowo autorski program doktoranta, *pdbclust*, który wydaje się być flagowym osiągnięciem w recenzowanej rozprawie. W drugiej części scharakteryzowane są 3 struktury z PDB, wykorzystane w obliczeniach, oraz programy do dynamiki molekularnej, Amber (dynamika pełno-atomowa) i UNRES (dynamika gruboziarnista), przy pomocy których wykonane zostały symulacje REMD. Rozdział jest napisany klarownie i merytorycznie poprawnie, nie mam do niego zastrzeżeń.

W rozdziale kolejnym, Wynikach, doktorant dokonuje analizy przeprowadzonych symulacji, skupiając się na następujących parametrach wynikowych: energii całkowitej, współczynnikowi żyroskopowemu oraz statystykom RMSD względem struktur natywnych. Następnie omawiana jest analiza skupień oraz tworzenie modelu Markova, oba obliczenia wykonane z wykorzystaniem programu *pdbclust*. W tej części, na początkowych 2 stronach opisany jest protokół analiz trajektorii z symulacji REMD, a następnie ich dokładny opis dla trajektorii z symulacji wszystkich 3 białek, najpierw w programie i polu siłowym UNRES, a potem w Amberze. Łącznie, przedstawienie uzyskanych wyników oraz ich dyskusja są poprowadzone systematycznie i szczegółowo. Doktorant

koncentruje się na omówieniu poszczególnych, mniej lub bardziej stabilnych stanów przejściowych oraz relacji pomiędzy nimi (przejścia w jedną i/lub drugą stronę). Zdaniem doktoranta, wyniki analiz każdej wykonanej symulacji potwierdzają główną hipotezę badawczą, wskazując na różnorodność ścieżek zwijania prowadzących do struktury natywnej oraz nieultymatywną stabilność struktury natywnej, pozostającej w równowadze z różnymi, częściowo zwiniętymi konformacjami należącymi do stanów realizujących przejścia z/do stanu struktury natywnej. W tym miejscu chciałbym zwrócić uwagę na pewne drobniejsze błędy/niedociągnięcia napotkane w Wynikach:

1) Pomysł zmieszczenia na jednym wspólnym wykresie określonych parametrów/metryk (energia całkowita, temperatura, współczynnik żyroskopowy, RMSD) z 8 replik danej symulacji REMD skutkuje przepełnieniem wykresów, które z tego powodu stają się słabo czytelne. W związku z tym np. zdanie na stronie 92, „Widać też że zależność energii od temperatury repliki jest bardzo wyraźna w programie UNRES, a w polu sił programu AMBER jej nie widać" może być przyjęte głównie na wiarę, gdyż wykresy są za mało przejrzyste by klarownie potwierdzić tezę doktoranta;

2) Dyskusja wyników RMSD wprowadza pewne zamieszanie. Parametr ten może służyć jako ogólna miara podobieństwa 2 struktur, jednak szeroka skala wartości uzyskana w symulacjach doktoranta pozwala na jedynie bardzo zgrubny, wysoce nieostry obraz podobieństwa. W tej sytuacji pomocne byłyby chociaż przykładowe rysunki ukazujące nałożenie struktur, dla którego liczone były RMSD;

3) Na stronie 97 doktorant podaje iż „W niektórych przypadkach (kiedy uznałem, że to może dostarczyć dodatkowych informacji) przygotowałem drugi model złożony ze wszystkich struktur.", co pozostaje w opozycji do sytuacji gdy dla każdej symulacji „przygotowałem model Markova ze struktur o temperaturze 310 K, która jest najbliższa temperaturze fizjologicznej spośród wykorzystanych w symulacji." Bez wyjaśnienia czym miałyby być dodatkowe informacje proceder wygląda na działanie czysto arbitralne, brakujące racjonalnego imperatywu;

4) W tekstach szczegółowo analizujących trajektorie razi odrobinę bezrefleksyjne hierarchizowanie stabilności poszczególnych stanów w zależności od procentowej obecności w nich struktur pochodzących z przedziału temperatur 280-310 K. Generalnie taka hierarchizacja wydaje się wprawdzie sensowna, jednak czasem werdykty w ten sposób zasądzone uwidaczniają tym lepiej problemy czasowe analizowanych symulacji (niedostateczna długość), jak np. gdy mowa w kategoriach dużej stabilności o stanie ze strukturą natywną obejmującym zaledwie 0,3 % struktur użytych w analizie skupień (1bdd w polu Amber). W tym wypadku układ prawdopodobnie nie zdążył osiągnąć w większej ilości struktur stanu natywnego ze względu na zbyt szybko zakończoną symulację;

5) Na stronie 121 doktorant pisze o napotkanych strukturach przypominających struktury drugo-rzędowe: „Być może są one wprost β-kartkami, które zostały na tyle zaburzone przez proces odbudowy łańcucha, że program PyMOL przestał je wykrywać." Jakim narzędziem do wykrywania struktury drugo-rzędowej dysponowała wersja programu PyMOL, którą posługiwał się doktorant? W razie wątpliwości, można odwołać się programu dssp obecnego w różnych dystrybucjach, między innymi na serwerze online;

6) Na stronie 132, przeprowadzając analizę trajektorii białka 1l2y w polu siłowym Amber, doktorant pisze: „Wskazuje to na raczej prawidłowe przygotowanie modelu Markova." Stwierdzenie „raczej prawidłowe" nie jest fortunne; być może doktorant mógłby wywieść swoje końcowe przekonanie w oparciu o inne jeszcze przesłanki prócz bliskiej symetrii macierzy przejścia, dla której jednak występują wartości znacząco różniące się od swoich odpowiedników w macierzy?;

7) Przedstawienie struktur centralnych oraz struktury natywnej w kolorze zielonym na ciemnoszarym tle nie gwarantuje odpowiednio dobrego kontrastu by wygodnie obserwować struktury. W szczególności na grafach przejść struktury wtłoczone w małe prostokąty są słabo widoczne.

W rozdziale Dyskusja doktorant podsumowuje uzyskane wyniki oraz przeprowadza dokładniejszą dyskusję zaproponowanej przez siebie wersji metody analizy skupień oraz parametrów które decydują o praktycznym sukcesie metody. Formułuje również propozycje/wnioski na przyszłość, koncentrujące się na ulepszeniu jakości uzyskiwanych modeli Markova oraz ulepszeniu wyników analizy skupień. Na podstawie przeprowadzonej w rozprawie w końcowym podrozdziale Wyników Analizy wydajności programu *pdbclust*, doktorant stwierdza ponadto że „możliwa jest analiza setek tysięcy i więcej struktur w czasie rzędu godzin", co uznane jest za dobry wynik. Moje trzy uwagi do Dyskusji:

8) na stronie 148 doktorant podaje przykład oznaczenia pojedynczej, ogólnej ścieżki zwijania, dla stryktury „o kodzie ID w polu sił UNRES", zapominając jednak zapodać kod (zapewne chodzi tu o strukturę 1l2y, choć poprawniej byłoby powiedzieć że zaproponowane były tu 2 alternatywne ogólne ścieżki);
9) doktorant pisze na stronie 150 o wyborze liczby grup wykorzystywanych jako stany w konstruowaniu modelu Markova: „Z moich obserwacji wynika, ze 10 jest w wielu przypadkach rozsądną wartością". Wygląda to na rodzaj kompromisu i/lub wartość przybliżoną. Przydatne byłoby rozwinięcie wskazujące źródła takiego oszacowania;
10) czy doktorant zna alternatywne do *pdbclust* programy, które wykonywałyby podobne zadania? Jeśli tak, pożyteczne byłyby statystyki porównawcze wydajności programów...

W rozprawie doktorant powołuje się na 108 pozycji literaturowych z różnego okresu czasu, zarówno historycznych jak i nowszych. Lista sprawia wrażenie dobrze zestawionej i dowodzi biegłej znajomości literatury przedmiotu przez doktoranta. Sugerowałbym jednak, aby w wykazie literatury powoływać się w przypadku struktury z bazy Protein Data Bank nie tylko na kod w tej bazie, lecz o ile jest dostępna, korzystać z oryginalnej cytacji. Taka sytuacja ma miejsce w przypadku wszystkich 3 białek wykorzystanych w symulacjach REMD przez doktoranta.

Przejdę teraz do oceny generalnej. Struktura rozprawy doktorskiej p. mgr Mateusza Pikuły jest uporządkowana i klarowna, choć doktorant nie uniknął szeregu niezręczności stylistycznych czy też błędów interpunkcyjnych, które sugerują pewną niestaranność edytorską, np. na stronie 42 zamiast „będącym" winno być „będącemu", na stronie 60 zamiast „Zmniejsza się go..." - „Zmniejsza się je...", na stronie 64 zamiast „Algorytmy" - „Algorytmu", na stronie 76 zamiast „maksymalna" - „maksymalną", na stronie 102 zamiast „helisie" - „helisy" itd. itd., listę można mnożyć. Doktorant zadeklarował zrealizowanie swojego zamierzenia badawczego w 5 na 6 przypadków, tj. wyznaczonych wstępnie symulacji REMD struktur z PDB. Wypracowanie oryginalnej drogi badania ścieżek zwijania białek, poparte napisaniem programu komputerowego w znacznym stopniu ułatwiającego (automatyzującego) wykonanie zadań obliczeniowych przewidzianych w algorytmie postępowania, zasługuje z pewnością na słowa uznania. Odnosząc się jednak wprost do celu badawczego, nie sposób nie wspomnieć o pewnych niedociągnięciach w zaplanowaniu badań. I tak, nie jest jasne czym kierował się doktorant w doborze:

1) 3 białek wykorzystanych do testowania zarówno hipotezy badawczej jak i funkcjonowania swojego programu (poza ogólnym stwierdzeniem że wybór dotyczył niewielkich białek) oraz
2) innego ważnego parametru, tj. czasu dedykowanego symulacjom dynamiki molekularnej.

Wybór multipleksowej dynamiki molekularnej z wymianą replik można zrozumieć dążeniem do zagwarantowania penetracji możliwej szerokiej przestrzeni konformacyjnej poprzez unikanie zakleszczenia układu w lokalnych minimach energetycznych. Jednak wyznaczenie *a priori* czasu symulacyjnego w wymiarze jednakowym dla wszystkich 6 symulacji REMD, bez zważania na rozmiar

układu (20 reszt aminokwasowych w najmniejszym układzie względem 112 w największym) wydaje się niezrozumiałe. Warto w tym wypadku być elastycznym i programować czas symulacji stosownie do potrzeb. Zdaniem recenzenta, niepowodzenie w przypadku analizy trajektorii pełno-atomowej symulacji białka 2mq8 w polu siłowym Amber można było antycypować, nawet pomimo „subiektywnie" okazałego czasu takiej symulacji (łącznie 64 repliki po 96 ns każda), biorąc pod uwagę ogrom zadania obliczeniowego dla zbadania zwijania średniej wielkości białka z dużą liczbą stopni swobody. Prócz tego, gruboziarnista symulacja białka 2mq8 w polu UNRES wydaje się także „niedokończona", zważywszy na zupełnie znikomą populację stanu ze strukturą natywną (~0,04 %) przy zaledwie 30 % struktur w tak małej próbce wywodzących się z temperatury fizjologicznej lub niższych, tj. z zakresu 280-310 K. Co więcej, również symulacja pełno-atomowa w polu Amber dotycząca układu 1bdd przyniosła wyniki stawiające pod poważnym znakiem zapytania wystarczalność czasową symulacji REMD ze względu na bardzo wątłą populację stanu natywnego wynoszącą 0,3 %. Choć więc doktorant poprowadził do końca w sumie 5 analiz, to jednak wyniki 2 z nich (UNRES dla 2mq8 i Amber dla 1bdd) budzą poważne wątpliwości czy czas symulacyjny był w ich przypadku dostatecznie długi by oszacować z pewnością najważniejsze możliwe ścieżki fałdowania do struktury natywnej oraz sieci powiązań pomiędzy stanami przejściowymi. Pryncypialnie podchodząc do sprawy, takiej pewności w symulacjach dynamiki molekularnej uzyskać nie sposób. Ale dobrym pomysłem byłoby wypróbować dla każdego z badanych układów po kilka różnych czasów symulacji by przekonać się, jak ten kluczowy parametr wpłynie na wyniki analizy skupień oraz generowanie Łańcucha Markowa. Na końcu tego rozumowania jest naturalnie pytanie o zasoby obliczeniowe którymi dysponował doktorant, jednak obiektywnie **sytuację z niedostatecznym czasem niektórych symulacji uważam za podstawowy mankament recenzowanej rozprawy.** Oczekiwałbym stosownego komentarza doktoranta na ten temat podczas obrony pracy doktorskiej.

Podsumowując stwierdzam że rozprawa doktorska p. mgr Mateusza Pikory jest napisana w sposób dojrzały i dowodzi dobrego opanowania warsztatu badawczego przez doktoranta. Uważam, że pomimo pewnych niedociągnięć w zaplanowaniu badań (w kilku przypadkach niedostateczny czas symulacji dla zbadania nakreślonych problemów), rozprawa doktorska p. mgr Mateusza Pikuły broni się przez zawarte w niej pozytywy, takie jak przede wszystkim wypracowanie algorytmu postępowania w analizie trajektorii dynamiki molekularnej celem zbadania mechanizmów fałdowania białek. Szczególnym osiągnięciem recenzowanej rozprawy jest napisanie i przekazanie do domeny publicznej pożytecznego programu do analizy skupień i budowania modelu Markova na podstawie trajektorii dynamiki molekularnej. Doktorant deklaruje, i pozostaje Mu w to wierzyć, że program *pdbclust* jest przygotowany do szybkich analiz w trybie równoległym na bardzo dużych zbiorach danych, dzięki czemu zakres jego potencjalnego zastosowania wydaje się być szeroki. Wobec tego stwierdzam ostatecznie, że przedłożona mi do recenzji rozprawa doktorska spełnia warunki określone w Ustawie 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, z późn. zm.) i na tej podstawie wnoszę do Rady Międzyuczelnianego Wydziału Biotechnologii Uniwersytetu Gdańskiego oraz Gdańskiego Uniwersytetu Medycznego (MWB UG-GUMed) o dopuszczenie p. mgr Mateusza Pikory do dalszych etapów przewodu doktorskiego.

A. Jarmuła