prof. dr hab. Krzysztof Lewiński
Wydział Chemii
Uniwersytetu Jagiellońskiego

Kraków, 8.08.2014

Recenzja rozprawy doktorskiej mgr. Przemysława Porębskiego
"Improvement of model building and refinement of macromolecular crystal structures"

Liczba struktur makromolekuł dostępnych w Protein Data Bank przekroczyła w bieżącym roku 100 000. Około 90% z nich stanowią struktury białek wyznaczone metodą dyfrakcji promieniowania rentgenowskiego. W ostatnich latach liczba deponowanych struktur wynosi około 10 tys. rocznie i systematycznie rośnie. Wynika to z wielu czynników, wśród których należy szczególnie podkreślić coraz doskonalsze metody prowadzenia i przetwarzania pomiarów, rozwiązywania i udokładniania struktur a także coraz większą liczbę naukowców zaangażowanych w ten proces. Rozwój oprogramowania, w którym coraz większa część procedur jest zautomatyzowana i nie wymaga podejmowania decyzji przez człowieka sprawia, że coraz częściej struktury są wyznaczane przez osoby będące specjalistami w dziedzinach innych niż krystalografia makrocząsteczek. Z jednej strony pozwala to znacznie przyspieszyć postęp w zakresie biologii strukturalnej, z drugiej strony jednak niesie za sobą potencjalne niebezpieczeństwo deponowania i publikowania struktur o jakości niższej niż wynikająca z zebranych danych eksperymentalnych. Częściowym rozwiązaniem tego problemu są coraz bardziej wyrafinowane metody weryfikacji wyznaczanych struktur zarówno na etapie rozwiązywania jak i deponowania, jednak ich rola sprowadza się tylko do sygnalizowania nieprawidłowości, których przyczyny niejednokrotnie są trudne do ustalenia i usunięcia. Dlatego, coraz istotniejsze jest rozwijanie nowych narzędzi do wyznaczania i udokładniania struktur, które skuteczniej będą pozwalały unikać błędów na wczesnych etapach budowy modelu struktury. Właśnie taki jest cel stworzonego w ramach recenzowanej rozprawy doktorskiej programu *Fitmunk*.

Praca doktorska pana Przemysława Porębskiego powstała w wyniku współpracy specjalizującej się w modelowaniu molekularnym grupy badawczej kierowanej przez profesor Martę Pasenkiewicz-Gierulę na Wydziale Biochemii, Biofizyki i Biotechnologii Uniwersytetu Jagiellońskiego w Krakowie z grupą badawczą w University of Virginia (Charlottesville, USA) kierowaną przez profesora Władka Minora. W grupie prof. Minora należącej do konsorcjum Midwest Center for Structural Genomics realizowane są liczne projekty badawcze związane zarówno z systematycznymi badaniami strukturalnymi białek jak również z rozwijaniem oprogramowania służącego do przetwarzania danych dyfrakcyjnych i rozwiązywania struktur makrocząsteczek. Pan Porębski włączył się w te prace w trakcie odbywania długoterminowego stażu badawczego w University of Virginia.

Rozprawa ma klasyczny podział na wprowadzenie, materiały i metody oraz wyniki i ich dyskusję, zawiera także liczącą nieco ponad 100 pozycji bibliografię, streszczenia w języku polskim i angielskim oraz podsumowanie i cztery dodatki. Przypuszczalnie ze względu na to, że rozprawa powstawała w trakcie pobytu Pana Porębskiego w USA a także ze względu na możliwość bezpośredniego użycia znacznych jej fragmentów jako dokumentacji programu, rozprawa jest zredagowana w języku angielskim.

Zasadniczym celem pracy była próba opracowania nowego algorytmu, który poprzez zastosowanie metod modelowania molekularnego, umożliwił by automatyczne rozpoznawanie i dopasowywanie łańcuchów bocznych aminokwasów do eksperymentalnie wyznaczonej gęstości elektronowej. Obecnie istnieje kilka programów, które umożliwiają takie dopasowanie automatycznie bądź wspomagają manualny wybór optymalnej konformacji łańcucha bocznego na etapie budowy lub przebudowy modelu. Programy te i algorytmy na których bazują zostały dość szczegółowo omówione we wstępie do rozprawy oraz w części "Materials and methods". Tak więc zadaniem doktoranta było nie tylko opracowanie jeszcze jednego programu ale przede wszystkim stworzenie takiej metody dopasowywania, która będzie prowadziła do uzyskiwania lepszego modelu struktury niż przy zastosowaniu dotychczasowych procedur. W tym celu połączył on ze sobą w nowym algorytmie stosowane dotychczas metody wyszukiwania najlepszych konformerów oraz wprowadził nową parametryzację i biblioteki konformerów.

Pierwszym etapem pracy było przygotowanie biblioteki konformerów w oparciu o wysokorozdzielcze struktury krystaliczne zdeponowane w PDB. Ze względu na olbrzymią liczbę dostępnych struktur uwzględnione zostały tylko struktury białek o homologii poniżej

40% zdeponowane w okresie 6 lat do wiosny 2013 roku i spełniające odpowiednie kryteria jakościowe wyrażone przez wskaźniki R i Rfree. Ze zbioru przeszło 438 tys. reszt doktorant uzyskał nieco ponad 83 tys. nieredundantnych konformerów stanowiących bazę do stworzenia dalszych uproszczonych bibliotek.

Kolejnym etapem pracy było połączenie w jedną całość procedur dopasowania bazujących na metodach wykorzystywanych przez inne programy takie jak *ARP/wARP*, *Buccaneer* oraz *RESOLVE*. Zastosowane zostały: dopasowanie konformera z biblioteki w oparciu o kryterium energetyczne i probabilistyczne, eliminacja kolizji z użyciem algorytmów A* i DEE oraz optymalizacja dopasowania z uwzględnieniem oddziaływań elektrostatycznych i van der Waalsa pomiędzy sąsiednimi łańcuchami. Dwuetapowy opis tego algorytmu przedstawiony w streszczeniu różni się nieco od strony formalnej od opisu trójetapowej procedury przedstawionej na rysunku 3 i w rozdziale "Materials and methods". Tym niemniej, oba opisy obejmują takie same metody obliczeniowe co pozwala przypuszczać, że jest to tylko formalna niekonsekwencja a nie dwa różne algorytmy.

Znaczna część rozprawy poświęcona jest opisowi metod optymalizacji pięciu parametrów programu oraz testowania nowego algorytmu. W tym celu wykorzystano łącznie 550 struktur treningowych i 1100 struktur testowych o dobranych rozdzielczościach. Parametry były optymalizowane dla każdego z osiemnastu aminokwasów (z wyłączeniem glicyny i alaniny) i każdego z jedenastu przedziałów rozdzielczości. Jako kryterium oceny przyjęta zastał procentowy wzrost liczby konformerów wmodelowanych w pozycje odległe średnio o mniej niż 0.5 Å od pozycji oryginalnych. Biorąc pod uwagę liczbę użytych struktur, przedziałów rozdzielczości i typów łańcuchów bocznych, wymagało to wykonania olbrzymiej pracy choćby tylko w celu przygotowania obliczeń. O złożoności tego zadania najlepiej świadczą zamieszczone głownie w dodatkach spisy struktur treningowych i testowych oraz ponad 160 wykresów obrazujących wpływ parametrów na jakość dopasowania każdego typu aminokwasu.

Struktury użyte w zbiorach treningowych i testowych pochodziły z Protein Data Banku i przeszły przez wcześniejsze sito weryfikacji poprawności. Nie można wykluczyć, że wśród nich znajdowały się struktury zawierające nieprawidłowe konformacje łańcuchów bocznych, jednak stanowiły przypuszczalnie niewielki odsetek a zatem ocena skuteczności działania programu poprzez zgodność konformacji wmodelowanej i wyznaczonej jest uzasadniona. Ciekawszym zagadnieniem jest jednak zastosowanie programu *Fitmunk* do automatycznej

poprawy struktury na wcześniejszym etapie udokładniania i wyznaczenie wpływu wprowadzonych zmian na krystalograficzne wskaźniki poprawności, takie jak między innymi R i Rfree. Test taki został wykonany dla zbioru 36 struktur i zaobserwowano poprawę wskaźników średnio o 1.3%. Największa poprawa obserwowana była, jak łatwo przewidzieć, dla struktur których modele pochodzące z podstawienia molekularnego nie były wcześniej przebudowywane. Ciekawszym wynikiem jest poprawa, choć często nieznaczna, wskaźników dla struktur wcześniej przebudowanych przy użyciu innych programów. Świadczy to o skuteczności programu *Fitmunk* i jego pewnej przewadze nad podobnymi programami.

Oprócz wynajdywania optymalnych konformerów, program *Fitmunk* może być wykorzystywany również do rozpoznawania sekwencji łańcucha polipeptydowego. Ta interesująca możliwość może być bardzo cenna w przypadku struktur nowych białek pozyskiwanych z naturalnych źródeł, gdyż często ich sekwencja nie była wcześniej wyznaczana lub jest niekompletna. Doktorant wskazuje też na możliwość zastosowania programu w przypadku gdy przypisana znana sekwencja jest przesunięta w stosunku do rzeczywistej sekwencji łańcucha polipeptydowego.

Ponieważ "produktem końcowym" pracy jest program komputerowy *Fitmunk*, spodziewałem się, że rozprawa będzie zawierała jakieś szczegóły dotyczące jego budowy, wymagań systemowych oraz dystrybucji programu. Niestety, gdyby nie to, że znaczna część rozprawy to prezentacja wyników obliczeń uzyskanych przy pomocy tego programu, można by łatwo przegapić fakt jego powstania. W rozprawie wymieniony jest tylko język programowania - *Python* oraz użyte biblioteki. Dopiero w ostatnim zdaniu rozprawy, a właściwie jej podsumowania, podana jest informacja o serwerze umożliwiającym zdalne użycie programu. Skorzystałem z takiej możliwości aby przekonać się, że rzeczywiście program działa, jednak na obecnym etapie, czyli przypuszczalnie do czasu opublikowania wyników, jest to serwis wyłącznie dla "wtajemniczonych".

Podsumowując, rozprawa jest przygotowana starannie aczkolwiek w kilku miejscach pozostawiono puste miejsca nawet na blisko pół strony. Wskazuje to na daleko idące przeróbki tekstu na końcowym etapie przygotowywania rozprawy, już po jej wydrukowaniu. Nie chciał bym się odnosić w recenzji szczegółowo do poprawności językowej pracy gdyż nie uważam się za specjalistę w tym zakresie a jedynie użytkownika. Generalnie język jest poprawny, aczkolwiek sporadycznie spotyka się nieprawidłowo napisane wyrazy a w wielu

miejscach użył bym innych konstrukcji gramatycznych, tym niemniej uważam, że tekst jest zrozumiały i jednoznaczny.

Rozprawa zawiera istotne elementy nowości naukowej, nie ogranicza się zebrania znanych metod i zaimplementowania ich w jeszcze jednym programie. Doktorant opracował oryginalną bibliotekę konformerów, dokonał optymalizacji parametrów istotnych dla efektywnego działania algorytmu i przetestował ich skuteczność na setkach struktur. Opracowany w ramach niniejszej rozprawy algorytm dopasowywania konformerów można uznać za skuteczniejszy od dotychczas dostępnych. Ponadto, może on być wykorzystywany również do automatycznego rozpoznawania sekwencji aminokwasowej nowego białka w oparciu o mapę gęstości elektronowej. Program może być (i pewnie będzie) dalej rozwijany, doktorant sygnalizuje taką możliwość szczególnie w zakresie wpasowywania alternatywnych konformacji, jednak już w obecnej postaci stanowi cenne narzędzie zasługujące na szerokie rozpowszechnienie.

W świetle powyższego, uważam, że rozprawa spełnia warunki określone w art. 13 ustawy z dnia 14 marca 2003 w stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. z 2003 r., nr 65 poz. 595 z późniejszymi zmianami) i stawiam wniosek o dopuszczenie pana mgr Przemysława Porębskiego do dalszych etapów przewodu doktorskiego.

Równocześnie, biorąc pod uwagę zarówno ogrom włożonej pracy jak też jej wysoki poziom a także potencjalnie istotny wpływ opracowanej procedury na rozwój metod współczesnej krystalografii makrocząsteczek, wnioskuję o wyróżnienie przedstawionej rozprawy.



prof. dr hab. Krzysztof Lewiński