Dr hab. inż. Marta Szachniuk
Prof. nadzw. IChB PAN
Instytut Informatyki
Politechnika Poznańska
mszachniuk@cs.put.poznan.pl

Poznań, 12.12.2017

Recenzja Rozprawy Doktorskiej

Tytuł: Supporting protein tertiary structure modelling with contact sites prediction methods

Autor: mgr inż. Paweł P. Woźniak

Promotor: dr hab. inż. Małgorzata Kotulska, prof. nazw. PWr

Kopromotor: prof. Gert Vriend

Tematyka badawcza

Recenzowana rozprawa dotyczy jednego z wiodących, a jednocześnie jednego z pierwszych problemów bioinformatyki struktur molekularnych jakim jest przewidywanie trójwymiarowej struktury białek za pomocą metod obliczeniowych. Powstające metody predykcji struktur białkowych próbują odpowiedzieć na pytanie o relację między kształtem cząsteczki a jej funkcją oraz umożliwić opisanie procesu zwijania się białek do ich formy natywnej, preferowanej energetycznie. Nadrzędnym celem rozwoju metod predykcji struktur jest jednak nie tylko zaspokojenie potrzeb poznawczych, ale – przede wszystkim – opracowanie mechanizmów pozwalających na modelowanie cząsteczek zwijających się do postaci zdefiniowanej *a priori* i pełniących określone funkcje. Dokładne metody modelowania struktur znacząco przyczynią się do postępu biomedycyny, zwłaszcza w zakresie diagnostyki medycznej oraz nowoczesnych metod leczenia chorób.

Pierwsze próby komputerowego modelowania białek były podejmowane już pod koniec lat 60-tych XX wieku (Scheraga, 1969). W 1973 roku Christian Anfinsen sformułował hipotezę termodynamiczną (znaną też jako dogmat Anfinsena) mówiącą, iż sekwencja białka jednoznacznie determinuje jego natywną strukturę przestrzenną i jest to prawdziwe przynajmniej dla małych cząsteczek globularnych. Wedle tej, nagrodzonej nagrodą Nobla, hipotezy, struktura natywna odpowiada minimum energii swobodnej układu, w którym znajduje się białko. Hipoteza stała się przyczynkiem do projektowania algorytmów

generujących struktury 3D białek na podstawie sekwencji aminokwasowej. Dodatkową motywacją dla rozwoju metod komputerowych były wysokie koszty eksperymentalnego rozwiązywania struktur, trudność w badaniu białek drogą eksperymentalną i niezadowalająca jakość wyników takich badań. Rozwój technologii informatycznych, coraz większa dostępność dużych mocy obliczeniowych oraz uruchomienie w 1994 r. konkursu CASP, Critical Assessment of Methods of Protein Structure Prediction (Moult *et al.*, 1995) poskutkowało znacznym przyspieszeniem badań nad modelowaniem białek *in silico*, powstaniem nowych algorytmów oraz zwiększeniem jakości i dokładności uzyskiwanych modeli strukturalnych.

Obecnie istnieje wiele metod przewidywania struktury trzeciorzędowej białek, rozwijanych w ramach najpopularniejszych podejść: predykcji *de novo*, modelowania komparatywnego, bądź połączenia tych dwu. Jednak jakość i dokładność modeli białkowych generowanych *in silico*, choć wysoka, wciąż jeszcze nie pozwala na ich bezpośrednie wykorzystanie np. w medycynie molekularnej. Dlatego dziś kładzie się duży nacisk już nie na nowe algorytmy predykcji, ale na rozwój metod, które – zintegrowane z algorytmami predykcji – pozwolą na znaczące udokładnienie przewidywanych trójwymiarowych modeli białek oraz odfiltrowanie modeli nieprawidłowych. Jednej z takich metod została poświęcona recenzowana rozprawa doktorska.

Zakres pracy i wkład autora

W ramach pracy badawczej, autor recenzowanej rozprawy skupił się na analizie miejsc kontaktowych oraz ich wykorzystaniu w komputerowej predykcji struktur białkowych. Miejsce kontaktowe definiowane jest jako para aminokwasów oddalonych w sekwencji o co najmniej cztery pozycje, których atomy C_β leżą od siebie nie dalej niż w odległości 8Å. Aminokwasy te pozostają w kontakcie, co oznacza, że mutacja jednego z nich powoduje automatycznie mutację drugiego. Występowanie takiej zależności jest warunkiem koniecznym uznania pary aminokwasów za miejsce kontaktowe. Miejsca te uważa się za punkty w strukturze białka, które mają istotne znaczenie dla jej zachowania. Metody, które przewidują strukturę białka na podstawie miejsc kontaktowych działają dwuetapowo. W pierwszym etapie dokonują predykcji miejsc kontaktowych na podstawie sekwencji białkowej, a następnie odtwarzają całą strukturę białka poprzez obudowanie miejsc kontaktowych pozostałymi aminokwasami.

Istnieje szereg metod przewidujących miejsca kontaktowe w strukturze białek. Największą efektywnością wykazują się te, które wykorzystują analizę sprzężeń bezpośrednich

(DCA) mutacji zależnych oraz wielosekwencyjne dopasowanie (MSA) sekwencji aminokwasowej badanego białka. Ich skuteczność oscyluje wokół 40% na sto przewidzianych miejsc kontaktowych, co nie jest zadowalającym wynikiem w przypadku, gdy zidentyfikowane miejsca mają stanowić fundament dla procedury odbudowującej całą strukturę.

Autor niniejszej rozprawy postawił sobie za cel opracowanie metod obliczeniowych, które poprawią efektywność algorytmów przewidujących miejsca kontaktowe, a jednocześnie zwiększą wiarygodność przewidzianych miejsc. W drodze do osiągnięcia w/w celu zaplanowano i zrealizowano kilkuetapowe badania, których wymiernym efektem było zwiększenie skuteczności predykcji o 5% dla kontaktów przewidzianych z największym prawdopodobieństwem. Dodatkowo opracowano metodę pozwalającą na odfiltrowanie fałszywych modeli białka na podstawie miejsc kontaktowych oraz metodę szacującą skuteczność predykcji kontaktów dla konkretnego białka.

W ramach badań sformułowano i zweryfikowano trzy oryginalne hipotezy badawcze:

1. Wykorzystanie charakterystyki miejsc kontaktowych poprawia skuteczność ich predykcji.
2. Przewidziane miejsca kontaktowe pozwalają odróżnić strukturę prawidłową białka od jego fałszywego modelu.
3. Dla zadanego białka można przewidzieć skuteczność predykcji miejsc kontaktowych.

W celu zweryfikowania pierwszej z postawionych hipotez autor przeanalizował charakterystykę kontaktów z uwzględnieniem miejsca kontaktowego oraz grupy strukturalnej białka. Następnie wyselekcjonował parametry opisujące związek między typami aminokwasów mogących stanowić kontakt a klasami i topologią struktur białkowych. Wykorzystując wyniki powyższej analizy, autor stworzył algorytm dopasowujący charakterystykę kontaktów danego białka do charakterystyki skojarzonej z klasą tego białka. Algorytm ten, zintegrowany z metodą predykcji mfDCA poprawił skuteczność predykcji kontaktów średnio o 5%. Weryfikacja drugiej hipotezy nastąpiła w drodze porównania liczby poprawnie przewidzianych kontaktów dla modeli białkowych o targetach zdeponowanych w Protein Data Bank. Kontakty zostały przewidziane algorytmem gplmDCA. Wykazano 100% skuteczności w identyfikowaniu fałszywych modeli w przypadku instancji zaprojektowanych na potrzeby eksperymentu. Podczas weryfikacji trzeciej hipotezy autor zaprojektował modele regresyjne służące do oszacowania jaka jest skuteczność predykcji kontaktów w konkretnej strukturze białkowej. Model przewidział skuteczność kliku algorytmów z dość wysoką dokładnością pokazując, że można w ten sposób uwiarygodnić przewidziane kontakty.

Ocena strony merytorycznej

Recenzowana rozprawa doktorska ma formę spójnego tematycznie zbioru artykułów opublikowanych w czasopismach naukowych z dyscypliny biocybernetyki, bioinformatyki oraz biologii obliczeniowej. Najważniejszą część rozprawy stanowią trzy artykuły naukowe, których pierwszym autorem jest mgr inż. Paweł P. Woźniak. Pierwszy artykuł został opublikowany w 2014 r. w *Journal of Molecular Modeling* (czasopismo w III kwartylu; Impact Factor=1,425; Punkty MNiSW=20). Dwa pozostałe artykuły opublikowano w 2017 r. w dwutygodniku *Bioinformatics* (czasopismo w I kwartylu; Impact Factor=7,307; Punkty MNiSW=45). Fakt, iż uzyskane wyniki badawcze zostały zaakceptowane do publikacji w wiodącym czasopiśmie z dziedziny, jakim jest *Bioinformatics*, świadczy, iż autorzy opublikowanych artykułów uprawiają naukę na najwyższym światowym poziomie, a opracowane przez nich metody oraz rezultaty ich zastosowań są bardzo wartościowe.

Merytoryczna strona rozprawy nie budzi żadnych zastrzeżeń. Oceniam ją bardzo wysoko. Temat pracy został trafnie dobrany, a rozpatrywany problem badawczy sformułowany poprawnie i zrozumiale. Autor pracy we właściwy sposób zdefiniował cel prowadzonych badań oraz wyczerpująco objaśnił sposób realizacji zadań badawczych. Wybór metod badawczych wskazuje na dobre merytoryczne przygotowanie autora rozprawy do podjęcia zaplanowanych działań. Postawione zostały trzy hipotezy, które potwierdzono w 100% stosując metody obliczeniowe i analityczne. Trzy artykuły naukowe stanowiące główną część pracy zawierają kompletny i poprawny merytorycznie materiał naukowy. W streszczeniach stanowiących dodatek do publikacji zauważyłam jedną usterkę, mianowicie definicje miejsca kontaktowego w wersji angielskiej i polskiej różnią się nieznacznie: w streszczeniu polskim napisano, że „Miejsce kontaktowe jest (...) oddzielone w sekwencji przynajmniej czterema aminokwasami.", natomiast w streszczeniu angielskim podano, iż „A residue-residue contact is (...) separated in sequence by more than 4 residues." („more than 4" oznacza więcej niż 4, czyli co najmniej 5, a nie co najmniej 4 jak napisano w wersji polskiej). Brakuje też nieco obszerniejszego przedstawienia dalszych pomysłów i planów badawczych. Autor oczywiście podaje, iż zostało jeszcze wiele do zrobienia w kwestii poprawienia skuteczności predykcji miejsc kontaktowych jest to jednak wypowiedź natury bardzo ogólnej.

Ocena strony redakcyjnej

Główną część rozprawy stanowią trzy artykuły naukowe opublikowane w *Journal of Molecular Modeling* (listopad 2014) oraz *Bioinformatics* (maj 2017, czerwiec 2017). Dodatkowo praca zawiera 8 krótkich rozdziałów, w których przedstawiono dorobek autora, wykaz cytowanych publikacji, obszerne streszczenie wyników badań w języku angielskim oraz krótkie streszczenie w języku polskim. Z wyjątkiem ostatniego z wymienionych rozdziałów, praca została spisana po angielsku. Język pracy jest poprawny i zrozumiały. Tekst opatrzony został kolorowymi ilustracjami, wykresami i tabelami. Całość zamyka się w 84 stronach. Struktura pracy nie budzi zastrzeżeń. Źródła informacji zostały prawidłowo wybrane i zacytowane. Poniżej wymieniono zauważone usterki redakcyjne i typograficzne:

1. W pracy nie zachowano jednolitej interlinii. Na stronach 7-10 oraz na stronie 23 zastosowano interlinię 1,5 wiersza, natomiast w pozostałej części pracy posłużono się odstępem na 1 wiersz. Przy tym większy odstęp występuje w miejscach, gdzie wymienione są publikacje i osiągnięcia autora pracy oraz jej główne tezy. Sprawia to wrażenie jakby autor chciał nadrobić niewystarczającą (w swoim mniemaniu) liczbę osiągnięć długością akapitu. Chcę w tym miejscu zaznaczyć, iż według mnie autor pracy ma bardzo dobre osiągnięcia naukowe i stosowanie takich wybiegów uważam za niepotrzebne.
2. Str. 9: W tytule ostatniej wymienionej na tej stronie publikacji jest błąd. Artykuł nosi tytuł „Database of Peptides …", a nie „Databases of Peptides …".
3. Zgodnie z zasadami składu tekstu nie należy stosować spacji przed znakami interpunkcji (str. 9-10, wstawiono spacje przed dwukropkiem).
4. Str. 9-10: W wykazie publikacji nie stosuje się jednolitego formatowania cytowanych prac, np. w różnorako podane jest miejsce konferencji („Wrocław, 22-25 September 2016" oraz „3-6 October – Będlewo"), różnie oznaczane są strony („s. 19-24" oraz „pp: 20-27").
5. Str. 35-41: W wykazie cytowanej literatury nie zastosowano jednolitego formatowania:
   a. Nazwy czasopism pisane są w jednym miejscu z użyciem oficjalnych skrótów, a w innym – pełną nazwą, np.: „Nucleic Acids Research" (poz. 4) oraz "Nucleic Acids Res" (poz. 5).
   b. W zapisie skróconych nazw czasopism raz stosuje się kropki, a innym razem nie (lub stosuje się wybiórczo), np.: "J. Mol. Biol." (poz. 1), „J Mol Biol." (poz. 2) oraz „JMol Biol" (poz. 8).

c. Przy niektórych pozycjach podany jest numer doi, przy innych go nie ma (mimo, iż numer ten jest nadany).
d. Po nazwie czasopisma występuje kropka, przecinek lub nie ma żadnego znaku interpunkcyjnego, np.: „Bioinformatics. 27(11):1573-4", „Protein Eng, 6:593-604", „Proteins 15:191-204".
e. Wolumen, numer i strony formatowane są dowolnie, np.: „27(11):1573-4", „33(10), 1497-1504", „76:176-183".
f. Wykaz publikacji jest nieporządkowany stylistycznie. Tytuły większości publikacji zapisuje wg reguły: pierwszy wyraz wielką literą a pozostałe małą. Jednak w kilku miejscach z nieuzasadnionych przyczyn zastosowano: wszystkie wyrazy wielką literą.

6. Str. 40: Trzy ostatnie prace podane na tej stronie są to artykuły będące podstawą recenzowanej pracy doktorskiej. Dwie z nich oznaczono literami (A), (B), trzeciej nie oznaczono wcale. Prawdopodobnie oznaczenia literowe miały być zgodne z oznaczeniami podanymi na stronie 7, ale nie są.
7. Str. 43: Jest „ich występowania (...) może być wskazane", powinno być „ich występowanie (...) może być wskazane".
8. Str. 43: Jest „skuteczność predykcji metod" a lepiej brzmiałoby „skuteczność predykcji" lub „skuteczność metod". W obecnej formie można to odczytać, iż chodzi o przewidywanie metod.
9. W wielu miejscach na końcu linii pojawiają się spójniki, co jest dość typowym, a jednocześnie łatwym do uniknięcia błędem typograficznym.

Powyższe usterki nie mają znaczącego wpływu na czytelność pracy i nie umniejszają jej wartości. Nie zmieniają również mojej ogólnej wysokiej oceny recenzowanej dysertacji.

Wnioski końcowe

Uważam, iż autor recenzowanej pracy wykazał się umiejętnością poprawnej i przekonywującej prezentacji wyników przeprowadzonych badań oraz trafnością wnioskowania. Dowiódł, iż w wystarczającym stopniu poznał dotychczasowy stan wiedzy o podjętym temacie (przewidywanie struktur białkowych z wykorzystaniem miejsc kontaktowych), przedstawiany w przedmiotowej literaturze krajowej i zagranicznej. Posiada również ogólną wiedzę teoretyczną w dyscyplinie Informatyki, Biocybernetyki i Inżynierii Biomedycznej oraz wykazuje się umiejętnością samodzielnego prowadzenia pracy naukowej

Recenzowana praca zawiera oryginalne rozwiązanie problemu naukowego. Uzyskane przez autora wyniki badań zostały opublikowane w wiodącym czasopiśmie z dziedziny. Autor wielokrotnie prezentował je również na konferencjach naukowych w kraju i zagranicą, zyskując uznanie i liczne nagrody za prezentacje. Pracę oceniam bardzo wysoko i składam wniosek o jej wyróżnienie.

Stwierdzam, że praca pt. „Supporting protein tertiary structure modelling with contact sites prediction methods" spełnia wymagania stawiane rozprawom doktorskim określone w aktualnie obowiązującej w Polsce Ustawie o Stopniach Naukowych i Tytule Naukowym (Art. 13) i stanowi oryginalne rozwiązanie przez autora zagadnienia naukowego.

Marta Szachniuk