Dr hab. inż. Szymon Wąsik
Google Polska
szymon.wasik@gmail.com

4 stycznia 2022

Recenzja rozprawy doktorskiej

Mgr Joanny Miśkiewicz

zatytułowanej:

Bioinformatics Methods of Motif Analysis in RNA Structure

1. Problem badawczy i jego znaczenie

Rozprawa doktorska oparta jest o zbiór sześciu artykułów naukowych, z których 5 zostało już opublikowanych w czasopismach z listy WoS. Wspólnym motywem tematycznym łączącym wszystkie artykuły jest analiza motywów strukturalnych w cząsteczkach RNA. Na podkreślenie zasługuje wysoki sumaryczny IF czasopism (29.060) oraz fakt, że w większości artykułów autorka rozprawy jest pierwszym autorem. W zbiorze tym szczególnym osiągnięciem jest publikacja w czasopiśmie *Briefings in Bioinformatics*, w którym publikowane są jedynie wybitne publikacje przeglądowe, a znalezienie się w gronie autorów wymaga bardzo obszernej wiedzy i ogromnej pracy. Jak już wspomniano powyżej, tematyka artykułów dotyczy metod analizy motywów biologicznych i przedstawia badania, które zostały przeprowadzone dla konkretnych organizmów w oparciu o konkretne metody bioinformatyczne. Tak więc rozwiązywane problemy badawcze zostały zdefiniowane na bardzo szczegółowym poziomie (na poziomie poszczególnych organizmów i określonych motywów strukturalnych), co może być nie do końca jasne po przeczytaniu tytułu rozprawy, który jest bardzo ogólny i odnosi się do metod bioinformatycznych w analizie motywów strukturalnych bez żadnego uszczegółowienia. Nie umniejsza to w żaden sposób wartości badań, uważam jedynie, że tytuł rozprawy może być trochę mylący. Może lepszym tytułem byłoby na przykład *Bioinformatics Methods for Solving Selected Problems of RNA Structure Analysis.*

Tematyka przedstawionych badań może być zaklasyfikowana do trzech grup tematycznych. Pierwsza dotyczy analizy motywów występujących w roślinnych mikroRNA. Autorce udało się zaobserwować schemat powtarzających się małych pętli wewnętrznych w *Arabidopsis thaliana*, a następnie rozszerzyć badania na całe królestwo roślin zielonych. Kolejna grupa tematyczna, to analiza sekwencji pochodzenia ludzkiego. W badaniach tych przeanalizowano jak różne motywy występujące w sekwencjach mogą wpływać na ich funkcjonowanie w kompleksach z białkami. W końcu najnowsze badania dotyczyły analizy motywu kwadrupleksów. Ich wynikiem jest opracowanie nowych formatów oraz klasyfikacji kwadrupleksów oraz bazy danych do ich przechowywania i wyszukiwania.

Najlepszym potwierdzeniem naukowego charakteru rozprawy jest fakt, że jej wyniki zostały opublikowane w pięciu czasopismach naukowych z listy WoS, która obejmuje jedynie wiodące, światowe czasopisma naukowe, w których nie ma możliwości opublikowania artykułów o innym charakterze. Potwierdza to również przedstawiona w każdym artykule analiza istniejącego stanu wiedzy oraz szczegółowa bibliografia.

Jak już wspomniano powyżej, problematyka badawcza ma zdecydowanie praktyczne zastosowanie. Poszczególne artykuły przedstawiają wyniki badań dla konkretnych sekwencji, dla konkretnych, rzeczywistych organizmów i prezentują wyciągnięte na podstawie przeprowadzonych analiz wnioski

biologiczne. Każde z omówionych badań może być w przyszłości zaadoptowane to przeprowadzenia podobnych analiz dla innych sekwencji i wyciągnięcia kolejnych wartościowych wniosków.

Wkład autora

Ponieważ wszystkie przedstawiony artykuły są wieloautorskie, a rozdział 2 podsumowujący wyniki badań jest pisany z punktu widzenia osiągnięć wszystkich autorów, w niniejszej sekcji chciałbym odnieść się do wkładu autorki na podstawie informacji o wkładzie w badania przedstawionej na stronach vi-vii. W zasadzie we wszystkich przeprowadzonych badaniach wkład autorki widoczny jest na wszystkich etapach prowadzenia badań – od zebrania danych przez opracowanie metod i algorytmów po ich analizę i opracowanie manuskryptów. Najniższe zaangażowanie widać na etapie opracowywania założeń dla przeprowadzanych badań oraz na etapie wyciągania wniosków praktycznych. Jest to w pełni zrozumiałe na etapie studiów doktoranckich, gdy etapy te muszą być w znacznym stopniu inspirowane i wspomagane przez bardziej doświadczonych naukowców z bardziej rozległą wiedzą. Inne niż powyższe znaczące kontrybucje innych autorów ograniczają się do współpracy przy projektowaniu algorytmu opisanego w publikacji A1 oraz przy projektowaniu bazy danych opisanej w A6. Jednak w tym drugim przypadku na uwagę zasługuje wejście w rolę kierowniczą i mentorowanie dwóch współautorów implementujących bazę danych według opracowanego projektu. Taki wpływ na innych zdecydowanie wykracza poza wymagania stawiane doktorantom. Podsumowując, uważam że wkład autorki jest w zupełności wystarczający do obrony niniejszej rozprawy autorskiej, a ze względu na liczbę i objętość publikacji (w sumie ponad 100 stron dość skoncentrowanego tekstu) nawet przewyższa wymagania stawiane zwyczajowo doktorantom.

2. Poprawność

Ponieważ przedstawiona rozprawa jest podsumowaniem zbioru opublikowanych w większości artykułów naukowych (z wyjątkiem A6) najtrudniejsze zadanie weryfikacji poprawności badań przedstawionych w rozprawie zostało wykonane już przez zewnętrznych ekspertów recenzujących poszczególne artykuły. Ponieważ wszystkie artykuły zostały opublikowane w czołowych czasopismach naukowych (wszystkie znajdują się na liście WoS) można było podejrzewać, że jakość recenzji była bardzo wysoka. Potwierdza to analiza przeprowadzona przez autora tej recenzji, który doszukał się jedynie kilku drobnych nieścisłości i niedomówień, które nie mają jednak istotnego wpływu na poprawność rozprawy. Większość z nich wystąpiła zresztą w rozdziale 1 omawiającym podstawy teoretyczne badań. Są one następujące:

1. Myślę, że w kontekście bardzo dynamicznego rozwoju bioinformatyki, a w szczególności obliczeniowych metod analizy struktur biologicznych stwierdzenie, że „wiedza o nich jest głównie czerpana z eksperymentów laboratoryjnych” (str. 1) jest bardzo mocne i jeżeli faktycznie jest zgodne z przekonaniem autorki, to powinno być lepiej uzasadnione.
2. Informacja o tym, że obliczeniowe modelowanie struktur 3D kwasów nukleinowych jest cały czas wyzwaniem podparta jest publikacją sprzed dekady (str. 5) co w kontekście bioinformatyki jest okresem niezwykle długim, w czasie którego rozwiniętych mogło być bardzo wiele nowych metod. W szczególności zastosowanie głębokiego uczenia maszynowego pozwoliło w ostatnich czasach znacząco podnieść jakość predykcji, również dla nowych struktur. Przydałaby się w tym miejscu dyskusja najnowszych osiągnięć biologii obliczeniowej.
3. Zdanie „In the relation to the recent threat situation caused by SARS-CoV-2 virus, we can also search for epidemic motifs” może sprawiać wrażenie jakby poszukiwanie motywów

epidemiologicznych nabrało znaczenia dopiero w kontekście aktualnej pandemii, co nie do końca jest prawdą. Myślę, że poprawniej byłoby napisać, że od dłuższego czasu metody te są już rozwijane (choćby przytoczona publikacja z 2009 roku), a w ostatnim czasie nastąpił ich szczególny rozwój w kontekście pandemii COVID (przydałaby się jakaś referencja).

4. Trochę nie rozumiem zdania „Each motif search algorithm is based on a specific data format, thus we need to bear in mind that outputs (patterns), for even the same data set, can be different.". Z jakiego powodu wyniki mogą być inne? Co ma z tym wspólnego to, że każdy algorytm może wymagać danych w trochę innym formacie? Myślę, że mógłbym znaleźć co najmniej kilka możliwych odpowiedzi na każde z tych pytań, ale warto by było, żeby autorka rozprawy uszczegółowiła co miała na myśli.
5. Warto byłoby podać statystyki odwiedzin opracowanej bazy ONQUADRO – ile osób odwiedza ją miesięcznie i ile wykonuje wyszukiwań. Byłoby to wartościowe potwierdzenie użyteczności bazy, szczególnie w kontekście braku ostatecznej publikacji opisującego ją artykułu.

Oprócz powyższych nieścisłości udało mi się znaleźć również drobne błędy edycyjne i językowe. Należy jednak podkreślić że są one bardzo nieliczne, a cała rozprawa została zredagowana w sposób bardzo przejrzysty, czytelny i staranny:

1. „The above suggestions" – myślę, że lepszym słowem byłoby „examples", bo powyższe formaty są przykładami, a nie sugestiami.
2. W stwierdzeniu „A 2D structure is often represented" słowo "often" jest dosyć nieprecyzyjne. Czy oznacza ono, że to najpopularniejsza metoda? Jeżeli tak, to dlaczego nie została opisana jako pierwsza?
3. Myślę, że słowo „proposition" lepiej byłoby zastąpić przez „proposal" (str. 9) bo mówimy tu o pewnej sugestii zgłoszonej jako propozycja, a nie konkretnym kontrakcie.
4. „atom distribution" zamiast „atoms distribution" (str. 11).
5. „An example (…) networks are" zamiast "An example (…) networks is" (str. 14).
6. "can contribute to find" zamiast "to finding" (str. 14).
7. "mismatch" (str. 17) – nie powinno się tworzyć listy wypunktowanej jeżeli ma się składać tylko z jednego elementu.

3. Wiedza kandydata

Ogólny stan wiedzy autora w zakresie bioinformatyki potwierdza rozdział 1 zaprezentowanej rozprawy (21 stron tekstu). Omawia on szczegółowo w kolejnych sekcjach podstawy biologiczne tematyki poruszonej w rozprawie (RNA i jego rola) oraz bioinformatyczne (formaty reprezentacji RNA, zastosowanie analizy motywów w naukach o życiu oraz bardziej szczegółowe omówienie motywów strukturalnych oraz metod ich znajdowania). Przedstawiony opis jest wyczerpujący, czytelnie podzielony na metody stosowane dla każdego poziomu szczegółowości reprezentacji struktury (pierwszo-, drugo- i trzeciorzędowa) i z całą pewnością potwierdza, że autorka dobrze zna obszar badawczy, który porusza w rozprawie (należący do dyscypliny Informatyka techniczna i telekomunikacja). Świadczy o tym fakt, że z drobnymi, mało istotnymi wyjątkami (patrz poniżej) rozdział ten porusza podstawy teoretyczne wszystkich omawianych w rozprawie zagadnień.

Bibliografia zawiera 152 dobrze dobrane pozycje naukowe, związane tematycznie z tematyką rozprawy. Sformatowana jest w spójny, czytelny sposób. Co jednak najważniejsze, każde stwierdzenie przedstawione w pracy podparte jest odpowiednią pozycją z bibliografii, co niestety nie zawsze jest pewnikiem nawet w artykułach opublikowanych w dobrych czasopismach naukowych.

Jedyna tematyka, której zabrakło mi w rozdziale pierwszym (konkretnie 1.3) to:

1. Definicja czym jest motyw. Doceniam szeroki przegląd obszarów i dyscyplin nauk o życiu, które zajmują się wykrywaniem i analizą motywów, ale warto byłoby jasno podkreślić, że każda z nich może definiować motyw w nieco inny sposób i podeprzeć to stwierdzenie kilkoma przykładowymi definicjami. W szczególności, biorąc pod uwagę dyscyplinę pracy, warto byłoby podać matematyczną definicję choćby najprostszego motywu strukturalnego.
2. Dedykowanej podsekcji w sekcji 1.3, która opisałaby praktyczne zastosowania metod analizy modeli (czyli po co te metody są właściwie rozwijane).

4. Inne uwagi[1]

Brak.

5. Podsumowanie

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez artykuł 13 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym (z późniejszymi zmianami)[2] moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problem naukowego? (wybierz jedną opcję stawiając znak **X**)

[x]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie **Informatyka techniczna i telekomunikacja**?

[x]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej?

[x]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

Ponadto, biorąc pod uwagę wysoką rangę czasopism naukowych, w których opublikowane zostały artykuły (szczególnie A4 i A5) rekomenduję wyróżnienie rozprawy doktorskiej[3].

Podpis

[1] Opcjonalnie
[2] http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf
[3] Oczywiście to zdanie jest opcjonalne.