dr hab. inż. Witold Dyrka
Politechnika Wrocławska
witold.dyrka@pwr.edu.pl

6.12.2022

Recenzja rozprawy doktorskiej
mgr. Jakuba Wiedemanna
zatytułowanej:
Combinatorial Analysis of RNA Tertiary Structures with the Use of Angular Representations

1. Problem badawczy i jego znaczenie

Rozprawa Pana mgr. Jakuba Wiedemanna dotyczy metod i narzędzi analizy *in silico* cząsteczek kwasów rybonukleinowych (RNA). Tematem przewodnim rozprawy jest efektywne wykorzystanie reprezentacji kątowej do analizy trzeciorzędowej struktury RNA. Funkcjonalność biopolimerów takich jak białka i kwasy rybonukleinowe ściśle zależy od ich struktury przestrzennej. Jednocześnie uzyskanie takiej struktury metodami laboratoryjnymi jest dużo trudniejsze i wiąże się z większym nakładem środków w stosunku do nieomal rutynowo dziś wykonywanego sekwencjonowania. Warunkuje to konieczność stosowania komputerowego przewidywanie struktury przestrzennej na podstawie sekwencji. O ile ostatnie lata przyniosły bardzo duży postęp w przewidywaniu struktury trzeciorzędowej białek, dzięki dostępności dużej ilości danych eksperymentalnych, rozwojowi metod maszynowego uczenia, ale także zorganizowanemu wysiłkowi społeczności naukowej, to określenie struktury trzeciorzędowej RNA zwykle pozostaje wyzwaniem. Tym samym podjęty przez mgr. Jakuba Wiedemanna problem badawczy pozostaje jednym z największych wyzwań bioinformatyki AD 2022. Można się spodziewać, że podobnie jak stało się w przypadku białek, kompleksowe rozwiązanie problemu przewidywania struktury trzeciorzędowej RNA miałoby przełomowe znaczenie dla biologii i medycyny. W tym procesie zapewnienie odpowiedniej reprezentacji sekwencji, adekwantych miar jakości rozwiązań, oraz aktualnych zasobów – co jest bezpośrednim przedmiotem omawianej rozprawy – ma bardzo duże znaczenie praktyczne, dostarczając jednocześnie bardzo przydatnych narzędzi dla biologów obliczeniowych do stosowania w analizie konkretnych przypadków.

2. Wkład autora

Jakkolwiek tezy rozprawy nie zostały *explicite* sformułowe (co ma tę zaletę, że uniknięto dzięki temu dość powszechnej sztuczności tego typu konstrukcji), to już na podstawie streszczenia rozprawy można wskazać trzy zasadnicze elementy wkładu Autora w rozwiązanie problemu badawczego: 1) pokazanie, że porównanie lokalnych motywów strukturalnych ma zasadnicze znaczenie dla oceny podobieństwa cząsteczek RNA, 2) opracowanie metody znajdowania najdłuższych ciągłych segmentów o wysokim podobieństwie w reprezentacji kątowej RNA, LCS-TA, 3) opracowanie aktualizowanej na bieżąco bazy pętli wieloramiennych w RNA.

Pierwszy element wkładu odnosi się do pracy pt. *StructAnalyzer – a tool for sequence vs. structure similarity analysis*, którą mgr Wiedemann opublikował razem z dr. Maciejem Miłostanem w 2016 r. w perdiodyku Acta Biochimica Polonica. Przedmiotem tego projektu było zaprojektowanie oraz implementacja oprogramowania pozwalającego znajdować oraz uwidaczniać różnice struktury trzeciorzędowej RNA występujące pomimo bardzo wysokiego podobieństwa sekwencji. Badania

przeprowadzone z wykorzystaniem oprogramowania StructAnalyzer doprowadziły m.in. do przytoczonej wyżej konkluzji. Nie będąc aktywnym członkiem społeczności zaangażowanej w analizę obliczeniową struktur przestrzennych RNA, nie potrafię stwierdzić na ile wniosek dot. znaczenia lokalnym motywów strukturalnych dla określenia podobieństwa sekwencji był w czasie publikacji tej pracy oryginalnym. Umiarkowany status periodyku oraz analiza cytowań sugerują, że recepcja artykułu była ograniczona głównie do środowiska naukowego Autora rozprawy. Jednocześnie praca ta z pewnością stanowiła bardzo ważną podwalinę dla kolejnych działań naukowych doktoranta. Również dla czytelnika omawianej rozprawy stanowi ona dobre wprowadzenie do problematyki.

Najważniejszym wynikiem otrzymanym przez Autora rozprawy wydaje się być oryginalna metoda LCS-TA. Ten element wkładu mgr. Wiedemanna opisany został w pracy pt. *LCS-TA to identify similar fragments in RNA 3D structures* opublikowanej razem z dr.dr. Tomaszem Żokiem, Maciejem Miłostanem oraz prof. Martą Szachniuk w 2017 roku czasopiśmie BMC Bionformatics. Bazę dla metody stanowiły wcześniejsze prace dr. Tomasz Żoka oraz prof. Marty Szachniuk nad wykorzystaniem reprezentacji kątowej do porównywania cząsteczek RNA. Celem badań mgr. Wiedemanna było znalezienie najdłuższych ciągłych segmentów o wysokim podobieństwie w reprezentacji kątowej RNA, na wzór analogicznych koncepcyjnie rozwiązań wykorzystywanych przy ocenie podobieństwa struktur przestrzennych białek. Mając na uwadze złożoność obliczeniową problemu, Autor zaprojektował eleganckie rozwiązanie oparte o klasyczne podejście dziel i zwyciężaj (ang. *divide and conquer*), dzięki czemu znacząco ograniczył oczekiwany czas obliczeń względem długości sekwencji. Co istotne, metoda LCS-TA została wykorzystana jako jedna z metryk podobieństwa na platformie RNA-Puzzles, służącej międzynarodowej społeczności badań obliczeniowych RNA do prowadzenia cyklicznego eksperymentu oceny metod przewidywania struktury RNA (na wzór CASP-u, który przyczynił się do postępu w przewidywaniu struktur białek). Warto zauważyć, że metoda LCS-TA była prezentowana przez mgr. Wiedemanna m.in. podczas prestiżowej konferencji RECOMB w Paryżu w 2018 r. Liczba cytowań publikacji wynosi 11, w tym tylko jedno obce. Prawdopodobnym jest, że użytkownicy platformy RNA-Puzzles ograniczają się do cytowania jej jako całości: publikacja pt. *RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools*, opublikowana w prestiżowym Nucleic Acid Research w 2020, której mgr Jakub Wiedemann jest współautorem, była już cytowana 36 razy (według Google Scholar). Reasumując, omawiany wkład łączy wysokie walory informatyczne z praktyczną użytecznością.

Jako trzeci element wkładu zostało wskazane opracowanie bazy RNAloops zawierającej wszystkie wieloramienne pętle RNA dostępne w bieżącej edycji bazy struktur molekularnych PDB. Z oświadczeń autorskich wynika, że baza powstała we współpracy pomiędzy mgr.mgr. Jakubem Wiedemannem i Jackiem Kaczorem i przy udziale szerszego zespołu kierowanego przez prof.prof. Martę Szachniuk oraz Macieja Antczaka. Opis bazy został opublikowany pt. *RNAloops: a database of RNA multiloops* w 2022 r. w czołowym periodyku *Bioinformatics*. W przypadku tego wkładu najważniejsze jest znaczenie praktyczne: jest to jedyny aktualizowany automatycznie na bieżąco zasób umożliwiający analizę wieloramiennych pętle RNA. Potencjalne korzyści zostały przekonujące zilustrowane opisem przewidywanych przypadków użycia w ww. artykule. Na uwagę zasługuje także autorski algorytm ekstrakcji wieloramiennych pętli opracowany przez Kandydata. Jakkolwiek jest za wcześnie, aby wyrokować o recepcji bazy RNAloops, to otrzymane przez jej współtwórców nagród konferencji RNA Society oraz Polskiego Towarzystwa Bioinformatycznego stanowią obiecujący prognostyk.

3. Poprawność

Muszę przyznać, że w omawianej rozprawie nie dopatrzyłem się znaczących błędów. Moje uwagi krytyczne będą zatem dotyczyły raczej tego, czego w pracy mi zabrakło, przy czym staram się uwzględniać formułę rozprawy jako swego rodzaju przewodnika po cyklu artykułów naukowych. Najogólnie: oczekiwałbym szerszej dyskusji osiągniętych wyników, także w perspektywie rozwoju dziedziny, który nastąpił od czasu publikacji do momentu złożenia rozprawy.

Na przykład w rozdziale 2.1, opisującym badania nad zależnościami pomiędzy sekwencją a strukturą przestrzenną RNA, brakuje odniesienia do aktualnego stanu wiedzy w tym zakresie. Co więcej analizy struktur prowadzące do postawienia tezy, iż porównanie lokalnych motywów strukturalnych ma zasadnicze znaczenie dla oceny podobieństwa cząsteczek RNA, ograniczają się do pojedynczych w sumie przypadków. O ile taka oszczędność jest częściowo zrozumiała w przypadku artykułu stanowiącego bazę dla tego rozdziału, jako że w zasadzie prezentuje on oprogramowanie, to w kontekście tematu rozprawy wskazane byłoby poszerzenie tych aspektów. Jednocześnie chciałbym zauważyć, że samo pytanie badawcze stojące u podstaw tej pracy, tj. *jak zróżnicowane strukturalnie są podobne do siebie sekwencje RNA?*, jest dobrze sformułowane, a czytelny opis oprogramowania StructAnalyzer wskazuje na jego duże możliwości. Tym bardziej warto byłoby przedstawić informację o jego wykorzystaniu w dalszych badaniach.

W rozdziale 2.2, który traktuje o metodzie znajdowania najdłuższych ciągłych segmentów o wysokim podobieństwie w reprezentacji kątowej RNA, chciałbym podkreślić przekonujące umotywowanie zaproponowanego podejścia oraz przejrzysty opis samej metody. Natomiast opis przeprowadzonych testów metody jest w moim odczuciu zbyt mocno skrócony w rozprawie w stosunku do publikacji stanowiącej podstawę rozdziału. Także i w niej brakuje szerszego i bezpośredniego porównania miary LCS-TA (dla progu parametru MCQ) z innymi bardziej tradycyjnymi miarami. Oczekiwałbym precyzyjnej rekomendacji w jakich sytuacjach i połączeniu z jakimi innymi miarami stosować miarę LCS-TA? Zastanawia przy tym duża wrażliwość LCS-TA na wartość progową MCQ. Autorzy publikacji zdają się być świadomi tych znaków zapytania, ponieważ podsumowując swój tekst piszą, że *[z]amierza[ją] przeprowadzić testy metody na dużą skalę w celu określenia wiarygodnych progów MCQ[; oraz, że] [p]lanuj[ą] przeanalizować związek między wynikami LCS-TA a motywami struktury drugorzędowej analizowanych struktur RNA*. Z rozprawy nie można jednak wywnioskować, czy awizowane kroki działania zostały podjętę.

Co do rozdziału 2.3 moje uwagi krytyczne będą miały charakter redakcyjny (poniżej). Uznanie natomiast budzi sam projekt i funkcjonalność opracowanego serwisu udostępniającego dane o pętlach wieloramiennych w strukturach RNA z PDB. W artykule stanowiącym podstawę tego rozdziału chciałbym zwrócić uwagę na wysoką jakość ilustracji, a przede wszystkim – raz jeszcze – na przekonujące przykłady zastosowań zasobu. Chronologicznie patrząc, w kontekście powyższych uwag, rzuca się w oczy rosnąca jakość prac pierwszoautorskich stanowiących podstawę rozprawy.

Przechodząc do kwestii szczegółowych, głównie redakcyjnych, zwróciłem uwagę na kilka niewielkich niedociągnięć. Na przykład, na stronie 10. występuje stylistyczny problem w opisie wzorów (1.4) i (1.5). Na kolejnych stronach niezrozumiałe jest włączenie do przeglądu wiedzy we Wprowadzeniu (rozdz. Introduction) miary proponowanej przez Autora w niniejszej rozprawie. Dalej, na stronie 13., w opisie metody „podziel i ogranicz" wątpliwości budzi sformułowanie (w tłumaczeniu) *procedura ograniczająca pomija gałęzie drzewa, którego ścieżki nie zawsze prowadzą do optymalnego rozwiązania"*. W rozdziale 2 czcionka jest zbyt mała w etykietach rysunku 2.2. W rozdziale 3. natomiast brakuje nieco linii „oddechu" po definicjach procedur (strona 31), choć być może moje wrażenie wynika z przyzwyczajenia do standardów języka Python. Wreszcie, na stronach 7 i 32, nieco

razi użycie operatora splotu we wzorach 1.1 i 2.1. Natomiast godny pozytywnego podkreślenia jest zwięzły i przejrzysty styl formułowania myśli.

4. Wiedza kandydata

Ogólny stan wiedzy został omówiony w rozdziale 1., który przedstawia biologiczne i bio-informatyczne podstawy dot. struktury i funkcji *RNA* (1.1), wprowadza reprezentację kątową (1.2), problematykę oceny jakości modeli przestrzennych RNA (1.3) oraz podstawy informatyczne proponowanych metod i narzędzi – z zakresu algorytmiki oraz systemów bazodanowych (1.4). Wstęp jest zwięzły, a przy tym – co stwierdzam jako bioinformatyk, który na co dzień zajmuje się białkami i peptydami, a nie RNA – bardzo użyteczny w dalszej lekturze rozprawy. Ponadto o praktycznych kompetencjach Autora w zakresie informatyki technicznej świadczy fakt zaprojektowania oraz implementacji trzech narzędzi (StructAnalyzer, biblioteki LCS-TA w ramach pakietu MCQ4Structures, RNAloops). Umiejętny sposób ich opisu niedwuznacznie wskazuje na dużą wiedzę i „obycie" informatyczne Kandydata.

Bibliografia, która liczy niespełna 60 pozycji, jest bogata np. w zakresie miar podobieństwa cząsteczek RNA. Spodziewałbym za to nieco szerszego potraktowania bieżącego rozwoju dziedziny analizy i przewidywania struktury RNA. Niewykluczone jednak, że moje oczekiwania są wygórowane ze względu na spoglądanie przez pryzmat bliższej mi bioinformatyki strukturalnej białek, która przeżywa dość gwałtowny rozwój. Powyższa wątpliwość nie podważa wypływającego z lektury rozprawy i cyklu publikacji przekonania co do adekwatnie wysokiego poziomu wiedzy Autora z zakresu informatyki.

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)[1] moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problem naukowego?

[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*

Podpis

[1] http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf