Dr hab. Paweł Górecki
Wydział Matematyki, Informatyki i Mechaniki
Uniwersytet Warszawski
gorecki@mimuw.edu.pl

21 luty 2018

Recenzja rozprawy doktorskiej

Tomasza Żoka

zatytułowanej:

Algorithmic Aspects of RNA Structure Similarity Analysis

1. Problem badawczy i jego znaczenie

Recenzowana rozprawa doktorska mgr inż. Tomasza Żoka poświęcona jest teoretycznym i praktycznym zagadnieniom porównywania struktur cząsteczek RNA (kwasu rybonukleinowego). Z biologicznego punktu widzenia, RNA to jedne z kluczowych związków chemicznych biorących udział w wielu ważnych procesach zachodzących w komórkach. Cząsteczka RNA zbudowana jest z ciągu czterech rodzajów nukleotydów połączonych wiązaniami fosfodiestrowymi. Stąd takie cząsteczki można reprezentować jako ciągi nad alfabetem czteroliterowym podobnie do bardziej rozpoznawalnego DNA. Zasadniczą różnicą między nimi jest jednoniciowość RNA (z wyjątkami). Sama rekonstrukcja sekwencji nukleotydów w RNA jest obecnie stosunkowo prosta z użyciem technologii sekwencjonowania (np. RNA-seq). Zatem z jednej strony RNA wydaje się podobna DNA, ale ze względu na prostszą konstrukcję (jednoniciowość), łatwiej tworzy złożone struktury przestrzenne. Poznanie struktury przestrzennej danej cząsteczki RNA umożliwia lepsze zrozumienie roli jaką pełni ona w komórce, co efekcie może prowadzić do odkryć w medycynie i biologii. Jednakże, rekonstrukcja tych struktur stanowi od lat wielkie wyzwanie dla badaczy z różnych dziedzin. Rozprawa mgr inż. Tomasza Żoka dotyczy porównywania tychże struktur. Jest to zagadnienie fundamentalne dla innych problemów, np. określenia podobieństwa dwóch sekwencji, znajdowania w bazach struktur fragmentów najlepiej pasujących do danego zapytania, czy rekonstrukcji struktur przy danym ciągu RNA. W podsumowaniu, bez wątpienia jest to ważna współcześnie tematyka, która dotyczy zastosowań m.in., w medycynie, biologii i biotechnologii. Ta tematyka doskonale wpisuje się w światowe trendy badawcze.

2. Wkład autora

Rozprawa składa się z pięciu rozdziałów. Główny wkład Autora jest przedstawiony rozdziale trzecim i czwartym, które można traktować w pewnym stopniu niezależnie.

Rozdział trzeci wprowadza nowe metody porównywania struktury drugorzędowej RNA. Na początku Autor przedstawia znane modele zapytań pozwalające definiować tzw. *motywy* reprezentujące drugorzędowe strukturalne fragmenty RNA. Wady tych modeli stanowią motywację do zaproponowania bardziej ogólnego modelu, w którym motywy lub fragmenty sekwencji z elementami drugorzędowej struktury są wyrażane w postaci grafu (*mixed graph*). W tym grafie węzeł to nukleotyd, a krawędź to albo wiązanie pierwszorzędowe (pomiędzy sąsiednimi nukleotydami w nici RNA) albo oddziaływanie drugorzędowe. Przy zastosowaniu takiego modelu, Autor proponuje algorytmy do wyszukiwania doskonałych i przybliżonych dopasowań, a następnie pokazuje przykłady

P. G

zastosowań tych algorytmów do znajdowania brakujących motywów oraz do predykcji struktury RNA. O ile same algorytmy i model są stosunkowo nieskomplikowane, przykładowe zastosowania są interesujące i jednocześnie nietrywialne.

Rozdział czwarty dotyczy struktury trzeciorzędowej (3D) RNA. Autor opracował metodę porównywania struktur trzeciorzędowych RNA, w których zamiast współrzędnych atomów, używa się tzw. kątów torsyjnych, czyli zbioru kątów między odpowiednimi ciągami atomów definiowanych dla każdego nukleotydu w sekwencji. Ponieważ te kąty są niezależne od ustawienia danej cząsteczki RNA, potencjalnie łatwiej jest porównywać struktury 3D zakodowane w przestrzeni kątów torsyjnych. W oparciu o kąty torsyjne w rozprawie przedstawione są dwie miary MCQ (średni kąt torsyjny) i MedCQ (medianowy kąt torsyjny) wraz ze szczegółową dyskusją oraz algorytmem typu 'sliding-window' do porównywania struktur o różnych rozmiarach. Następnie Autor pokazuje jak rozwiązać problem porównywania struktur o wielu łańcuchach z użyciem algorytmu wielomianowego dla problemu przypisania (metoda węgierska). W kolejnej części opisany jest projekt MCQ4Structures, w którym m.in. zawarte są implementacje wszystkich metod z rozprawy. Rozdział czwarty kończy się zestawem zastosowań metod do klastrowania i porównywania struktur RNA w konkursie RNA-puzzle.

W rozprawie przeprowadzono kilka obliczeniowych eksperymentów ze strukturami RNA. W mojej opinii wykonane analizy są często nietrywialne i wymagają dobrej wiedzy dziedzinowej. Oprócz metod, ten element rozprawy uznaję za bardzo wartościowy.

Mocną stroną rozprawy są zastosowania metod. Szczególnie wartość wyników jest wyróżniająca w kontekście konkursu RNA-puzzles, gdzie zaproponowane miary są używane do oceny modeli. Oznacza to, że metody zaproponowane przez mgr inż. Tomasza Żoka znalazły uznanie w środowisku naukowym, czego wyrazem są m.in. cykliczne publikacje-raporty z konkursu w czasopiśmie *RNA* (IF=4.6).

Do ważnego praktycznego aspektu rezultatów rozprawy zaliczam także projekt MCQ4Structures. Jest to zbiór narzędzi autorstwa mgr inż. Tomasza Żoka, który jest dostępny w publicznych repozytoriach (github).

Części rozprawy są związane z publikacjami, które Autor zamieścił w Dodatku. Pierwsza to praca opublikowana wspólnie z promotorami w *Central European Journal of Operations Research* w 2014, wprowadza miarę MCQ. Druga to praca w *International Journal of Applied Mathematics and Computer Science z 2015,* dotyczy zastosowań tej miary przy klastrowania struktur RNA. Kolejne dwie to wspomniane wcześniej wieloautorskie prace z czasopisma *RNA* dotyczące konkursu RNA-puzzles. Wskaźniki bibliometryczne z indeksu Web of Science pokazują na wyróżniający dorobek naukowy. Sama praca o MCQ z 2014 roku posiada 16 cytowań, co jest bardzo dobrym wynikiem dla pracy o charakterze stosunkowo teoretycznym. W indeksie WOS mgr inż. Tomasz Żok posiada 13 publikacji cytowanych 127 razy (h-index 6), które ukazywały się w dobrych i bardzo dobrych czasopismach (oprócz czasopisma *RNA*, m.in. artykuł typu application note w *Nucleic Acid Research* z IF=10.1, także *Bioinformatics*, *BMC Bioinformatics*).

P. L

3. Poprawność

Algorytmy w pracy są łatwe do zrozumienia i poprawne z dokładnością do drobnych uwag, które zamieszczam w cz. 5 recenzji. Poza ogólnymi opisami w tekście brakuje dokładnego opisu niektórych procedur z Alg. 2 (IMMG). Brakuje też elementów formalnych, choć nie są one najbardziej istotne w tej rozprawie, np. uzasadnienia poprawności, podania złożoności algorytmów zarówno niektórych wspominanych we wprowadzeniu jak i własnych algorytmów Autora.

Kilka definicji wymaga dopracowania, np. definiując pojęcie powinno dostarczyć się od razu wszystkich komponentów potrzebnych do jej zrozumienia. Przykładem jest definicja f w formule (4.5), która przy podanych warunkach jest trywialnie zerem gdy wszystkie x_ij są 0. Innym przykładem jest Def. 3.3.1 (uwagi zamieściłem w części 5 recenzji).

Przeprowadzone eksperymenty obliczeniowe nie budzą zastrzeżeń. Oczywiście trudno zweryfikować wyniki złożonych obliczeń bez ich ponownego przeprowadzenia, ale sama metodologia i opis jest przekonujący i formalnie poprawny. Jako recenzent jestem ciekawy np. w jaki sposób został wybrany przykład z Rozdziału 3, w którym strukturalny fragment RNA jest ręcznie modyfikowany w eksperymencie pokazującym, że zaproponowany model może ulepszać predykcję struktur.

Powyższe uwagi nie zmieniają pozytywnego odbioru tej rozprawy, mimo pewnych nieformalnych uproszczeń i drobnych uwag redakcyjnych rozprawę czyta się bardzo dobrze. Ponadto, rozprawa jest zilustrowana licznymi obrazkami i tabelami, które przez bardzo dobry dobór i wykonanie znacząco ułatwiają zrozumienie.

4. Wiedza kandydata

Rozdział pierwszy zawiera łagodne wprowadzenie do bioinformatyki i RNA, a stan wiedzy dotyczący RNA i porównywania struktur jest opisany w rozdziale drugim. Jakość tych rozdziałów nie budzi zastrzeżeń. Są one opisane w sposób klarowny z dobrze zilustrowanymi przykładami. Bibliografia składa się z ok. 80 pozycji i również nie budzi zastrzeżeń.

Interdyscyplinarna tematyka podjęta w rozprawie wykracza poza klasyczną informatykę. Samo wykonanie obliczeniowych eksperymentów, wymaga dobrej wiedzy dziedzinowej. Autor wykazał się nie tylko umiejętnościami projektowania algorytmów, ich implementacji i wdrażania, a także dobrym rozumieniem nietrywialnych zagadnień z pogranicza dziedzin takich jak biologia molekularna, chemia, bioinformatyka i trygonometria.

5. Inne uwagi

Uwagi formalne.

Str. 18. Figure 2.12. Niektóre z atomów po prawej nie pasują do atomów w schematycznej strukturze RNA strukturze (ale można to odtworzyć).

P. G

Str. 35. Def. 3.3.1. W Pkt. 3 definicji, indeksy i, i+1 są niezdefiniowane. Pkt. 3 oznacza, że nie wszystkie krawędzie pierwszorzędowe z S muszą wystąpić w A (warunek jest w jedną stronę); zakładam, że intencją Autora jest *„<x,y> w A wtw y jest następnikiem x w S”*. Podobnie w pkt. 4. W szczególności obecna definicja dopuszcza by mixed-graf reprezentujący strukturę S nie miał krawędzi nawet jeśli S je posiada.

Str. 31. Czy ta notacja jest jednoznaczna? Np. dla 3-way junction, jest alternatywny sposób łączenia - gdzie G (1) łączy się z C (15).

Str 42. W Algorytmie 1 w 17 linii powinno być „(begin,end) in E_G” (zbiór krawędzi). Podobnie w Algorytmie 18, str 43, linia 10.

Algorytm 2 jest bardzo podobny do Algorytmu 1 (część jest kopią). Ponadto poza krótkim opisem w tekście brakuje procedury ElongateShortestStrand, ScoreFragments i SortBySimilarity.

Str. 67. Definicja „assignment problem” wymaga (drobnej) korekty. Przy takim sformułowaniu problem ma rozwiązania tylko gdy n=m. Gdy m<>n nie istnieje macierz [x_ij] spełniająca warunki z tej definicji. Dla wyników rozprawy ten ogólny przypadek jest potrzebny.

Str. 19. Formuła (2.1) powinna być dokładniej wyjaśniona.

Wybrane uwagi językowe

Abstract. "reveal the importance" zamiast "reveal importance".
Str. 21. "At first glance" zamiast "At a first glance".
Str. 22. "the same values" zamiast "exactly the same values".
Str. 22. "has also been" zamiast "has been also"; jest więcej podobnych.
Str. 22. powinno być "a tree graph".
Str. 22. podwójne "as".
Str. 25 "has a different meaning" zamiast "has different meaning".
Str. 25 "Thus, a new measure" zamiast "Thus, new measure".
Str. 29. "the first three characters".
Str. 29. Sformułowanie "The most innovative and powerful search is the search engine of RNA FRABASE", dot. badań w których brał udział Autor, wymaga uzasadnienia.
Str. 39. "biological point of view" brakujący przedimek; podobnych przypadków jest więcej.
Str. 44. "composed from" - warto sprawdzić alternatywne sformułowania.
Str. 47. Tabelka 3.2 posiada za mały rozmiar czcionek.
Str. 48. "the the".
Str. 48. "allows inserting" zamiast "allows to insert".
Str. 48. "exact structural equivalent" zamiast "exact structure equivalent".
Str. 66. Powinno być „under the assumption”.
Str. 68. Powinno być „a user has”.
Str. 72-87. Kilka razy. Powinno być „were taken” zamiast „where taken”.
Str. 78. Powinno być „the model from Das”.
W niektórych fragmentach wzmocnienia huge, extremely, itp., można pominąć bez straty dla tekstu.
Jest nieco problemów interpunkcyjnych z przecinkami.
Występują liczne słowa zarówno z języka angielskiego brytyjskiego jak i amerykańskiego.

P. L

6. Podsumowanie

Mimo pewnych usterek redakcyjnych, zawartość merytoryczną ocenianej rozprawy uważam za interesującą i wartościową. Potwierdzam także, że Autor potrafi zaproponować nowe metody porównywania struktur drugorzędowych i trzeciorzędowych RNA oraz skutecznie je zaimplementować w postaci narzędzi informatycznych, które pozwalają na użycie metod w praktyce. Stwierdzam, że recenzowana przeze mnie praca spełnia wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy i wnoszę o dopuszczenie magistra inżyniera Tomasza Żoka do dalszych etapów przewodu doktorskiego.

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) 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 lub Automatyka i Robotyka?

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

C. Czy kandydat umiejętność samodzielnego prowadzenia pracy naukowej?

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

Ponadto, biorąc pod uwagę znaczenie interdyscyplinarnych wyników już obecnie uznanych i stosowanych przez środowisko naukowe, opublikowanych w bardzo dobrych czasopismach o zasięgu międzynarodowym, oraz implementacje z praktycznymi zastosowaniami rekomenduję wyróżnienie rozprawy doktorskiej.

Podpis