

Prof. dr hab. inż. Sebastian Deorowicz,
Katedra Algorytmiki i Oprogramowania
Wydział Automatyki, Elektroniki i Informatyki
Politechnika Śląska
44-100 Gliwice, ul. Akademicka 16

Gliwice, 08.07.2021

Recenzja rozprawy doktorskiej

Tytuł rozprawy:

Application of the transcriptomic data and evolutionary conservation in search and characterization of long non-coding RNAs in animals

Autor:

mgr Oleksii Bryzghalov

Promotor:

dr hab. Michał Szcześniak, prof. UAM

I. Problematyka naukowa oraz przedmiot rozprawy

Tematyka rozprawy dotyczy długich niekodujących cząsteczek RNA (ang. long noncoding RNA, lncRNA). Wiadomo, że cząsteczki te mogą uczestniczyć w różnych procesach komórkowych takich jak transkrypcja, splicing, translacja i inne. Niestety do tej pory zdecydowana większość zidentyfikowanych lncRNA nie została scharakteryzowana, co znacznie utrudnia poznanie ich funkcji w organizmach żywych. Fragmenty te nie są zbyt dobrze zakonserwowane, co powoduje trudności w stosowaniu narzędzi badawczych używanych np. do analizy fragmentów kodujących geny. Recenzowana praca dotyczy właśnie tej problematyki.

II. Analiza treści rozprawy oraz uzyskanych wyników

1. Treść rozprawy

Rozprawa ma charakter krótkiego wprowadzenia oraz cyklu trzech artykułów opublikowanych w czasopismach indeksowanych w bazie Journal Citation Reports (JCR).

Wprowadzenie, liczące 5 stron, zwięźle wprowadza w tematykę badań lncRNA i związanych z tym problemów. Autor omawia tutaj m.in. dlaczego badania tych struktur są trudne. Charakteryzuje także skrótowo zawartość wybranej bazy danych cząsteczek lncRNA pokazując jak mało informacji o funkcjach można w niej znaleźć. Ponadto, Doktorant krótko omawia zawartość trzech publikacji, stanowiących dalszy ciąg rozprawy.

Druga część składa się z trzech publikacji stanowiących sedno rozprawy. Prace te to:

- A1 Bryzghalov O., Szcześniak M.W., Makałowska I., Retroposition as a source of antisense long non-coding RNAs with possible regulatory functions. *Acta Biochimica Polonica* 2016; 63(4):825–833.
- A2 Bryzghalov O., Szcześniak M.W., Makałowska I., SyntDB: defining orthologues of human long noncoding RNAs across primates. *Nucleic Acids Research* 2020, 48(D1):D238–D245.
- A3 Bryzghalov O., Makałowska I., Szcześniak M.W., lncEvo: automated identification and conservation study of long noncoding RNAs. *BMC Bioinformatics* 2021, 22:59.

Praca [A1] dotyczy badania mechanizmu retropozycji prowadzącego do powstawania lncRNA. Fragmenty lncRNA powstałe w ten sposób oraz geny, które są ich źródłem mogą podlegać ekspresji w sposób skorelowany. W ramach pracy przeprowadzono analizę retrokopii oraz ich genów rodzicielskich dla niektórych ludzkich genów. Zbadano także kwestię konserwacji tych lncRNA.

Zgodnie z deklaracją Doktoranta jego wkład w powstanie tej publikacji obejmował filtrowanie i analizę danych bioinformatycznych, analizę literatury, studia przypadków oraz współudział w przygotowaniu tekstu publikacji. Udział procentowy został oszacowany na 40%. Według bazy Web Of Science praca ta była cytowana (na dzień sporządzenia recenzji) 8 razy. Publikacja ta, podobnie jak pozostałe, została opracowana w 3-osobowym zespole.

W ramach pracy [A2] opracowano potok umożliwiający odkrycie długich niekodujących fragmentów RNA u człowieka oraz 11 innych naczelnnych. Wynikiem pracy oprócz tego potoku jest także baza danych SyntDB zawierająca zarówno znalezione fragmenty RNA jak i informacje o ich ortologach u innych naczelnnych. Baza ta zawiera ok. 78 tys. ludzkich lncRNA oraz mniejsze, ale także istotne ilości takich fragmentów u innych naczelnnych. Informacje o zakonserwowaniu niektórych fragmentów mają istotne znaczenie dla dalszych badań, których celem może być odkrycie funkcji, które pełnią niektóre z tych fragmentów.

Udział Doktoranta w publikacji polegał na: realizacji potoku analitycznego dla konserwacji ewolucyjnej, analizie danych, przygotowaniu bazy danych oraz współpracy przy opracowywaniu tekstu. Udział procentowy Doktoranta szacowany jest na 40%. Według bazy Web Of Science praca ta była cytowana (na dzień sporządzenia recenzji) 5 razy.

Badania opisane w pracy [A2] miały swoją kontynuację, co zaowocowało publikacją [A3]. Celem tych prac było opracowanie potoku, który automatyzowałby odkrywanie lncRNA. Do tego celu wykorzystana została technologia Nextflow. Opracowany potok jest przeznaczony do uruchamiania na zwykłych komputerach oraz w środowisku chmurowym takim jak np. AWS.

Udział Doktoranta w publikacji polegał na: rozbudowie i wdrożeniu potoku analitycznego, testowaniu i optymalizacji tego potoku oraz współpracy przy opracowywaniu tekstu artykułu.

Udział procentowy Doktoranta w przypadku tej pracy szacowany jest na 70%. Praca ta nie została jeszcze zacytowana (według bazy Web of Science), ale wiąże się to z faktem, że została ona opublikowana w bieżącym roku.

2. Najważniejsze wyniki przedstawione w rozprawie

W mojej ocenie pierwszym z głównych osiągnięć rozprawy jest opracowanie potoku umożliwiającego automatyczną detekcję lncRNA. Dzięki wykorzystaniu adnotacji z bazy ENSEMBL możliwe jest łatwe poszerzanie zakresu gatunków, dla których wyznaczane są te długie niekodujące fragmenty RNA. Warte podkreślenia jest to, że narzędzie to zostało opracowane z wykorzystaniem nowoczesnych technologii, może pracować w środowisku chmurowym i dzięki temu powinno się dobrze skalować.

Drugim z głównych osiągnięć rozprawy jest opracowana baza danych zawierająca informacje o zakonserwowanych lncRNA w genomach ludzkich i innych naczelnych. Zarówno opracowane narzędzie jak i baza danych dostępna jest publicznie, co pozwala mieć nadzieję, że w niedługiej przyszłości oba zyskają na popularności i będą wykorzystywane w badaniach w różnych ośrodkach naukowych.

3. Uwagi merytoryczne

Zasadniczą część rozprawy stanowią publikacje, które podlegały już procesowi recenzji w uznanych czasopismach naukowych. Stąd też trudno o sformułowanie jakichś zasadniczych uwag krytycznych. Uważam, że zaproponowana metoda wyznaczania lncRNA, jak i uzyskana dzięki niej baza danych są wartościowe. Doktorant wykazał się także umiejętnościami z zakresu wykorzystania różnych istniejących narzędzi i bibliotek do stworzenia z ich wykorzystaniem efektywnych potoków.

Mam jednak pewne mniej istotne spostrzeżenia i uwagi:

1. W pracy [A2] znajduje się stwierdzenie: "We also used gzip compression for transferring compressible responses and images". Jeśli chodzi o kompresję obrazów, to gzip nie jest najlepszym rozwiązaniem. W szczególności, obrazy są zwykle albo już w formatach skompresowanych i nie da się ich wtedy skompresować bardziej gzipem, albo też lepiej je kompresować do formatów powszechnie stosowanych jak PNG, JPEG.
2. Interesujące byłoby się dowiedzieć jaka platforma została wykorzystana do uzyskania wyników z Tabeli 1 z pracy [A3]. W szczególności interesujący jest aspekt jak skuteczne jest zrównoleglenie oraz jak wygląda skalowanie na większą liczbę maszyn, albo większą liczbę rdzeni. Byłbym wdzięczny za zaprezentowanie jakichś wyników w trakcie publicznej obrony.
3. W pracy [A3] podkreślana jest waga reprodukowalności wyników. W pełni się z tym zgadzam. Niemniej w praktyce problemem jest często nie tylko to jakie dokładnie wersje narzędzi zostaną wykorzystane. Niestety niektóre z narzędzi działają w sposób niedeterministyczny i przy kolejnych uruchomieniach mogą generować nieco inne wyniki. Nie mam wiedzy czy wszystkie stosowane w pracy narzędzia są deterministyczne (np. Bowtie2 ma opcję `--non-deterministic`, które w niektórych zastosowaniach jest rekomendowana), ale jeśli mocno podkreśla się jako zaletę kwestię reprodukowalności wyników, to warto byłoby tę kwestię zbadać, choćby wykonując potok wielokrotnie i sprawdzając czy uzyskiwane pliki pośrednie (oraz końcowe) są na pewno takie same. Ewentualny

brak determinizmu nie byłby wadą opracowanego potoku a jedynie jego cechą wynikającą z wykorzystanych komponentów. Co więcej, w praktyce niemożność odtworzenia tych samych wyników bierze się czasami także z tego, że po pewnym czasie dane przechowywane w różnych repozytoriach mogą zostać usunięte. Na to nie ma jednak już dobrej rady oprócz zrobienia lokalnych kopii plików, na podstawie których stworzono np. bazę SyntDB.

Kwestia, która, w mojej opinii, mogłaby pracę nieco ulepszyć, to stworzenie obszerniejszego wprowadzenia. W chwili obecnej to wprowadzenie, przynajmniej dla niespecjalisty jest nieco zbyt krótkie. Myślę także, że lekturę ułatwiłoby umieszczenie w tekście rozprawy załączników artykułów. W szczególności mam tutaj na myśli załącznik do pracy [A3], do którego w tekście artykułu są odwołania i który zawiera interesujące treści. Jego zawartość jest oczywiście dostępna on-line, ale dla kompletności mogłby on zostać wydrukowany i załączony.

4. Uwagi redakcyjne

Rozprawa została złożona dość starannie. W tekście wprowadzenia nie zauważyłem istotnych błędów. Jedyną drobną uwagę miałbym do składu bibliografii, który jest niespójny i sprawia wrażenie niedopracowanego.

5. Podsumowanie

Przedstawiona do oceny rozprawa składa się w przeważającej części z artykułów opublikowanych w prestiżowych czasopismach indeksowanych w bazie JCR. Prace te były opracowywane w zespole składającym z 3 osób, przy czym zawsze Doktorant był pierwszym autorem i jego wkład wynosił od 40% do 70%. Tematyka pracy dotyczy problematyki znajdowania długich niekodujących RNA. Pracę uważam za ambitną. Badania eksperymentalne były wszechstronne i z pewnością czasochłonne. Pozytywnie oceniam opracowane narzędzia. Doktorant wykazał się dobrymi umiejętnościami tworzenia potoków z wykorzystaniem współczesnych narzędzi.

Mimo drobnych niedociągnięć, o których wspominam w recenzji, pracę oceniam wysoko. Na zakończenie dodam, że choć nie uważam, aby wskaźniki bibliometryczne były istotne na tak wczesnym etapie kariery naukowej, to jednak wobec faktu, że są one wyróżniające, to je przytoczę. Baza Web of Science zawiera 4 prace, których współautorem jest Doktorant. Prace te były cytowane 13 razy (12 bez autocytowań) a indeks Hirscha Doktoranta wynosi 2.

III. Konkluzja

Rozprawa doktorska mgra Oleksiia Bryzghalova zawiera oryginalne i interesujące wyniki naukowe dotyczące wyznaczania długich niekodujących fragmentów RNA. Wykazuje również dużą ogólną wiedzę Doktoranta w dyscyplinie biotechnologia i świadczy o umiejętności samodzielnego prowadzenia pracy naukowej. Zawarte w recenzji uwagi krytyczne nie wpływają na moją ogólną bardzo dobrą ocenę rozprawy. Uważam, że zostały spełnione wymagania stawiane rozprawom doktorskim przez *Ustawę o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki z dnia 14 marca 2003 roku*. Wnoszę zatem o dopuszczenie wspomnianej rozprawy do publicznej obrony.

Ponadto, biorąc pod uwagę wysoką jakość uzyskanych wyników oraz fakt, iż rezultaty prac zostały już opublikowane w kilku artykułach w renomowanych czasopismach naukowych wnoszę o wyróżnienie pracy.

Sebn Perin