UNIWERSYTET WARSZAWSKI

Instytut Informatyki
Uniwersytet Warszawski
ul. Banacha 2
02–097 Warszawa
POLSKA

dr hab. Bartosz Wilczyński
profesor uczelni
Phone: +(48 22) 5544 577
Fax: +(48 22) 5544 400
e-mail: bartek@mimuw.edu.pl

Warszawa, 29. czerwca 2023 r.

Recenzja rozprawy doktorskiej pt. „Advanced data exploration techniques for augmented transcriptional landscape and its better quantification" przedstawionej przez mgr inż. Agatę Muszyńską

Recenzja niniejsza została sporządzona na zlecenie Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej zgodnie z wymogami ustawy dotyczącej procedur nadawania stopnia doktora.

Opis rozprawy

Przedstawiona rozprawa napisana jest w języku angielskim, składa się z 5 rozdziałów (Introduction, Background, Methods, Results, Summary) i liczy wraz z dodatkami 149 stron. Naturalnie, większość materiału istotnego dla oceny rozprawy znajduje się w rozdziałach trzecim i czwartym opisujących odpowiednio metodologię i wyniki rozprawy.

Dwa pierwsze rozdziały rozprawy wprowadzają do tematu i zarysowują tło przedstawionych badań. W szczególności rozdział 1 zawiera kilka fragmentów, które opisują zakres i cele rozprawy doktorskiej, a także przedstawia te aspekty rozprawy, które autorka uważa za nowatorski wkład naukowy. W szczególności, podrozdział 1.1 zawiera opis celu rozprawy, który autorka widzi jako przede wszystkim przeprowadzenie „stress-testów" obecnie stosowanych narzędzi do analizy transkryptomów i przedstawienie „rozwiązania" w postaci zestawu skryptów snakemake, które pozwalają na uruchomienie zestawu standardowych narzędzi w kilku różnych scenariuszach typowych dla analiz RNA-Seq (analiza różnicowej ekspresji genów, wykrywanie alternatywnego splicingu itp.) a także (w mniejszym zakresie) analiz danych mikromacierzowych. Podrozdział 1.2 zaś poświęcony jest nowatorskiemu wkładowi rozprawy, który, w ocenie autorki, polega na stworzeniu nowego zestawu skryptów snakemake, będących pewnym ulepszeniem obecnie dostępnych zestawów narzędzi do analizy RNA-Seq.

Rozdział trzeci zawiera opis metodologii wykorzystywanych w rozprawie. Obejmuje on zarówno podstawy technologi sekwencjonowania RNA/cDNA, podstawy technologii mikromacierzowych, jak i wszelkie techniki analizy danych, zarówno od strony statystycznej jak i

informatycznej. Jest on dość obszerny (strony 27-56 rozprawy), jednak pozostawia wrażenie pewnej pobieżności. Jest to w pewnym sensie nieuniknione, biorąc pod uwagę ogromną różnorodność narzędzi i metod wykorzystywanych w dziedzinie analizy danych transkryptomicznych. W mojej ocenie, autorka wybrnęła z tego, trudnego, zadania w sposób zadowalający, mimo, że każdy czytelnik, który ma doświadczenie w dziedzinie analizy danych transkryptomicznych znajdzie w tym opisie pewne luki. Niezależnie jednak od tych luk, widać jest, że autorka zapoznała się z metodami o dużym zakresie różnorodności i byłaby gotowa do tego, aby podjąć się w przyszłości w zasadzie dowolnej standardowej analizy transkryptomu. Być może musiałaby zapoznać się z jakąś nową metodą, ale zapewne miałaby w swoim doświadczeniu wystarczająco dużo punktów odniesienia, aby szybko douczyć się potrzebnych narzędzi.

Rozdział czwarty opisuje głównie wyniki eksperymentów przeprowadzonych przez autorkę. Obejmują one dość przekrojowe analizy kilku różnorodnych zbiorów danych i zastosowanie do nich różnych wariantów skryptów rozwiniętych przez autorkę. Zestaw tych wyników jest dość schematyczny: autorka przedstawia kolejne zbiory danych, stosuje do nich adekwatne metody analizy i przedstawia wynikowe wykresy, wraz z krótką analizą wyników. Wszystkie te analizy są obecnie dość standardowe, obejmując metody takie jak wykrywanie genów o różnicowej ekspresji, alternatywnego splicingu, normalizacja pomiędzy różnymi zestawami danych w przetwarzaniu „wsadowym" (ang. batch effect normalization). Nie widzę w tych analizach poważnych błędów rzeczowych, ale też i trudno dopatrzyć się w nich rzeczywiście nowatorskiego podejścia. Raczej określiłbym te wyniki jako pewnego rodzaju badania o naturze „przyrostowej", gdzie wraz z pojawiającymi się nowszymi wersjami standardowych metod w technologii osiągającej dojrzałość, takiej jak RNA-Seq, powstają kolejne wersje zintegrowanych narzędzi do analizy tego rodzaju danych, które pozwalają na coraz łatwiejszą i wymagającą mniej wiedzy fachowej analizę nowych danych. Biorąc pod uwagę, że dane RNA-Seq analizujemy od już niemal 15 lat (nie wspominając o mikromacierzach), proces ten osiągnął już etap gdy stworzenie narzędzia poprawiającego dostępne wersje wymaga nadal sporo wysiłku, jednak prawo „diminishing returns" powoduje, że przyrost funkcjonalności pozostaje niewspółmierny do wysiłku.

Rozdział 5 zawiera podsumowanie rozprawy i, co ważne, także w podrozdziale 5.2, podsumowuje artykuły naukowe, do których autorka przyłożyła się w czasie pracy nad doktoratem. O ile część podsumowująca, pozostawia raczej niedosyt, jako że wnioski płynące z analiz mają wymowę raczej negatywną (m.in. "currently there are no gold standards", oraz "great improvements can still be made ..."), ale z sekcji 5.2 płynie wyraźny sygnał, że autorka nie ograniczała się do prac opisanych w doktoracie, ale poświęciła też sporo czasu na rzeczywistą współpracę naukową w dużych zespołach, co doprowadziło do powstania dwóch publikacji już opublikowanych w czasopismach i jednej pracy przeglądowej w recenzjach (już opublikowanej, w momencie pisania recenzji). Wszystkie te prace są wieloautorskie, z umiarkowanym udziałem autorki (nie jest pierwszą autorką żadnej z prac), ale jedna z nich jest opublikowana w bardzo dobrym czasopiśmie (Genome Biol-

ogy).

Uwagi krytyczne

Tytuł rozprawy sugeruje, że autorka planowała stworzenie nowych „zaawansowanych" metod eksploracji danych, które miały prowadzić do „lepszych" metod kwantyfikacji transkryptomów. W istocie, po przeczytaniu rozprawy widać jest, że aż tak ambitne cele nie zostały osiągnięte. Szkoda, że nie udało się zmienić tytułu, aby bardziej odpowiadał zakresowi rozprawy. Podejrzewam, że być może taka zmiana byłaby kłopotliwa formalnie, ale myślę, że przy tak istotnym okrojeniu zakresu pracy, zmiana tytułu pozwoliłaby uniknąć efektu zbyt dużych oczekiwań wobec pracy, które taki ambitny tytuł wywołuje.

Praca napisana jest w języku angielskim i w niektórych miejscach widać jest, że nie jest to język ojczysty autorki. (np. sformułowania „confounding effects" zamiast „factors", „faster run time" zamiast „shorter..." itp. są używane niepoprawnie), ale zapewne wykorzystanie języka angielskiego podyktowane było udziałem prof. Kreila jako ko-promotora. Mimo tych, drobnych, niedociągnięć, praca daje się czytać dość dobrze i nie odbiega od poziomu typowego dla rozpraw doktorskich bronionych w Polsce.

Podział treści pomiędzy rozdziały budzi niekiedy wątpliwości. Np. transformata Burrows'a Wheelera opisana jest w podrozdziale 2.3 „NGS technologies summary". Wydaje się, że lepszym miejscem dla takiego opisu byłby rozdział 3, jako że nie jest to metoda przynależna do technik NGS, a raczej struktura danych pomocna m.in. w tej dziedzinie. Metoda ta jest potraktowana dość pobieżnie, co, być może, nie ma wpływu bezpośrednio na wyniki rozprawy, ale nie najlepiej świadczy o wiedzy kandydatki w dyscyplinie informatyka.

Podobnie, wykorzystanie rozkładu ujemnego dwumianowego (NB) w modelowaniu różnicowej ekspresji genów przedstawione jest w rozdziale wstępnym (2.4) i potraktowanie bardzo skrótowo. Autorka nie ustrzegła się przy tym sformułowań ocierających się o błąd, pisząc, że „Random sampling of RNA-Seq reads [..] can be modelled quite well by Poisson distribution" co nie jest prawdą, z czego autorka zapewne zdaje sobie sprawę pisząc dalej o wykorzystaniu rozkładu NB do tego celu.

Opisując pakiet snakemake w podrozdziale 3.1, autorka używa sformułowania „comes with a set of additional advantages", co jest dość zaskakujące, bo nie dowiadujemy się jakie to zalety, a mowa nie o jakimś marginalnym narzędziu, a głównym, wybranym przez autorkę, rozwiązaniu.

Opisując metodę TMM autorka używa określenia „trimmed mean of M-values values", co jest zapewne po prostu błędem pisarskim.

W podrozdziale 3.5.5, autorka wymienia metody poprawek związanych z testowaniem wielu hipotez, ale wspomina wyłącznie o metodzie Holma (zwanej zwykle Holma-Bonferroniego) i o podejściu FDR, pomijając prostszą i często stosowaną metodę Bonferroniego.

Ryciny 3.1 i 4.1 wydają się być identyczne. Być może powtórzenie jest zamierzone, ale jest to moim zdaniem zła praktyka w pracach naukowych.

W rozdziale 4, najważniejszym dla oceny rozprawy, rozczarowuje brak rzeczywistej integracji wyników różnych narzędzi w sposób inny niż wizualizacja różnych analiz dla tego samego zestawu danych. Wydaje się, że częściowo zgodnie z diagramem 3.1/4.1, autorka stworzyła narzędzie pozwalające na wykorzystanie różnych metod do otrzymania genów o różnicowej ekspresji i/lub alternatywnych miejsc splicingu, ale nie dokonała integracji wyników. Jeśli już tworzyć "zintegrowane narzędzie" do takich analiz, warto byłoby spróbować zintegrować wyniki różnych analiz, choćby w postaci wspólnego raportu, a jeszcze lepiej, w postaci analizy, która rozważa alternatywny splicing wspólnie z różnicową ekspresją genów. Szkoda, że autorka nie tylko nie dokonuje próby takiej, głębszej, integracji wyników, ale też nie podaje powodów, które ją przed tym powstrzymały.

Podobnie, wydaje się, że w tak złożonym narzędziu, decyzja o wykorzystaniu narzędzi do analizy efektów „wsadowych" (ang. „batch effects") jedynie w przypadku analizy różnicowej ekspresji, jednak nie w przypadku alternatywnego splicingu, pozostawia pewien niedosyt. Wydaje się, że normalizacja odczytów związanych z AS mogłaby także pomóc w lepszej powtarzalności wyników.

Podsumowanie

Podsumowując, praca mgr inż. Agaty Muszyńskiej miała niezwykle ambitne cele (wprowadzić nowe metodologie i ulepszyć ilościowe pomiary w analizie danych transkryptomicznych), jednak uzyskane wyniki są natury raczej przyrostowej i znaczącej jedynie w lokalnym wymiarze.

Rozprawa wykazuje, że autorka zdobyła wiedzę o wielu różnych metodach analiz danych transkryptomicznych, choć być może bardziej dogłębna analiza mniejszej ilości metod dałaby ciekawsze wyniki. Niemniej, widoczne jest, że jej wiedza jest wystarczająca, aby w przyszłości mogła wykonywać tego rodzaju analizy w środowisku typowych projektów bioinformatycznych.

Praca zawiera pewne niedociągnięcia formalne, a także drobne uchybienia metodologiczne, jednak nie ma w niej poważnych błędów rzeczowych.

Wyniki tej pracy pozostawiają w czytelniku pewne poczucie niedosytu, choć zapewne częściowo można to wytłumaczyć nadmiernie ambitnym tytułem, któremu trudno byłoby sprostać w środowisku analiz RNA-Seq, gdzie metodologie są już w tej chwili dość ustabilizowane.

Rozważając, czy rozprawa ta spełnia wymagania ustawowe wobec rozpraw doktorskich, trzeba przyznać, że ze względu na przyrostowy charakter wyników, odpowiedź pozytywna nie jest oczywista. Jednak, w moim odczuciu na korzyść doktorantki przemawia jej udział w kilku projektach naukowych, gdzie wykazała się praktycznymi umiejętnościami w analizie danych RNA-Seq, co pozwoliło jej uzyskać istotne wyniki naukowe. W związku z tym

uważam, że – mimo wspomnianych słabości - **rozprawa wypełnia ustawowe wymagania stawiane wobec rozpraw doktorskich** i może zostać skierowana do kolejnych etapów postępowania w kierunku nadania stopnia doktora.

Bartosz Wilczyński