Warszawa, 19.06.2023

dr hab. inż. Tomasz Gambin, prof. uczelni
Instytut Informatyki
Politechnika Warszawska
ul. Nowowiejska 15/19, 00-665 Warszawa
tomasz.gambin@pw.edu.pl

Recenzja rozprawy doktorskiej **Pani mgr inż. Agaty Muszyńskiej** pt. "Advanced data exploration techniques for augmented transcriptional landscape and its better quantification"

Tematyka i układ pracy

Recenzowana rozprawa doktorska posiada charakter interdyscyplinarny i dotyczy obszarów informatyki, bioinformatyki, transkryptomiki. Jej przedmiotem jest rozwój i walidacja kompleksowego procesu analizy danych RNA-seq oraz mikromacierzy.

Układ pracy jest przejrzysty i nie budzi większych zastrzeżeń. Rozprawa została napisana w języku angielskim i składa się z 6 rozdziałów oraz 4 załączników. W pierwszym rozdziale został przedstawiony cel pracy oraz wkład jaki praca wnosi w badania nad analizą danych z RNA-seq i mikromacierzy. W rozdziale drugim zostały omówione podstawy biologiczne, w tym techniki mikromacierzowe, metody sekwencjonowania następnej generacji oraz wyzwania związane z analizą RNA-seq stanowiące jednocześnie motywację do podjęcia badań prezentowanych w rozprawie. W części trzeciej została przedstawiona konstrukcja opracowanego w ramach rozprawy potoku przetwarzania wraz ze szczegółowym omówieniem narzędzi, które weszły w jego skład. Rozdział czwarty zawiera wyniki analiz przeprowadzonych na trzech zbiorach danych (dwóch zbiorów RNA-seq i jednego zbioru danych mikromacierzowych). Rozdział piąty zawiera podsumowanie oraz informacje o kodzie źródłowym. W rozdziale szóstym umieszczono podziękowania.

Cel rozprawy

Celem tej pracy doktorskiej jest zbadanie i próba rozwiązania głównych problemów jakie występują w analizie danych z RNA-seq i mikromacierzy, w szczególności w kontekście reprodukowalności uzyskiwanych wyników. Pani Agata Muszyńska opracowała kompleksowy potok przetwarzania do ilościowej (analizy ekspresji) i jakościowej (wykrywania alternatywnego splicingu) analizy danych RNA-Seq, zweryfikowanych na niezależnych zestawach danych rzeczywistych zebranych w nowych eksperymentach. Oprogramowanie zostało utworzone przy uwzględnieniu najlepszych praktyk wypracowanych m.in. przez konsorcja SEQC oraz MAQC. Potok przetwarzania ma na celu zintegrowanie wielu etapów analizy danych RNA-seq, w tym kontroli jakości, wstępnego przetwarzania, dopasowania, kwantyfikacji, różnicowej ekspresji genów (DGE) i analizy wzbogacania zestawu genów oraz analizy alternatywnego splicingu (AS).

Rozpatrywane zagadnienie badawcze jest istotne z punktu widzenia dalszego rozwoju analizy danych transkryptomicznych. O ile główny cel został stosunkowo jasno zdefiniowany przez Autorkę to w pracy brakuje dokładniejszego opisu celów szczegółowych, którymi mogłyby być m.in.: (i) porównanie do innych narzędzi; (ii) szczegółowa analiza czynników, które wpływają na ograniczenie reprodukowalności wyników na różnych etapach potoku przetwarzania.

Ocena piśmiennictwa

Analiza światowej literatury i bieżącego stanu wiedzy w omawianym obszarze zostały przeprowadzone w sposób właściwy i świadczą o dostatecznej wiedzy Autorki w tej dziedzinie. Przegląd literatury znajduje się głównie w rozdziałach drugim i trzecim. Cytowania dotyczą przede wszystkim narzędzi, które wchodzą w skład opracowanego przez Autorkę potoku przetwarzania, metod służących do oceny wyników, a także samych wyników badań porównawczych różnych grup algorytmów związanych z analizą danych RNA-seq/mikromacierzy i wynikających z nich najlepszych praktyk. W mojej opinii w pracy brakuje odniesienia i porównania do istniejących potoków przetwarzania do analizy RNA-seq i mikromacierzy.

Ocena zastosowanych metod badawczych

Szczegółowy opis opracowanego przez Autorkę potoku przetwarzania, w tym wykorzystanych narzędzi oraz zbiorów testowych służących do walidacji potoku został zaprezentowany w rozdziale trzecim. Potok został zaimplementowany z wykorzystaniem oprogramowania typu Workflow Management System o nazwie Snakemake. Jest to jeden z bardziej popularnych systemów wykorzystywanych przy implementacji potoków bioinformatycznych obok Workflow Description Language, Common Workflow Language oraz NextFlow.

Opracowany potok przetwarzania składa się z czterech modułów. Pierwszy jest przeznaczony do wykonania wstępnej analizy (uliniowienie, kontrola jakości). Moduł drugi służy do przeprowadzania analiz ekspresji różnicowej. W module trzecim dokonywana jest detekcja alternatywnego splicingu i dalsza analiza uzyskanych wyników. Moduł czwarty zajmuje się wizualizacją.

Moduł pierwszy zawiera narzędzia do uliniowienia odczytów (HiSat2 i Kallisto) oraz kontroli jakości (FastQC, MultiQC). Drugi moduł został opracowany w dwóch wariantach dla danych z mikromacierzy oraz RNA-seq. Zawiera on m.in.: (i) narzędzia do usuwania czynników zakłócających (ang. confounding factors) (SVA oraz SVA-seq); (ii) narzędzia do przeprowadzenia analizy ekspresji różnicowej (limma, DESeq2, EdgeR); (iii) metody analizy funkcjonalnej (pakiety go.DB i topGO). Moduł trzeci rozpoczyna swoje zadanie od wykrywania zdarzeń splicingowych przy użyciu narzędzia SplAdder. Następnie do analizy uzyskanych wyników wykorzystywane są narzędzia Bisbee oraz InterProScan. Ostatni moduł służy do wizualizacji wyników przy użyciu pakietów języka R.

Dobór narzędzi na poszczególnych etapach przetwarzania został uzasadniony w oparciu o wyniki przeprowadzonych analiz literaturowych. Podsumowując, potok przetwarzania został skonstruowany z wykorzystaniem najlepszych praktyk opisanych w najnowszych pracach naukowych. Dla każdego etapu analizy zostały dobrane odpowiednie zestawy algorytmów i dobór ten nie budzi zastrzeżeń.

Ocena części rozprawy doktorskiej dotyczącej omówienia wyników badań

Opis wyników został umieszczony w rozdziale czwartym i został podzielony na trzy części. W pierwszej (4.1) omówiono wyniki analiz dla dużego zbioru RNA-seq (badania bólu neuropatycznego u myszy). W drugiej zaprezentowano analizę działania fragmentu potoku (wykrywanie alternatywnego splicingu) dla zbioru benchmarkowego z konsorcjum SEQC2. W części trzeciej przedstawiono wyniki analiz ekspresji różnicowej dla zbioru danych mikromacierzowych pochodzących z badań pacjentów z chorobą Parkinsona.

Najbardziej obszerna część (4.1) prezentuje w pierwszej kolejności wyniki analizy ekspresji różnicowej (4.1.1). Interesującym wkładem autorki są przeprowadzone badania wpływu różnych ustawień narzędzia SVAseq na reprodukowalność ostatecznych wyników ekspresji różnicowej. Niestety, pomimo uzyskania lepszej separowalności na wykresie PCA dla jednego z ustawień, nie udało się potwierdzić istotnej poprawy reprodukowalności wyników. Jak piszę w dalszej części recenzji, w mojej opinii warto byłoby kontynuować ten wątek badawczy na większej liczbie różnorodnych zbiorów danych (z możliwie różnym stosunkiem sygnału do szumu).

W dalszej części (4.1.2) Autorka przeprowadziła analizę detekcji alternatywnych zdarzeń splicingowych. Uzyskane wyniki wskazują na znacznie wyższy poziom reprodukowalności niż w przypadku analizy ekspresji różnicowej. W sekcjach 4.1.3-4.1.5 analizowana jest liczba genów oraz termów z ontologii GO, które posiadają jednocześnie zdarzenia splicingowe różnego typu. Badanie przeprowadzone zostało osobno dla różnych rodzajów isoform (znanych i nowych) i wskazuje na niewielki stopień powtarzania się genów i termów dla zdarzeń splicingowych różnego typu. Autorka zaobserwowała również, że zdarzenia splicingowe występują częściej w grupach genów powiązanych z układem nerwowym. Jest to spodziewana obserwacja ze względu na typ analizowanej tkanki. W opisie wyników nie znalazłem informacji o poziomie istotności tej obserwacji. Warto byłoby również wskazać do czego tego typu analiza jest przydatna dla użytkownika.

Sekcja 4.1.6 opisuje wyniki analizy funkcjonalnej konsekwencji wybranych zdarzeń splicingowych i wskazuje na duże znaczenie nowych zdarzeń, nie uwzględnionych w referencyjnych adnotacjach.

W sekcji 4.1.7 gdzie Autorka zaprezentowała wizualizacje wybranych zdarzeń w genach związanych z układem nerwowym wysuwając jednocześnie hipotezę, że dla badanego zbioru danych zdarzenia w tych genach występują częściej. Niestety hipoteza ta nie została poparta żadnym testem statystycznym.

W sekcji 4.2 Autorka przeprowadza testy działania programu SplAdder na benchmark'owym zbiorze danych SEQC2. Wyniki analizy potwierdzają wysoką jakość działania programu. Wskazane zostały też potencjalne ograniczenia w detekcji zmian splicingowych wynikające z niedoskonałości technologii sekwencjonowania.

W sekcji 4.3 zostały zaprezentowane wyniki analiz dla danych mikromacierzowych. W szczególności zbadano wpływ różnych metod normalizacyjnych na uzyskane wyniki ekspresji różnicowej.

Podsumowując, prezentacja uzyskanych wyników pomaga zrozumieć mocne i słabe strony opracowanego potoku przetwarzania. Jednocześnie część analiz zyskałaby na znaczeniu gdyby zostały one przeprowadzone w sposób bardziej kompleksowy, w tym dla większej liczby zbiorów danych oraz z wykorzystaniem narzędzi statystycznych pozwalających na weryfikację hipotez zawartych w tej części rozprawy.

Słabe strony rozprawy

Opracowany przez Doktorantkę potok przetwarzania został zweryfikowany na trzech wybranych zbiorach danych, w tym kompleksowa analiza wszystkich elementów potoku została wykonana tylko na jednym zbiorze danych pochodzącym z rzeczywistych badań dotyczących bólu neuropatycznego u myszy. Chociaż z jednej strony jest to mocna strona, jeśli chodzi o wykazanie praktycznego zastosowania potoku, to uwzględnienie niewielkiej liczby zbiorów danych znacznie ogranicza możliwość uogólnienia wyników. Korzystne byłoby przetestowanie potoku na szerszym zakresie zestawów danych.

Jedną z konkluzji w kontekście analizy ekspresji różnicowej (zarówno w przypadku RNA-seq jak i mikromacierzy) był niski poziom uzyskanej reprodukowalności wyników. Taki wniosek w niewielkim stopniu pogłębia wiedzę na temat praktycznych możliwości zastosowań tej części potoku na rzeczywistych (nie benchmarkowych) zestawach danych. Jednocześnie, wydaje się, że bardziej pogłębiona analiza problemu, uwzględniająca większą liczbę zbiorów danych, inny dobór i lepszą kalibrację narzędzi mogłaby pozwolić lepiej zidentyfikować czynniki, które wpływają na ograniczenie reprodukowalności. Ciekawym wątkiem badawczym wydaje się również próba zdefiniowania właściwości zbiorów danych, które muszą zostać spełnione aby reprodukowalność wyników została zachowana.

W mojej opinii w pracy zabrakło również szerszej dyskusji i porównania do istniejących potoków przetwarzania. Choć Autorka stwierdza, że jej potok jest bardziej kompleksowy od istniejących rozwiązań (obejmuje zarówno analizę ilościową jak i jakościową), to nie przedstawia bezpośredniego porównania do konkretnych programów. W szczególności cenne byłoby porównanie funkcjonalności do najbardziej popularnych metod takich jak potoki z bazy nf-core (https://nf-co.re/pipelines?q=rna-seq). Warto również zwrócić uwagę, że niektóre potoki (np. https://github.com/kids-first/kf-rnaseq-workflow) poza możliwościami wykonywania różnicowej analizą ekspresji genów oraz wykrywania alternatywnego splicingu, pozwalają także na poszukiwanie fuzji genowych. W pracy Autorka nie poruszyła również problemów związanych z wydajnością i skalowalności opracowanego potoku.

Autorka w swoich badaniach (sekcje 4.1.3-4.1.5) zaobserwowała niski poziom powtarzalności genów i termów dla różnych zdarzeń splicingowych. Niestety, nie znalazłem w pracy informacji jakie znaczenie posiadają uzyskane wyniki. W mojej opinii ryciny i tabele w sekcjach 4.1.3-4.1.5 było by łatwiej zinterpretować gdyby zostały zestawione razem.

W sekcji 4.1.6 Autorka zaobserwowała duże znaczenie nowych zdarzeń splicingowych przy interpretacji konsekwencji funkcjonalnych wybranych zdarzeń splicingowych. Przydatne byłoby oszacowanie istotność statystycznej dla powyższej hipotezy.

Przy rycinach dość często brakuje wystarczającego opisu. Przykładowo ryc. 4.15 zawiera kilka fragmentów, które częściowo są opisane w głównym tekście rozprawy ale brakuje ich w legendzie.

Zastosowania praktyczne

Praktyczne zastosowania opracowanego przez autorkę potoku analizy danych RNA-seq wydają się być znaczące. Potok przetwarzania ma potencjał do zastosowania w szerokim zakresie badań transkryptomicznych, w szczególności dotyczącymi alternatywnym składaniem transkryptów i funkcjonalnymi analizami genów. Potok ten został zweryfikowany na rzeczywistych zestawach danych, wykazując jego gotowość do zastosowania w praktycznych badaniach. Autorka również

zwraca uwagę na dostępność danych i kodu, co jest kluczowe dla reprodukowalności i dalszego rozwoju jej pracy przez innych naukowców.

Ocena czy rozprawa stanowi oryginalne rozwiązanie problemu naukowego?

W mojej opinii rozprawa stanowi oryginalne rozwiązanie problemu naukowego. Autorka zaprojektowałą, zaimplementowała i przetestowała na rzeczywistych i benchmarkowych zbiorach danych kompleksowy potok przetwarzania. Przeprowadzona analiza wykazała jego ograniczenia (niska replikowalność wyników analiz ekspresji różnicowej) oraz mocne strony (wiarygodna detekcja zdarzeń splicingowych).

Ocena czy rozprawa prezentuje ogólną wiedzę teoretyczną oraz umiejętność samodzielnego prowadzenia pracy naukowej?

Przeprowadzony przegląd literatury, dobór odpowiednich metod i konstrukcja w pełni funkcjonalnego potoku przetwarzania, wykonane eksperymenty oraz wyciągnięte z nich wnioski wskazują na ogólną wiedzę teoretyczną i umiejętność samodzielnego prowadzenia pracy naukowej przez Autorkę. Jednocześnie warto zaznaczyć, że doktorantka jest współautorką trzech publikacji w czasopismach Genome Biology, Frontiers in Genetics oraz Journal of Molecular Sciences.

Wniosek końcowy

Wystawiam pozytywną ocenę rozprawie doktorskiej mgr inż. Agaty Muszyńskiej pt. „Advanced data exploration techniques for augmented transcriptional landscape and its better quantification". Stwierdzam, że praca w mojej opinii spełnia wymagania i warunki nakładane przez ustawę o stopniach naukowych i wnoszę o dopuszczenie doktorantki do obrony pracy w celu uzyskania stopnia doktora nauk technicznych w dziedzinie nauk inżynieryjno-technicznych w dyscyplinie informatyka techniczna i telekomunikacja.

Tomasz Gambin