Warszawa, 20 czerwca 2024

prof. dr hab. Anna Gambin
Instytut Informatyki
Uniwersytetu Warszawskiego

**RECENZJA ROZPRAWY DOKTORSKIEJ**

***mgra Przemysława Latocha***

**zatytułowanej**

ANALIZA EKSPRESJI GENÓW W SZEREGACH CZASOWYCH NA PODSTAWIE DANYCH Z SEKWENCJONOWANIA TRANSKRYPTOMÓW ORAZ TRANSLATOMÓW

## 1. Problem badawczy i jego znaczenie

Rozprawa doktorska prezentuje nowe narzędzie służące do analizy danych RNAseq oraz RIBOseq. Technologie RNAseq i RIBOseq służą do sekwencjonowania RNA, umożliwiając analizę ekspresji genów. RNAseq mierzy ilości mRNA, pozwalając na identyfikację aktywnych genów i ich poziomów ekspresji, natomiast RIBOseq sekwencjonuje cząsteczki RNA związanego z rybosomami, co pozwala na badanie translacji białek. Opracowana przez Autora rozprawy biblioteka w języku R i nazwana `TimeSeqR` pozwala na analizę jednocześnie wykonywanych eksperymentów RNAseq i RIBOseq, co umożliwia identyfikację wzorców ekspresji genów oraz kwantyfikację efektywności translacji.

Funkcjonalność prezentowanego narzędzia obejmuje wstępne etapy analizy danych, takie jak kontrola jakości, oczyszczanie, mapowanie do genomu referencyjnego i zliczanie odczytów. W swojej głównej części zaimplementowana biblioteka pozwala na normalizację danych, analizę różnicowej ekspresji genów w czasie, klasteryzację i analizę funkcjonalną interesujących grup genów.

Rozprawa zawiera instrukcje instalacji, konfiguracji, szczegóły implementacji, opis interfejsu graficznego oraz przykłady użycia biblioteki dla danych z *Bacillus subtilis, Saccharomyces cerevisiae* oraz danych symulowanych.

Tytuł rozprawy uważam za dość mylący, gdyż problem badawczy jaki jest w niej zawarty dotyczy opracowania zintegrowanego narzędzia do analizy danych z RNAseq i RIBOseq, umożliwiającego badanie ekspresji genów w szeregach czasowych, a sama analiza ekspresji genów zaprezentowana w rozprawie jest dość uboga i nie zawiera biologicznie istotnych wniosków.

Nie neguję potrzeby skuteczniejszego łączenia i analizowania danych z obu wspomnianych technologii, aby lepiej zrozumieć dynamikę ekspresji genów i translacji w czasie, jednak zaproponowane narzędzie nie zostało moim zdaniem wystarczająco porównane z istniejącymi rozwiązaniami.

## 2. Wkład Autora

Główne dokonanie Autora rozprawy dotyczy zaprojektowania oraz implementacji, w języku do obliczeń statystycznych R, biblioteki `TimeSeqR`. Biblioteka ta została stworzona do analizy sekwencji czasowych i oferuje zestaw funkcji umożliwiających obróbkę danych tego typu.

W rozprawie przedstawiono opis funkcjonalności narzędzia, zawierający omówienie jego modułów oraz możliwości. Opisano metody i algorytmy zaimplementowane w bibliotece, które pozwalają na przeprowadzanie analiz statystycznych oraz wizualizacji danych.

Autor przedstawił również kilka przykładów użycia `TimeSeqR`, pokazując zastosowania biblioteki w różnych scenariuszach badawczych. Przykłady te ilustrują, jak za pomocą `TimeSeqR` można analizować dane sekwencji czasowych, co podkreśla wszechstronność narzędzia.

Dodatkowo, w rozprawie opisano protokół instalacji biblioteki, co umożliwia jej wdrożenie na różnych systemach operacyjnych. Przedstawiono również interfejs graficzny, który został zaprojektowany w celu ułatwienia pracy użytkownikom preferującym interaktywne środowisko analityczne. Opisano jego funkcje i sposób obsługi, co pozwala na korzystanie z narzędzia.

Na koniec, Autor omówił możliwe rozszerzenia biblioteki, sugerując kierunki przyszłych prac nad jej rozwojem. Wskazano na potencjalne usprawnienia oraz dodatkowe funkcje, które mogą zostać wprowadzone, aby zwiększyć jej funkcjonalność.

Mgr Latoch jest współautorem sześciu publikacji powiązanych z rozprawą. Wszystkie publikacje są wieloautorskie (od 5 do 15 autorów) a mgr Latoch dzieli status pierwszego autora w dwóch z nich. W pozostałych nie jest pierwszym autorem, co sugeruje, że jego wkład koncentrował się głównie na wspieraniu zespołów badawczych poprzez przeprowadzanie analiz bioinformatycznych.

Publikacje nie stanowią spójnego cyklu i mają dość różnorodną tematykę. Poruszane zagadnienia obejmują metody wzbogacania mRNA i analizę transkryptomu, badania mikrobiomu pszczół miodnych oraz analizy paleolimnologicznych jezior i ich historii ekologicznej. Trzy z nich ukazały się w dobrych czasopismach (*Nucleic Acid Research, Scientific Reports, Biochimica et Biophysica Acta* ), pozostałe trzy w dość przeciętnych (*Journal of Visualized*

*Experiments, Pathogens, Data in Brief*).

Podsumowując wkład Autora, oceniam go jako minimalnie spełniający wymagania stawiane rozprawom doktorskim. Stworzona biblioteka `TimeSeqR` wykorzystuje zaproponowane wcześniej narzędzia i w żaden sposób nie rozwija metod obliczeniowych służących analizie danych z technologii sekwencjonowania. Biblioteka opiera się na istniejących technikach, nie wprowadzając nowych rozwiązań ani innowacji w zakresie analiz bioinformatycznych.

Ze względu na to, że rozprawa jest przedstawiana w dyscyplinie nauk biologicznych, powyższe nie stanowiłoby znaczącego zarzutu, gdyby dzięki użyciu nowego narzędzia zaprezentowano ciekawe odkrycia związane z biologią molekularną. Takie odkrycia mogłyby stanowić wartość dodaną, rekompensując brak innowacji w samych metodach analizy. Niestety, rozprawa nie prezentuje takich wyników, co ogranicza jej znaczenie naukowe.

Dodatkowo, rozprawa nie zawiera wystarczającego porównania z istniejącymi metodami analizy. Bardzo ograniczone porównanie z innymi narzędziami uniemożliwia ocenę, czy `TimeSeqR` oferuje jakiekolwiek przewagi w stosunku do już dostępnych rozwiązań.

W związku z powyższym, wkład Autora, choć technicznie poprawny i zgodny z wymaganiami formalnymi, nie wnosi istotnego postępu ani w dziedzinie biologii molekularnej, ani w metodologii analiz bioinformatycznych

## 5. Inne uwagi

Poniższe uwagi uzasadniają moje zdanie o rozprawie oraz zawierają sugestie poprawek.

1. Rozdział wstępny jest zbytnio rozbudowany w opisie technologii sekwencjonowania, repozytoriów danych i algorytmów nie powiązanych bezpośrednio z pracą. Często brakuje struktury, co powoduje, że opis staje się nieczytelny (Rozdział 1.3.3). Prezentacja algorytmów i metod statystycznych nie zawsze jest poprawna. W języku polskim używamy określenia prawdopodobieństwo (nie szansa), rozkład (nie dystrybucja), P-wartość (nie wartość P) oraz liczba (nie ilość) w przypadku policzalnych wielkości. Natomiast opis metod, które zostały użyte w bibliotece jest bardzo chaotyczny i niewystarczający. Metody analizy różnicowej, zwłaszcza te, które zostały zaimplementowane w bibliotece powinny być przedstawione bardzo starannie i najlepiej z użyciem pseudokodu.

2. Rozdział 3 zawiera zarówno opis funkcjonalności narzędzia jak też treści, które umożliwiają instalację i rozbudowę biblioteki. Taki układ jest bardzo nieczytelny. Czysto techniczne rozwiązania związane z implementacją (generowanie wykresów etc) oraz listingi kodu funkcji powinny znaleźć się w Apendiksie, jako podręcznik użytkowni-

ka, czyli integralna część stworzonego narzędzia. Dotyczy to również opisu interfejsu graficznego oraz zaprezentowanych na stronach 98 – 106 przykładów wizualizacji.

3. Nie jest jasne, jaki jest status dostępności biblioteki. Czy została ona zgłoszona jako pakiet R do repozytorium CRAN? jako pakiet na platformie Bioconductor ? Na jakiej licencji ?

4. Kluczowa funkcjonalność zaproponowanej biblioteki obejmuje analizę różnicową genów. Do wykonania tej pierwszej użyto narzędzia `ImpulseDE2`. W pracy wymieniono niektóre alternatywne podejścia (`DESeq2`, `edgeR`, `limma-voom`, `Cuffdiff2`, `EBSeq`, `maSigPro`) pomijając inne (np. `GP-Seq`). Niestety działanie biblioteki w analizie ekspresji genów porównano jedynie z metodą `DESeq2`, a porównanie to ogranicza się do przedstawienia diagramów Venna (Rys 31 i 35 w pracy). Sam Autor stwierdza, że w przypadku genów znalezionych jedynie przez jego narzędzie konieczna jest dalsza analiza. Bez zaprezentowania tej analizy czytelnik może mieć obawy, czy zidentyfikowane geny nie są tzw fałszywymi pozytywami.

5. Kolejna funkcjonalność biblioteki obejmuje klasteryzację genów. Do tego zadania wybrano narzędzie `mfuzz` oparte o algorytm rozmytych k-średnich. W tym przypadku zbiór alternatywnych podejść jest jeszcze liczniejszy, a wybór `mfuzz` nie został w żaden sposób uzasadniony, mimo znanych ograniczeń metody (niejasność interpretacji, duża złożoność, wrażliwość na parametry). Nie znalazłam też porównania z innymi metodami grupowania wzorców ekspresji, choć z zaprezentowanych w Rozdziale 4 wizualizacji wynika jednoznacznie, że zaproponowane przez `mfuzz` klastrowanie jest bardzo trudne w interpretacji.

6. Jeśli sugerujemy się tytułem, to Rozdział 4 zawiera główne wyniki rozprawy. W przypadku pierwszego zestawu danych (*Bacillus Subtilis*) po dość szczegółowym opisie wykonania eksperymentu (przygotowanie próbek, sekwencjonowanie), następuje lakoniczny opis użycia biblioteki. Czytelnik dowiaduje się jedynie, ile genów zostało zidentyfikowanych jako istotne przez daną metodę. Rysunek 26 jest dość niepokojący, gdyż różnica symetryczna dla zbiorów genów identyfikowanych przez `TimeSeqR` i metodę `DESeq2` jest znacząca. Narzuca się pytanie jakie geny znajdują się w tej różnicy symetrycznej? W jakie procesy molekularne są zaangażowane i jakie funkcje pełnią ? Czy ich interpretacja w kontekście danego eksperymentu może potwierdzić wyższość prezentowanej biblioteki nad istniejącymi metodami?

7. Drugi zbiór danych analizowany w Rodziale 4 pochodzi z organizmu *Saccharomyces cerevisiae*. Wyniki analiz nie są w żaden sposób zinterpretowane, a porównanie z

metodą `DESeq2` pokazuje znaczące rozbieżności. Nie zostało wyjaśnione, czy można je interpretować na korzyść `TimeSeqR`. Uważam, że konieczne jest scharakteryzowanie 1682 genów wskazanych przez bibliotekę i ich interpretacja w kontekście publikacji, w której został pierwotnie opisany eksperyment. Czy zastosowanie `TimeSeq2` pozwala na wzmocnienie wniosków z tej publikacji ? Czy raczej podważa postawione tam hipotezy badawcze ?

8. Ostatni przykład wykorzystania biblioteki obejmuje dane symulowane. Dane zostały wygenerowane za pomocą funkcji dostarczanej przez narzędzie `ImpulseDE2`. Opisano jedynie parametry generatora, a najważniejsze informacje o tym z jakiego rozkładu generowane są dane zostały pominięte. Analiza rozpoczyna się od ustalenia **liczby** klastrów. Podobnie jak w poprzednich przypadkach maksymalna liczba klastrów ustawiona jest na 20. Brakuje uzasadnienia tego ograniczenia. Następnie okazuje się, że uzyskane metodą `mfuzz` klastrowanie jest dość niskiej jakości. Narzuca się tutaj rozszerzenie biblioteki przez implementację co najmniej jednej alternatywnej metody klastrowania. Porównanie z metodą `DESeq2` wykazuje jej przewagę nad `TimeSeqR` w identyfikacji różnicujących genów. W tym przypadku nie ma możliwości wykonania analizy funkcjonalnej, jednak pominięcie 139 genów wydaje się niepokojące. Czy można to wytłumaczyć specyfiką generatora oraz modelem impulsowym zastosowanym przez `ImpulseDE2`, który jest wykorzystywany w `TimeSeqR` ?

9. Rozdział 5 rozprawy jest poświęcony dyskusji różnych narzędzi służących do analizy danych RNASeq. Niektóre z nich zostały użyte przez Autora do wstępnego procesingu danych a inne weszły w skład biblioteki. Jest też zbiór narzędzi, które są jedynie wymienione, ale ich funkcjonalność nie została porównana z proponowanym rozwiązaniem. Rozdział w znacznym stopniu jest redundantny z rozdziałem wprowadzającym. Problemem jest też brak struktury tekstu, który jest z tego powodu bardzo nieczytelny. Wymienione narzędzia powinny być adekwatnie pogrupowane względem funkcjonalności, a następnie porównane pod kątem jasno sprecyzowanych i adekwatnych kryteriów. Stosowane kryteria powinny być odpowiednio uzasadnione, a ostateczna walidacja narzędzia musi obejmować interpretację uzyskanych wyników w kontekście rozważanego zagadnienia biologii molekularnej.

Poniższe uwagi mają mniejszą wagę i dotyczą czytelności rozprawy:

1. strona 17: ilość odczytów → liczba odczytów; w języku polskim nie stosuje się powszechnie przedimków – np w wyrażeniach „ten nanopor", „ ta mobilność";

2. strona 29: ilość znaków → liczba znaków;

3. strona 31: poprawny zapis formuł rekurencyjnych w algorytmach Needlemana-Wunscha oraz Smitha-Watermana powinien zawierać funkcję maksimum oraz nie używać notacji wektorowej;

4. strona 32: nie mówimy, że algorytmy nie są wykonalne (bo są), tylko, że ich złożoność obliczeniowa jest (zbyt) duża; ilość sekwencji $\rightarrow$ liczba sekwencji;

5. strona 33: metody iteracyjne zbiegają do pewnego rozwiązania, najczęściej wymagamy, żeby były zbieżne i staramy się oszacować szybkość zbieżności;

6. strony 34 – 38: opis metod mapowania sekwencji NGS jest nieczytelny, brakuje struktury tekstu, odpowiedniego grupowania przedstawianych metod i porównania ich za pomocą sensownych kryteriów;

7. strona 42: czy „moc wykrywania" oznacza moc testu statystycznego ? jak jest zdefiniowana?

8. strona 43: wielkość $\Phi$ nie została zdefiniowana, podobnie jak $\alpha$;

9. strona 44: wykorzystanie Twierdzenia Bayesa w metodzie `DESeq2` jest niejasno opisane;

10. strony 48 – 52: opis kluczowej metody, czyli `ImpulseDE2` jest niestaranny; dlaczego stosujemy algorytm BFGS dla problemu optymalizacyjnego? jaki jest warunek stopu algorytmu? jak wygląda test ilorazu wiarygodności w analizie ekspresji? dlaczego wykorzystujemy rozkład Chi-kwadrat do obliczenia p-wartości? dla kompletności należy też przedstawić metodę korekty w przypadku testowania wielu hipotez;

11. strona 55: co oznacza w kontekście omawiania algorytmu c-średnich sformułowanie *iteracja Picarda*?

12. rozdział 3 (strony 57-107): listingi zawierające kod poszczególnych fragmentów biblioteki powinny (w adekwatnym kontekście) znaleźć się w dokumencie zwanym zazwyczaj podręcznik użytkownika, a w tekście rozprawy kluczowe algorytmy mogą zostać zaprezentowane w postaci pseudokodu;

13. rysunki 8 i 9 prezentujące konstrukcję biblioteki (strony 84 i 85 w rozprawie) są nieczytelne;

14. w Rozdziale 4.1.3 sformułowania ilość klastrów/punktów należy zastąpić przez liczba klastrów/punktów;

15. jak interpretować wykresy PCA dla klastrów ? czy nakładanie się klastrów nie dyskwalifikuje metody użytej klastrowania rozmytego?

16. jakie grupy genów reprezentują klastry przedstawione na Rysunkach 22 i 23 ?

17. rysunki prezentujące wizualizacje uzyskiwane dzięki bibliotece (strony 102 i 103 w rozprawie) są nieczytelne;

18. używane w rozprawie oznaczenie *padj* na oznaczenie granicznej wartości p-warości jest bardzo nieczytelne;

19. jakie grupy genów reprezentują klastry przedstawione na Rysunku 28 ?

**8. Podsumowanie**

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez artykuł 13 Ustawy z dnia 14 marca 2003 r. o stopniach i tytule naukowym (z późniejszymi zmianami [1]) stwierdzam, że recenzowana przeze mnie praca spełnia w stopniu minimalnym wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy i wnoszę o dopuszczenie magistra Przemysława Latocha do dalszych etapów przewodu doktorskiego.

[1] https://isap.sejm.gov.pl/isap.nsf/home.xsp