



Prof. Dr hab. Wojciech M. Karłowski  
Zakład Biologii Obliczeniowej  
Wydział Biologii  
Uniwersytet im. Adama Mickiewicza w Poznaniu

Poznań 04.06.2024

**Recenzja rozprawy doktorskiej mgr Przemysława Latocha pt. "Analiza ekspresji genów w szeregach czasowych na podstawie danych z sekwencjonowania transkryptomów oraz translatomów"**

Przedstawiona do recenzji praca doktorska została wykonana w Instytucie Nauk Biologicznych na Uniwersytecie Marii Curie-Skłodowskiej w Lublinie pod opieką dr hab. Agaty Starosty oraz dr Bartłomieja Balcerzaka. Opisany przez Doktoranta projekt obejmował stworzenie narzędzia obliczeniowego przeznaczonego do analizy danych pochodzących z eksperymentów sekwencjonowania RNA-Seq oraz Ribo-Seq. Napisana w języku R biblioteka TimeSeqR umożliwia normalizację danych, prowadzenie analiz różnicowej ekspresji genów w czasie, grupowanie genów wykazujących podobny wzór ekspresji oraz analizy funkcji molekularnych tych genów. W pracy przedstawiono szczegółowo instrukcje instalacji oraz konfiguracji środowiska, dostępne funkcje biblioteki oraz opis interfejsu graficznego. Ponadto przedstawiono przykłady zastosowania biblioteki TimeSeqR na danych symulowanych oraz uzyskanych z bakterii (*Bacillus subtilis*) oraz drożdży.

Praca zorganizowana jest w formie sześciu głównych części. Poza spotykanyymi w tego typu opracowaniach Wstępem, Dyskusją i Podsumowaniem zawiera trzy działy poświęcone opisowi przygotowania danych do analiz, implementacji biblioteki TimeSeqR oraz przykładów wykorzystania opracowanego narzędzia.

Wstęp jest bardzo rozbudowany i obejmuje wprowadzenie do technologii sekwencjonowania oraz opis technologii obliczeniowych związanych z analizą sekwencji. Z jednej strony Doktorant wprowadza elementarne techniki biologii molekularnej i pojęcia bioinformatyczne, a z drugiej zagłębia się w formalne aspekty działania algorytmów analizy sekwencji. W tej części pracy zauważyłem parę nieścisłości. Na przykład koszty poznania *de novo* sekwencji genomu są porównywane z kosztami re-sekwencjonowania (strona 13). Ponadto, technologia

sekwencjonowania PacBio jest zaliczana zwyczajowo do trzeciej generacji sekwencjonowania (technologie SMRT sequencing; strona 16).

Druga część pracy opisuje procedury stosowane przez Autora w przygotowaniu danych sekwencyjnych do analiz. Zaprezentowane procedury są zgodne z ogólnie przyjętymi zasadami wykonywania tego etapu analizy danych NGS. Autor stosuje tutaj popularne i ogólnie akceptowane narzędzie obliczeniowe. Podobnie jak w przypadku wstępu, opisy są bardzo szczegółowe.

Trzecia część pracy zawiera opis wdrożenia w języku R biblioteki TimeSeqR, z wykorzystaniem popularnych narzędzi obliczeniowych. Działanie biblioteki obejmuje: normalizację danych, różnicową analizę ekspresji genów w przebiegu czasowym, grupowanie genów wykazujących podobny profil ekspresji w czasie oraz adnotację funkcjonalną grup (klastrow) genów. Zgodnie z instrukcją zawartą w pracy (rozdział 3.2) podjąłem próbę instalacji biblioteki z wykorzystaniem załączonego na płycie CD skryptu instalacyjnego. Mimo pozytywnego komunikatu 'All done!' na końcu procesu instalacji część z bibliotek środowiska R nie została poprawnie zainstalowana. Napotkane przeze mnie problemy wskazują na konieczność dopracowania w przyszłości automatycznej procedury instalacji biblioteki. Mimo tych perturbacji, biblioteka TimeSeqR została przeze mnie zainstalowana i możliwe było uruchomienie graficznego panelu użytkownika. Panel użytkownika jest prosty i w sposób intuicyjny umożliwia prowadzenie obliczeń. Dynamicznie pojawiające się opcje w zależności od wybranego trybu analizy oferują funkcjonalność oczekiwaną w tego typu aplikacjach. Brak danych do testów niestety uniemożliwił dalszą ewaluację możliwości interfejsu biblioteki. Wnioskując jednak z opisu zamieszczonego w rozprawie doktorskiej, dostępne opcje wizualizacji oraz prezentacji danych powinny spełniać oczekiwania dla analiz ekspresji genów w przebiegach czasowych. W tej części pracy odniosłem jednak wrażenie, że w niektórych przypadkach opisy wykorzystanych programów stanowią dokładne tłumaczenie instrukcji załączonych do stosowanych narzędzi. Szczególnie jest to widoczne w przypadku programu GMT Helper (strona 80). Uważam, że wszystkie przypadki zamieszczenia tłumaczenia oryginalnego opracowania powinny być jednoznacznie zaznaczone w tekście rozprawy.

Ostatnia z głównych części rozprawy doktorskiej mgr Przemysława Latocha zawiera opis przykładów wykorzystania biblioteki z użyciem danych pochodzących z bakterii (*Bacillus subtilis*) oraz drożdży. Trochę mnie zaskoczył bardzo szczegółowy opis części eksperymentalnej w której, jak zrozumiałem z informacji zawartych w rozprawie (strona 108), Autor nie uczestniczył. Dodatkowo czuję pewien niedosyt wynikający z bardzo lakonicznego opisu wyników uzyskanych analiz. W związku chciałbym prosić o wyjaśnienie paru kwestii: (i) Czy można porównać jakie elementy (geny) wchodzi w skład poszczególnych grup przedstawionych na rys. 22 i 23? (ii) Jaka jest adnotacja funkcjonalna tych grup genów? (iii) Czy wyniki

przedstawione na rys. 26 oraz 31 można istotnie interpretować w kontekście przewagi jednej metody nad drugą? (iv) Czy analiza danych symulowanych nie jest dobrą okazją do przeprowadzenia formalnego porównania metod w oparciu o parametry statystyczne (czułość, precyzja, itp.)? Brak porównania parametrów statystycznych działania metody prezentowanej w rozprawie nie kwestionuje jednak jej użyteczności. Autor przedstawia listę trzech publikacji, które zawierają wyniki uzyskane z zastosowaniem prezentowanej metodologii, co jednoznacznie potwierdza jej wartość jako użyteczne narzędzie do badania zmian ekspresji genów w przebiegach czasowych w oparciu o dane z eksperymentów RNA-Seq i Ribo-Seq.

Rozprawa przygotowana jest starannie. Niestety w wielu miejscach Doktorant wykorzystuje terminy i hasła, które niepotrzebnie zostały zaczerpnięte z języka angielskiego lub zostały niefortunnie przetłumaczone. Na przykład: 'bazy' dla określenia zasad (strona 30, 60 i 64), wersja 'terminalowa' programu (strona 59), 'trymowanie' (strona 60), 'rodzic transkryptu dla eksonów' (strona 64), 'odciski rybosomalne' (strona 111), 'kody kreskowe' (strona 112). Nie jestem także przekonany, czy prezentacja kodu źródłowego funkcji i modułów w głównym tekście pracy jest rozwiązaniem optymalnym. Jeżeli Autor uważa drukowaną prezentację kodu jako konieczną, to moim zdaniem powinno się to znaleźć w osobnym załączniku do pracy.

Mimo pewnych niedociągnięć przedstawiona rozprawa potwierdza kompetencje Autora w zakresie tworzenia nowych i użytecznych narzędzi obliczeniowych oraz analiz danych pochodzących z sekwencjonowania NGS w kontekście różnicowej ekspresji genów w ciągach czasowych. Zgodnie z deklaracją zawartą w dokumencie, przygotowane narzędzie zostało wykorzystane w projektach, których wyniki ukazały się w trzech publikacjach. W związku z powyższym, stwierdzam, że przedstawiona do recenzji praca doktorska spełnia formalne wymagania stawiane przez Ustawę o stopniach i tytułach naukowych jak i zwyczajowe warunki stawiane rozprawom doktorskim i wnoszę do Rady Dyscypliny Nauki Biologiczne Uniwersytetu im. Marii Curie-Skłodowskiej w Lublinie o dopuszczenie mgr Przemysława Latocha do dalszych etapów przewodu doktorskiego.

Wojciech Korcusz