ul. Noskowskiego 12/14, 61-704 Poznań
tel.: centrala 61 852 85 03, sekretariat 61 852 89 19
faks: 61 852 05 32, e-mail: ibch@ibch.poznan.pl
REGON 000849327 NIP 777-00-02-062
http://www.ibch.poznan.pl

dr hab. Barbara Uszczyńska-Ratajczak
Zakład Biologii Obliczeniowej Niekodującego RNA
Instytut Chemii Bioorganicznej
Polskiej Akademii Nauk
Noskowskiego 12/14
61-704 Poznań

Recenzja
rozprawy doktorskiej mgr. Przemysława Latocha, zatytułowanej: *„Analiza ekspresji genów w szeregach czasowych na podstawie danych z sekwencjonowania transkryptomów oraz translatomów”*

Przedstawiona do recenzji praca doktorska Pana mgr. Przemysława Latocha została wykonana w Instytucie Nauk Biologicznych Uniwersytetu Marii Curie-Skłodowskiej w Lublinie. Promotorem rozprawy jest dr hab. Agata Starosta, a promotorem pomocniczym dr Bartłomiej Balcerzak.

Tematyka pracy koncentruje się na zaawansowanych i wysokoprzepustowych analizach transkryptomicznych, wykorzystujących dwa rodzaje danych: sekwencjonowanie RNA (RNA-seq) oraz profilowanie rybosomów (RIBO-seq). Sekwencjonowanie RNA umożliwia analizę ekspresji genów i struktur transkryptów, podczas gdy RIBO-seq, będące wyspecjalizowaną formą sekwencjonowania RNA, pozwala z dużą precyzją określić wydajność procesu translacji. Wynikiem tych badań jest opracowanie TimeSeqR – zintegrowanego podejścia w postaci nowej biblioteki w języku R, umożliwiającej jednoczesną analizę obu typów danych, a tym samym równoczesne śledzenie zmian ekspresji genów i aktywności ich translacji w czasie. Ponadto, zaprezentowano działanie narzędzia TimeSeqR z wykorzystaniem zarówno danych symulacyjnych, jak i rzeczywistych danych biologicznych, uzyskanych dla organizmów bakteryjnych (*Bacillus subtilis*) oraz eukariotycznych (*Saccharomyces cerevisiae*).

Przedstawiony w pracy problem badawczy jest niezwykle aktualny. W ostatnich latach jesteśmy świadkami szybkiego rozwoju nowych technologii sekwencjonowania RNA, które umożliwiają uzyskanie większej liczby dłuższych odczytów przy niższych kosztach. Metody te znajdują także zastosowanie w coraz nowszych kontekstach biologicznych. Jednakże głównym ograniczeniem stosowania technik sekwencjonowania RNA jest nadal złożony proces analizy danych. Potencjalne trudności wynikają często z dużego rozmiaru danych oraz braku jasno określonych standardów ich przetwarzania. Ponadto, dostępne narzędzia są zazwyczaj mocno wyspecjalizowane, a ich efektywne wykorzystanie często wymaga ugruntowanej wiedzy z zakresu bioinformatyki lub biologii obliczeniowej.

Opracowana w ramach niniejszej pracy biblioteka TimeSeqR to unikalne rozwiązanie umożliwiające jednoczesną analizę aktywności genów na poziomie transkrypcji i translacji. Równoczesna

analiza danych RNA-seq i RIBO-seq pozwala nie tylko zminimalizować zmienność techniczną wynikającą z różnic w przetwarzaniu danych, ale także efektywniej i precyzyjniej identyfikować geny wykazujące zmiany aktywności na każdym z badanych poziomów. Zaimplementowany w TimeSeqR scenariusz przetwarzania danych obejmuje wstępną analizę, w tym: normalizację danych, identyfikację genów o różnicowej ekspresji i ich klastrowanie. Dostępna jest także ścieżka analizy wyższego rzędu, uwzględniająca funkcjonalną charakterystykę genów, badanie współekspresji (ang. *co-expression analysis*) oraz graficzną reprezentację otrzymanych wyników, co znacząco ułatwia ich interpretację. TimeSeqR prowadzi zatem użytkownika przez wszystkie kluczowe aspekty wysokoprzepustowej analizy danych uzyskiwanych metodami sekwencjonowania drugiej generacji (ang. *next-generation sequencing*). Biblioteka ta charakteryzuje się wszechstronnością także na poziomie uruchamiania analizy. W zależności od preferencji i poziomu zaawansowania użytkownika, może być ona bezpośrednio uruchomiona w powłoce systemu Unix (ang. *bash shell*), w środowisku programistycznym RStudio lub za pomocą interfejsu graficznego w przeglądarce internetowej. Kluczowym aspektem użyteczności TimeSeqR jest także możliwość dostosowania procesu analizy do typu analizowanych danych, w tym określenia liczby punktów czasowych i serii, w jakich zbierano dane. Chociaż architektura TimeSeqR nie jest do końca modułowa, dostępne wersje scenariuszy analizy danych obejmujące (i) pełną analizę, (ii) analizę pojedynczego eksperymentu, (iii) klastrowanie genów lub (iv) proces ustalania liczby klastrów, zapewniają użytkownikowi swobodę działania. To właśnie elastyczność bioinformatycznego narzędzia gwarantuje jego użyteczność i niekiedy nawet ponadczasowy charakter. W tym kontekście chciałam zapytać, *jak wyobraża Pan sobie dalszy rozwój (a może i ewolucję) technik profilowania transkryptomu oraz translatomu na dużą skalę?* Coraz większą popularność zyskują metody sekwencjonowania trzeciej generacji (ang. *third generation sequencing, TGS*), które charakteryzują się znacznie dłuższymi odczytami, ale również mniejszą przepustowością sekwencjonowania (relatywnie niską całkowitą liczbą otrzymywanych odczytów). Rozwój technologii TGS aktualnie skupia się na zwiększeniu głębokości sekwencjonowania, co prawdopodobnie z czasem pozwoli wykorzystywać te metody także do badania ekspresji genów. W związku z tym chciałam zapytać, *czy TimeSeqR w obecnej formie byłby kompatybilny z charakterystyką danych TGS oraz jakich ewentualnych modyfikacji wymagałaby możliwość zastosowania tej biblioteki do analizy tego typu danych?* Kolejne z moich pytań dotyczy zdolności analizy genów wykazujących różnice aktywności na badanych poziomach. *Jak TimeSeqR radzi sobie z identyfikacją i klasyfikacją genów wykazujących skrajne różnice w aktywności na poziomie ekspresji (np. niska lub wysoka ekspresja) oraz translacji (np. wysoka lub niska aktywność translacyjna)?*

W kolejnej części pracy, Doktorant prezentuje przykłady możliwości użycia narzędzia TimeSeqR z wykorzystaniem dwóch zestawów danych biologicznych i jednego zestawu danych symulowanych. Pierwszy zestaw danych obejmuje 32 próbki RNA-seq oraz RIBO-seq pochodzące z organizmu *Bacillus*

subtilis. Próbki te uzyskano z linii potrójnego mutanta (3KO), pozbawionego aktywności genów kodujących paralogi białek rybosomalnych oraz ze szczepu dzikiego w procesie sporulacji trwającym ok. 7 godzin. Dane dla *Bacillus subtilis* zostały wygenerowane w ramach działalności naukowej laboratorium dr hab. Agaty Starosty, we współpracy z Genomed S.A. oraz Laboratorium Specjalistycznym Genomiki CENT na Uniwersytecie Warszawskim. Drugi zestaw to dane RNA-seq i RIBO-seq uzyskane przez Zheng Hu et al. dla *Saccharomyces cerevisiae* w ramach badań nad związkiem pomiędzy efektywnością translacji, a długością życia drożdży. Zestaw ten obejmuje 18 próbek zawierających komórki drożdży wyizolowane z normalnych hodowli (komórki młode), po 15 godzinach (komórki średniego wieku) oraz po 30 godzinach (komórki stare) hodowli w obecności estradiolu. Dane symulowane uzyskano przy wykorzystaniu biblioteki R – ImpulseDE2, aby zademonstrować działanie modułu „analizy pojedynczego eksperymentu". Dla każdego zestawu jasno opisano sposób przeprowadzenia analizy i dobór parametrów. Przedstawione wyniki jasno pokazują użyteczność narzędzia TimeSeqR. Niemniej jednak sposób przedstawienia wyników w pewnym stopniu zaniedbuje biologiczny aspekt analizy. Każdy z analizowanych zestawów danych obejmował pomiary ekspresji w ściśle określonych punktach czasowych. W związku z tym, *czy nie byłoby interesujące porównanie zmian aktywności transkrypcyjnej i translacyjnej badanych genów na przestrzeni tych punktów czasowych?* Istotne byłoby także poznanie przyczyny prowadzącej do uzyskania relatywnie wysokiej liczby genów specyficznych dla konkretnych rodzajów analizy: RNA-seq, RIBO-seq oraz efektywności translacji dla danych zestawów I oraz II. *Czy Pańskim zdaniem przyczyna ta ma podłoże biologiczne, czy techniczne?* Ponadto, *czy geny specyficzne dla np. efektywności translacji (212 genów) u Bacillus subtilis oraz specyficzne dla analizy RIBO-seq (1,480 genów) u Saccharomyces cerevisiae mają jakieś konkretne właściwości lub funkcje biologiczne, które sprawiają, że geny te nie zostały wykryte przez inny rodzaj analizy?* Podobne pytanie chciałabym zadać w kontekście porównania wyników analizy ekspresji różnicowej pomiędzy DESeq2 a TimeSeqR. Rozbieżności pomiędzy otrzymanymi listami genów są dość duże (361 genów specyficznych dla DESeq2 oraz 411 dla TimeSeqR przy jedynie 141 genach wykrytych przez obie metody). *Z czego może wynikać ta różnica i w jaki sposób klasyczna analiza różnicowa może wzbogacić wyniki otrzymane za pomocą TimeSeqR?*

Rozprawa doktorska obejmuje 160 stron i została napisana w języku polskim. Składa się z kilku części, zaczynając od wstępu, w którym krótko opisano problem badawczy i cele pracy, a także teoretyczne podstawy z zakresu technologii sekwencjonowania (DNA, RNA oraz analizy translatomu), jak również główne pojęcia bioinformatyczne dotyczące analizy danych, w tym formaty plików oraz główne algorytmy. Kolejny rozdział przedstawia przygotowanie danych oraz wstępne etapy analizy. Trzecia część pracy zawiera opis implementacji biblioteki TimeSeqR w języku R. Rozdziały cztery, pięć i sześć zawierają odpowiednio opis przykładów użycia biblioteki TimeSeqR na rzeczywistych oraz symulowanych danych

biologicznych, dyskusję oraz podsumowanie wraz z wnioskami. Istotnym elementem pracy są także załączniki, które zawierają aneks wraz z kodem funkcji do przeprowadzania normalizacji danych oraz płytę CD zawierającą kod źródłowy biblioteki TimeSeqR. Część literaturowa, obejmująca 139 odnośników, stanowi kompleksowy przegląd literatury z zakresu wysokoprzepustowych analiz transkryptomicznych, ze szczególnym uwzględnieniem analizy aktywności translacyjnej badanych genów.

Mgr Przemysław Latoch jest współautorem sześciu publikacji, w tym dwóch jako pierwszy współautor z równorzędnym wkładem naukowym, o czym mowa w pracy. Duża liczba publikacji w czasopismach o zasięgu międzynarodowym potwierdza dodatkowo interesujący charakter tematyki badawczej oraz aktualność podjętej problematyki. Usprawnienie procesu analizy danych transkryptomicznych poprzez jego standaryzację oraz poprawę powtarzalności otrzymywanych wyników stanowi jedno z ważniejszych wyzwań w dziedzinie genomiki i transkryptomiki. Niniejsza praca stanowi dodatkowy wkład w prowadzone wysiłki.

Podsumowując, stwierdzam, że przedstawiona mi do oceny rozprawa doktorska Pana mgr. Przemysława Latocha spełnia wymogi Ustawy z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce (Dziennik Ustaw 2020 r. poz. 85 z późniejszymi zmianami) i wnioskuję o dopuszczenie mgr. Przemysława Latocha do dalszych etapów przewodu doktorskiego.

Poznań, 03.06.2024

Podpisane elektronicznie przez Barbara Uszczyńska-Ratajczak (Certyfikat kwalifikowany) w dniu 2024-06-03.

Barbara Uszczyńska-Ratajczak