INSTYTUT CHEMII BIOORGANICZNEJ
Polskiej Akademii Nauk

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 inż. Agaty Muszyńskiej, zatytułowanej: „*Advanced data exploration techniques for augmented transcriptional landscape and its better quantification*”

(Zaawansowane techniki eksploracji danych do analizy rozszerzonego krajobrazu transkrypcyjnego i jego lepszej kwantyfikacji)

Przedstawiona do recenzji praca doktorska Pani mgr inż. Agaty Muszyńskiej została wykonana na Wydziale Automatyki, Elektroniki i Informatyki Politechniki Śląskiej. Promotorem rozprawy jest dr hab. inż. Paweł Łabaj z Małopolskiego Centrum Biotechnologii Uniwersytetu Jagiellońskiego, a promotorem pomocniczym dr David Kreil z Uniwersytetu Zasobów Przyrodniczych i Nauk o Życiu w Wiedniu (Universität für Bodenkultur Wien).

Tematyka pracy skupia się na zaawansowanych analizach transkryptomicznych z wykorzystaniem mikromacierzy DNA oraz sekwencjonowania RNA (RNA-seq) drugiej generacji. Szczególną uwagę poświęcono czynnikom technicznym, które mogą wpływać na proces analizy danych i jego powtarzalność. Głównym celem rozprawy jest ocena przydatności i znaczenia dobrych praktyk analizy z wykorzystaniem rzeczywistych zestawów danych w kontekście skutecznej minimalizacji wpływu zmienności technicznej na wyniki biologiczne. Praktyki te zostały przetestowane na trzech zestawach danych: zestawie mikromacierzy DNA oraz dwóch zestawach RNA-seq: rzeczywistym i referencyjnym. Kluczowym rezultatem przeprowadzonych badań jest stworzenie bioinformatycznego narzędzia, które obejmuje starannie zaprojektowany, kompleksowy proces analizy jakościowej i ilościowej danych transkryptomicznych, uwzględniających także różne typy zmienności technicznej.

Pierwszym problemem z jakim przyszło się zmierzyć Doktorantce jest efekt serii, wprowadzający do układu eksperymentalnego niezamierzone różnice techniczne wynikające z przygotowywania i przetwarzania próbek w poszczególnych grupach lub partiach. Zagadnienie to dotyczyło tzw. „rzeczywistego zestawu” danych RNA-seq, który obejmował 88 próbek. Zestaw ten zawierał próbki pochodzące z trzech różnych typów mutantów myszy oraz próbki kontrolne uzyskane z organizmów typu dzikiego. Wszystkie próbki przygotowane były w trzech seriach. Efekt serii wprowadzał dodatkową

zmienność techniczną, która znacząco zakłócała zmienność biologiczną pomiędzy próbkami (Rysunek 4.2a). W rezultacie analiza ekspresji różnicowej genów dawała fałszywie negatywny wynik. Zastosowanie narzędzia `SVAseq` pozwoliło zminimalizować efekt serii i poprawić jakość analizy różnicowej genów. Jednakże przeprowadzone analizy potwierdzają wcześniejsze doniesienia, iż usuwanie efektów technicznych, w tym także efektu serii, na etapie analizy dla danych jest procesem suboptymalnym. Pomimo przeprowadzonej korekty, otrzymane wyniki ekspresji różnicowej charakteryzowały się małą powtarzalnością pomiędzy seriami. Ponadto wykryte różnice ekspresji pomiędzy genami były na dość niskim poziomie.

Problem technicznej zmienności pojawia się także w przypadku analizy ekspresji genów z użyciem mikromacierzy DNA. Zestaw danych obejmuje 14 mikromacierzy DNA (Illumina HumanHT-12 v4) dla analiz transkryptomicznych 7 próbek uzyskanych od pacjentów z chorobą Parkinsona oraz 7 zdrowych ochotników. W tej części pracy, Doktorantka bada wpływ normalizacji danych oraz zmienności na poziomie próbek na proces detekcji genów różnicowych. Oba przypadki jednoznacznie wskazują, że korekta technicznej zmienności nie jest procesem trywialnym, ani całkowicie neutralnym dla różnic na poziomie biologicznym. Dlatego minimalizacja wszelkiej zmienności technicznej powinna odbywać się już na etapie planowania eksperymentu i przetwarzania próbek w laboratorium. Doktorantka zasadnie zwraca uwagę na kontekst biologiczny, postulując, że korekta zmienności technicznej powinna uwzględniać istotę pytania biologicznego oraz przestrzegać struktury zestawu danych, wynikającej bezpośrednio z układu eksperymentalnego. Całkowicie zgadzam się z główną konkluzją, że nie wszystkie błędy wynikające z przetwarzania próbek mogą być całkowicie wyeliminowane na etapie analizy danych.

Doktorantka, wykorzystując wcześniej opisany rzeczywisty zestaw danych RNA-seq (88 próbek), wykazała, że wskazany efekt serii ma mniejszy wpływ na analizę alternatywnego splicingu. Pomimo występowania znacznej zmienności technicznej w układzie, nadal możliwa jest wiarygodna analiza złożoności transkryptomicznej badanych próbek. Stosując narzędzie `Spladder`, Doktorantka przeprowadziła analizę pięciu najpopularniejszych typów zdarzeń alternatywnego splicingu (Tabela 4.3). Pomimo niskiej powtarzalności na etapie analizy ekspresji genów, wiele z wykrytych zdarzeń alternatywnego splicingu było współdzielonych pomiędzy badanymi próbkami. Szczególnie interesujące zdają się być dwa najmniej popularne, a zarazem najbardziej specyficzne typy zdarzeń: wzajemnie wykluczające się eksony (ang. *mutually exclusive exons*) oraz proces pominięcia wielu eksonów (ang. *multiple exon skip*). *Czy te typy zdarzeń dotyczą jakiejś szczególnej grupy genów? Mam na myśli tutaj ich biotyp (geny kodujące białko lub geny niekodujące), specyfikę kodowanego białka lub ich same właściwości genomiczne takie jak długość, liczba transkryptów danego genu lub liczba jego eksonów.* Korzystając z okazji chciałam zadać pytanie dotyczące analizy ekspresji różnicowej. *Skoro analiza alternatywnego splicingu jest znacznie bardziej powtarzalna od globalnej analizy ekspresji, czy próbowała Pani zbadać*

ekspresję genów stosując wyłącznie tzw. podzielone odczyty (ang. split reads), które nie posiadają ciągłego dopasowania i mapują się do dwóch różnych eksonów danego genu? Na ile Pani zdaniem optymalny jest proces filtrowania odczytów w porównaniu do korekcji zmienności technicznej?

W drugiej części pracy, Doktorantka skupia się wyłącznie na analizie alternatywnego splicingu z użyciem referencyjnego zestawu danych, który opracowany został w ramach konsorcjum SEQC2. Referencyjny zestaw danych zawiera próbki będące mieszaniną 10 różnych linii nowotworowych (A) oraz zdrowych osobników (B) w różnych proporcjach. Ponadto część próbek w ramach tego zestawu poddano ukierunkowanej metodzie sekwencjonowania, skupiając ten proces głównie na wybranych genach z panelu komercyjnego (1064 geny) oraz panelu niestandardowego (2125 genów). Zastosowanie celowanego sekwencjonowania zwiększa pokrycie badanych genów odczytami, pozwalając tym samym wydajniej i dokładniej wykrywać alternatywne formy transkryptów dla badanych genów. Przeprowadzone badania pokazują, iż poziom pokrycia genu jest istotny z punktu widzenia analizy alternatywnego splicingu, gdyż blisko 90% wykrytych połączeń eksonów pochodzi z genów ujętych w testowanych panelach. Ponadto detekcja ok. 100 tysięcy nowych połączeń eksonów wskazuje, iż wciąż nie do końca rozumiemy poziom złożoności transkryptomów ssaków. W związku z tym mam trzy pytania do tej części pracy: *(1) W jaki sposób Pani zdaniem będziemy identyfikować zdarzenia alternatywnego splicingu, które są istotne biologicznie? Byłabym wdzięczna za uwzględnienie scenariusza kiedy to konkretne typy zdarzeń mogą być powiązane z unikalnym zestawem funkcji danego transkryptu. (2) W jaki sposób, Pani zdaniem, powinniśmy radzić sobie z problemem niekompletności katalogów genów na etapie przetwarzania danych uzyskiwanych z użyciem sekwencjonowania RNA? Poprosiłabym o komentarz w kontekście badania alternatywnego splicingu, ale także i analizy ekspresji genów. (3) Jakie, Pani zdaniem, będą konsekwencje wzrostu katalogów genów poprzez identyfikację nowych genów oraz transkryptów dla analizy ilościowej i jakościowej transkryptomów?*

Przeprowadzone badania doprowadziły do stworzenia nowego narzędzia bioinformatycznego, które umożliwia niestandardową, kompleksową analizę danych transkryptomicznych. Chociaż sama idea nie jest nowa i obecnie dostępnych jest szereg podobnych rozwiązań, to proponowane narzędzie wyróżnia się możliwością przetwarzania różnych typów zestawów danych, takich jak dane uzyskiwane z użyciem mikromacierzy DNA oraz sekwencjonowania RNA metodą krótkich i długich odczytów. Taka wszechstronność jest możliwa ze względu na liczne podobieństwa w procesie analizy danych otrzymywanych z zastosowaniem mikromacierzy DNA oraz danych RNA-seq. Jednakże kompleksowa integracja tych kroków wymaga specjalistycznej wiedzy, którą bez wątpienia posiada Doktorantka. Istotnym atutem proponowanego narzędzia jest również implementacja bazy danych `InterProScan`, co umożliwia bezpośrednią ocenę wpływu alternatywnego splicingu na wytwarzane białka. *W tym miejscu chciałabym zapytać o przyszłość tego narzędzia oraz w jakim kierunku planuje je Pani rozwijać?*

Zakładając, że jednym z głównych kierunków rozwoju będzie analiza splicingu, chciałabym dopytać o rozbudowę wizualnej strony narzędzia.

Rozprawa doktorska obejmuje z 147 stron i została napisana w języku angielskim. Składa się z kilku części, zaczynając od wstępu, w którym krótko opisano cel i założenia pracy. Kolejny rozdział przedstawia teoretyczne podstawy, głównie wprowadzenie do transkryptomiki oraz opisuje założenia i wyzwania związane z wysokoprzepustowymi metodami analizy RNA. Dodanie kilku rysunków w części teoretycznej pozwoliłoby nieco przełamać techniczny charakter tego rozdziału i ułatwiłoby odbiór treści. Trzecia część pracy zawiera opis użytych metod oraz zestawów danych. Rozdziały cztery, pięć i sześć zawierają odpowiednio opis wyników, podsumowanie i podziękowania. Istotnym elementem pracy są także załączniki, które zawierają nośnik USB, listę rycin, tabel oraz dane uzupełniające (Supplementary Material). Część literaturowa, obejmująca 107 odnośników, stanowi kompleksowy przegląd literatury z zakresu wysokoprzepustowych analiz transkryptomicznych, ze szczególnym uwzględnieniem metod badania procesu alternatywnego składania transkryptów. Narzędzie bioinformatyczne, będące wynikiem niniejszej pracy jest publicznie dostępne za pomocą platformy github.

Doktorantka jest współautorką trzech publikacji, o czym mowa w pracy. To dodatkowe potwierdzenie interesującego charakteru tematyki badawczej oraz pozytywnej oceny przez zewnętrznych recenzentów. Szkoda jednak, że praca nie zawiera dodatkowego opisu naukowej aktywności Doktorantki. Choć taki opis nie jest formalnie wymagany, z pewnością stanowiłby dodatkowy atut, biorąc pod uwagę zaangażowanie Doktorantki w realizację projektów o znaczeniu międzynarodowym, takich jak udział w konsorcjum SEQC2.

Problem powtarzalności wyników w analizie danych biologicznych jest obecnie jednym z głównych wyzwań. W przyszłości konieczne będzie doskonalenie metod oraz narzędzi, które ocenią i poprawią powtarzalność ścieżek przetwarzania złożonych zestawów danych. Istotne będzie także ustalenie bardziej rygorystycznych standardów oraz rozwój zaawansowanych technik automatyzacji i integracji procesów analizy danych. Poprawa powtarzalności jest kluczowym krokiem w kierunku otrzymywania bardziej niezawodnych i wiarygodnych wyników badawczych w dziedzinie genomiki i transkryptomiki.

Podsumowując stwierdzam, iż przedstawiona mi do oceny rozprawa doktorska Pani mgr inż. Agaty Muszyńskiej spełnia wymogi Ustawy z dn. 20 lipca 2019 r. – Prawo o szkolnictwie wyższym i nauce (Dziennik Ustaw 2020 r. poz. 85 z późniejszymi zmianami) i wnioskuję o dopuszczenie mgr inż. Agaty Muszyńskiej do dalszych etapów przewodu doktorskiego.

Poznań, 23.06.2023

B. Ratajczak

Barbara Uszczyńska-Ratajczak