mgr. RDITT-05.10.2023
M. Skory

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
**Wydział Biologii**

dr hab. Michał Szcześniak, prof. UAM
Uniwersytet im. A. Mickiewicza w Poznaniu
Wydział Biologii
miszcz@amu.edu.pl

Poznań, 22 września 2023

Recenzja rozprawy doktorskiej mgr inż. Macieja Długosza pt.
**„Korekcja danych z sekwencjonowania genomów"**

## 1. Wprowadzenie

Rozprawa doktorska przedstawiona do recenzji została wykonana pod kierunkiem prof. dr hab. inż. Sebastiana Deorowicza na Wydziale Automatyki, Elektroniki i Informatyki Politechniki Śląskiej. Jest ona z zakresu technik informatycznych i bioinfromatyki i dotyczy opracowania, implementacji i testowania nowego algorytmu korekcji odczytów pochodzących z wysokoprzepustowego sekwencjonowania genomów w technologii Illuminy.

W pracy można wyróżnić następujące rozdziały:

1. *Wstęp*, w którym podsumowano treść pracy doktorskiej.
2. *Genomy i ich sekwencjonowanie*, gdzie nakreślono podstawy biologii molekularnej w kontekście sekwencjonowania materiału genetycznego, jak również zwięźle omówiono najbardziej popularne technologie sekwencjonowania.
3. *Kontekst informatyczny*, w którym omówiono problem korekcji odczytów od strony algorytmicznej.
4. *Algorytmy korekcji odczytów genomów*, w którym wykonano przegląd literatury pod kątem istniejących algorytmów korekcji odczytów z sekwencjonowania DNA, jak również algorytmów pomocniczych, na przykład służących do generowania symulowanych odczytów.

Kolejne dwa rozdziały to *Nowa metoda korekcji* oraz *Badania eksperymentalne*, które stanowią rdzeń rozprawy doktorskiej oraz podstawę do ubiegania się o stopień naukowy doktora. W rozprawie wyróżnić można ponadto *Spis treści*, *Podsumowanie*, *Bibliografię*, *Dodatek A: Wybrane pseudokody*, *Dodatek B: Szczegółowe informacje o eksperymentach*, *Dodatek C: Dodatkowe wyniki eksperymentów*, *Spis symboli i skrótów* (samych symboli naliczono 305), *Spis rysunków*, *Spis tabel* oraz *Spis pseudokodów*. Całość liczy 409 kolejno ponumerowanych stron. Poszczególne rozdziały są ze sobą spójne i tworzą logiczną całość. Błędy językowe czy

edytorskie pojawiają się niezwykle rzadko. Podsumowując, od strony formalnej praca jest napisana w sposób niemal perfekcyjny.

Poniżej przedstawiono krótką charakterystykę wyników i wniosków autora z rozdziałów *Nowa metoda korekcji* i *Badania eksperymentalne*.

## 2. Omówienie osiągnięć mgr inż. Macieja Długosza przedstawionych w rozprawie doktorskiej

Postawiono nastepujące tezy:

1. Zastosowanie niektórych nowoczesnych metod korekcji odczytów sekwencjonowania genomów pozwala na redukcję liczby błędów powstałych w trakcie sekwencjonowania oraz poprawę jakości wyników algorytmów przetwarzania odczytów, wliczając w to detekcję wariantów genomowych przy niskiej głębokości sekwencjonowania.
2. Możliwe jest opracowanie nowego algorytmu korekcji, zapewniającego lepszy od istniejących algorytmów bilans skuteczności względem zapotrzebowania na zasoby, a przy tym niskie ryzyko ewentualnej degradacji jakości danych.

W związku z drugą tezą, za cel postawiono opracowanie algorytmu, który będzie jak najbardziej skutecznie dokonywał korekty jakości danych, przy jednoczesnym zachowaniu szybkości przetwarzania i umiarkowanego zapotrzebowania na pamięć. Ważne było również, aby algorytm był mało wrażliwy na niedoskonałości modelowania danych sekwencjonowania i miał dobrą pesymistyczną czasową złożoność obliczeniową. Dążono też do tego, by ograniczyć do minimum ewentualne trudności techniczne związane z korzystaniem z algorytmu.

Opracowany algorytm, oparty na spektrum k-merów, określono mianem RECKONER.

Okazało się, że umożliwia on osiągnięcie wyników o dobrej jakości, a w niektórych przypadkach nawet najlepszych spośród szeregu analizowanych algorytmów. Jego implementacja cechuje się również relatywnie szybkim działaniem i zdolnością do skalowania dzięki możliwości przetwarzania równoległego, przy zachowaniu rozmiaru pamięci proporcjonalnego do sekwencjonowanego genomu. W trakcie przygotowania i implementacji algorytmu szczególną uwagę zwrócono na jego użytkowanie, w tym częściową automatyzację doboru podstawowych parametrów oraz eliminację problemów obserwowanych w innych algorytmach, a wyszczególnionych przez autora pracy.

W ramach tego algorytmu wprowadzono innowacyjne rozwiązania, m.in.:

1. Sposób pamięciowej reprezentacji rozwiązań częściowych.
2. Korekcję błędów typu indel przy pomocy metody z powrotami.
3. Uwzględniono liczebność k-merów w odczytach wejściowych.

4. Dopuszczono obecność w wynikowych odczytach sekwencji k-merów, które nie należą do danych wejściowych.
5. Wprowadzono staranną korekcję przy natrafieniu na symbole o niskiej jakości.

Aby osiągnąć wysoką szybkość algorytmu, zastosowano różne strategie, takie jak:

1. Ograniczenie liczby rozważanych przypadków lub uzyskiwanych wyników częściowych.
2. Wykorzystanie zrównoleglenia.
3. Delegowanie przetwarzanie zbiorów k-merów do specjalistycznych narzędzi, takich jak KMC i KMC tools.

Dodatkowo, uwzględniono możliwość kompresji wyników, jak również wykorzystano rozwiązania znane z istniejących wcześniej algorytmów, aby poprawić wyniki poprzez synergiczne ich połączenie, w szczególności wykorzystano oligomery o dwóch długościach w procesie korekcji i weryfikacji wyników oraz dostosowano metodę określania progu obcięcia spektrum k-merów.

Trzeba jednak zaznaczyć, że istnieje ryzyko zwiększonego zapotrzebowania na pamięć przy korekcji odczytów organizmów o bardzo dużych genomach, choć inne narzędzia oczywiście też nie są bez wad. Na przykład algorytm BFC cechuje się umiarkowaną skutecznością korekcji odczytów z sekwencjonatora NovaSeq, zaś algorytm Karect wymaga stosunkowo dużego zapotrzebowania na pamięć.

RECKONER, razem z dziewięcioma innymi narzędziami (znanymi wcześniej), został przetestowany eksperymentalnie, aby ocenić zużycie zasobów oraz uzyskiwane wyniki dotyczące samej korekcji odczytów, jak również jej wpływ na wyniki uzyskiwane w specjalistycznych aplikacjach bioinformatycznych, jak składanie genomów *de novo* czy wykrywanie wariantów genomowych. Testy przeprowadzono zarówno na rzeczywistych odczytach dostępnych w publicznych bazach danych, jak i na odczytach generowanych *in silico*. Na podstawie uzyskanych wyników stwierdzono, że korekcja zazwyczaj poprawia wyniki osiągane w eksperymentach bioinformatycznych, zwłaszcza w przypadku asemblacji *de novo* i przy niewielkiej głębokości sekwencjonowania - w tym świetle można uznać pierwszą tezę przedstawioną przez autora za potwierdzoną.

Wyniki uzyskane dla algorytmu RECKONER wskazują na jego dość dobrą skuteczność we wszystkich rozpatrywanych przypadkach (pogorszenie jakości danych zdarzało się sporadycznie). Razem z wyżej wymienionymi cechami algorytmu RECKONER można stwierdzić, że oferuje rozsądny kompromis między jakością wyników a zużyciem zasobów obliczeniowych. Wszystko to stanowi potwierdzenie drugiej tezy przedstawionej w pracy.

Warto nadmienić, że przedstawiony nowy algorytm oraz wyniki eksperymentów posłużyły do przygotowania artykułu opublikowanego w czasopiśmie *Bioinformatics*, a drugi artykuł jest w trakcie recenzji. Prace nad tymi tematami zaowocowały także rozdziałami w monografiach oraz plakatami prezentowanymi na konferencjach naukowych.

Rozprawa doktorska napisana jest w przejrzysty sposób, a od strony merytorycznej nie mam większych uwag. Chciałbym jednak, żeby autor rozprawy odpowiedział na następujące trzy pytania podczas publicznej obrony:

- Algorytmy i narzędzia, takie jak Reckoner, Lighter, Blue, nie przyjęły się jako komponenty standardowych potoków analitycznych służących do obróbki danych pochodzących z sekwencjonowania wysokoprzepustowego (NGS). Raczej wykonuje się filtrowanie względem współczynników jakości przypisanych poszczególnym nukleodydom (ang. *quality filtering*). Z czego to wynika? Czy potencjalne korzyści są niewielkie w stosunku do potrzebnych zasobów obliczeniowych, a może należałoby bardziej „wypromować" korekcję odczytów jako ważny składnik analizy danych NGS?
- Czy jest technicznie możliwe opracowanie uniwersalnego narzędzia, które wykonywałoby korekcję odczytów z dowolnego sekwenatora i dowolnej techniki (np. RNA-Seq, WGS), np. zakładając uprzednie wytrenowanie algorytmu na wszystkich możliwych kombinacjach takich wariantów? Ponieważ wciąż pojawiają się nowe technologie sekwencjonowania i nowe sekwenatory, taka cecha narzędzia byłaby pożądana.
- Błędy w odczytach są niekiedy cechą pożądaną w aplikacjach bioinformatycznych, jak np. wykrywanie modyfikacji epitranskryptomicznych w oparciu o sygnatury błędów generowanych na etapie odwrotnej transkrypcji, a więc w trakcie przygotowywania próbek RNA do sekwencjonowania (Tan KT et al. Repurposing RNA sequencing for discovery of RNA modifications in clinical cohorts. Sci Adv. 2021 Aug 4;7(32):eabd2605. doi: 10.1126/sciadv.abd2605). Ogólnie, dane transkryptmiczne są moim zdaniem trudniejsze pod kątem korekcji, niż genomowe – proszę o komentarz.

## 3. Wnioski

Zgodnie z obowiązującymi przepisami, wliczając *art. 187 ust. 3 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce*, recenzja rozprawy doktorskiej powinna zawierać jednoznaczną odpowiedź na poniższe trzy pytania:

1. **Ocena wraz z uzasadnieniem, czy rozprawa doktorska prezentuje ogólną wiedzę teoretyczną osoby ubiegającej się o nadanie stopnia doktora w określonej dyscyplinie albo dyscyplinach**
   Przedstawiona rozprawa doktorska w sposób jednoznaczny pokazuje wystarczającą wiedzę pana mgr inż. Macieja Długosza w zakresie technik informatycznych, algorytmice, a także wielkoskalowej obróbce danych oraz powiązanych aspektów biologii molekularnej, co wnioskuję na podstawie przedłożonych do oceny wyników, jak również wprowadzenia teoretycznego, stanowiącego wstęp do rozprawy.

2. **Ocena wraz z uzasadnieniem, czy rozprawa doktorska wykazuje umiejętność samodzielnego prowadzenia pracy naukowej lub artystycznej przez osobę ubiegającą się o nadanie stopnia doktora**

Autor rozprawy doktorskiej odegrał pierwszoplanową rolę w planowaniu i implementacji badań, jak również interpretacji uzyskanych wyników. Jest to dodatkowo potwierdzone czterema publikacjami i monografiami wymienionymi w pracy, w których doktorant jest pierwszym autorem.

3. **Ocena wraz z uzasadnieniem, czy rozprawa doktorska stanowi oryginalne rozwiązanie problemu naukowego, oryginalne rozwiązanie w zakresie zastosowania wyników własnych badań naukowych w sferze gospodarczej lub społecznej albo oryginalne dokonanie artystyczne.**
Uzyskane wyniki i ich walory merytoryczne, skrótowo scharakteryzowane wyżej, pozwalają jednoznacznie stwierdzić, że rozprawa doktorska przedstawia oryginalne rozwiązanie problemu badawczego. W szczególności należy tutaj wskazać zaprojektowanie i implementację szeregu innowacyjnych rozwiązań algorytmicznych, wraz z ich przetestowaniem na rzeczywistych i symulowanych danych biologicznych.

## 4. Podsumowanie

Przedstawiona do recenzji rozprawa doktorska, w tym wyniki uzyskane przez doktoranta, mogą być przedmiotem zainteresowania społeczności naukowej i potencjalnie stanowią wartościowy wkład do ulepszenia metod obróbki danych pochodzących z wysokoprzepustowego sekwencjonowania genomów. Stwierdzam, że rozprawa doktorska spełnia warunki określone w Ustawie z dnia 20 lipca 2018 roku prawo o szkolnictwie wyższym i nauce (Dz.U. z 2018 r. poz. 1668 ze zm.) oraz Ustawie z dnia 3 lipca 2018 r. Przepisy wprowadzające ustawę – Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2018 r poz. 1669 ze zm.) i **wnioskuję o dopuszczenie pana Macieja Długosza do dalszych etapów postępowania o nadanie stopnia doktora, jak również wyróżnienie stosowną nagrodą.**

Michał Szczerniak