Prof. dr hab. inż. Marta Szachniuk
Instytut Informatyki
Politechnika Poznańska
mszachniuk@cs.put.poznan.pl

Poznań, 17.08.2023

Recenzja Rozprawy Doktorskiej

Tytuł: Computational Methods for the Analysis of Chromosomal Rearrangements (Metody obliczeniowe w analizie rearanżacji chromosomowych)

Autor: mgr Barbara Agnieszka Poszewiecka

Promotorzy: prof. dr hab. Anna Gambin, dr Krzysztof Gogolewski

Problematyka badawcza rozprawy

Rearanżacje chromosomowe odegrały kluczową rolę w kształtowaniu ewolucyjnej trajektorii życia na Ziemi. Mechanizmy te obejmują inwersje, translokacje, duplikacje i delecje w sekwencji DNA skutkujące reorganizacją materiału genetycznego. W trakcie ewolucji gatunków rearanżacje chromosomowe zachodzą naturalnie. Są świadectwem dynamicznej natury genomów i trwającego procesu ewolucji. Ich konsekwencją są zmiany struktury i funkcji genomów, które wpływają na regulację ekspresji genów, pojawianie się nowych cech i fenotypów, fizjologię, adaptację organizmu do środowiska, itd. Badanie rearanżacji chromosomowych z wielu powodów jest uważane za krytyczne przedsięwzięcie współczesnych nauk o świecie ożywionym. Po pierwsze, rzuca ono światło na mechanizmy napędzające zmiany ewolucyjne. Analizując częstotliwość i rozkład rearanżacji u różnych gatunków jesteśmy w stanie odkryć historię ewolucji i relacje między organizmami. Wiedza ta pomaga w rekonstrukcji genomów przodków i śledzeniu rozbieżności gatunków w czasie. Po drugie, badanie zmian w genomach zapewnia wgląd w choroby i zaburzenia genetyczne. Nieprawidłowe rearanżacje mogą przyczyniać się do rozwoju nowotworów, zaburzeń rozwojowych i reprodukcyjnych. Odkrywanie przyczyn i konsekwencji rearanżacji chromosomowych ma kluczowe znaczenie dla diagnozowania, leczenia i profilaktyki takich schorzeń. Wreszcie rearanżacje chromosomowe bada się w kontekście praktycznych zastosowań w rolnictwie i biotechnologii. Rozumiejąc, w jaki sposób mechanizmy te

przyczyniają się do różnorodności genetycznej możemy opracować strategie ulepszania roślin uprawnych i zarządzania zagrożonymi gatunkami. Wgląd w rearanżacje jest podstawą do rozwoju technik inżynierii genetycznej i technologii edycji genomu.

Badanie rearanżacji chromosomowych na dużą skalę i w niespotykanej wcześniej rozdzielczości stało się możliwe dzięki pojawieniu się technologii sekwencjonowania oraz wspomagających je dedykowanych metod obliczeniowych. Ich połączenie zrewolucjonizowało naszą zdolność do rozszyfrowywania złożonych zmian genomowych leżących u podstaw procesów ewolucyjnych i chorobowych. Współczesne sekwenatory generują ogromne wolumeny danych, których nie da się analizować bez specjalizowanych narzędzi bioinformatycznych. Powstało już więc i nadal powstaje wiele takich narzędzi, w tym programy dedykowane do problemów związanych z rearanżacjami chromosomowymi. Są wśród nich algorytmy obliczeniowe, które analizują dane z sekwenatorów, aby wskazać miejsca rearanżacji w genomie. Rozwijane są strategie obliczeniowe, takie jak asemblacja *de novo* i sekwencjonowanie długich odczytów, które z coraz lepszą dokładnością potrafią zrekonstruować genomy dotknięte zmianami. Narzędzia analityczne ujawniają relacje ewolucyjne poprzez śledzenie rearanżacji u różnych gatunków. Powstają algorytmy przewidujące w jaki sposób rearanżacje wpływają na ekspresję i regulację genów, kierując badaniami funkcjonalnymi. Opracowywane są również platformy obliczeniowe, które integrują dane genomiczne i kliniczne w celu znalezienia korelacji między rearanżacjami a zaburzeniami genetycznymi. W końcu powstają zasoby obliczeniowe (bazy danych), które katalogują rearanżacje wspomagając tym ich analizę oraz rozwój algorytmów bioinformatycznych. W nurcie takich właśnie badań prowadzonych na pograniczu genomiki i nauk obliczeniowych powstała praca doktorska mgr Barbary Poszewieckiej. Doktorantka zajęła się problemem wykrywania oraz interpretacji zmian w architekturze genomu, które pojawiają się w wyniku rearanżacji chromosomowych. W ramach pracy doktorskiej opracowała narzędzie *PhaseDancer* do składania sekwencji DNA metodą wstępującą oraz stworzyła dwa powiązane z nim narzędzia do wizualizacji składania (*PhaseDancerViewer*) i do symulowania sekwencji segmentalnych duplikacji (*PhaseDancerSimulator*). Zaproponowała udoskonalony algorytm do analiz statystycznych UBCS oraz bazującą na nim metodę pozwalającą na datowanie odległości ewolucyjnej między gatunkami. Stworzyła wydajny algorytm generujący prawdopodobne scenariusze złożonych rearanżacji chromosomowych, a także opracowała aplikację *TADeus* do oceny wpływu zmian w konformacji chromatyny.

Ocena strony merytorycznej

Zawartość recenzowanej rozprawy doktorskiej jest zgodna z jej tytułem a przedstawione tezy są kompletne. Praca zawiera sześć rozdziałów. Pierwszy z nich stanowi wstęp teoretyczny. Autorka omawia w nim tło rozwiązywanego problemu badawczego. Przedstawia podstawowe informacje o budowie genomu oraz ekspresji genów. Następnie charakteryzuje technologie sekwencjonowania DNA pozwalające poznawać architekturę genomu. Rozpoczyna od pionierskich metod autorstwa W. Gilberta i F. Sangera z lat 70 XX wieku (sekwencjonowanie pierwszej generacji). Przedstawia technologie drugiej generacji na podstawie sekwenatorów firmy Illumina oraz sekwencjonowanie trzeciej generacji znane również jako sekwencjonowanie oparte na długich odczytach. W ramach tej ostatniej grupy Autorka prezentuje technologie oferowane przez Pacific Biosciences (PacBio) oraz Oxford Nanopore Technologies (ONT). W dalszej części rozdziału przedstawione są technologie wspomagające analizę danych z sekwencjonowania, w tym mapowanie optyczne oraz wychwytywanie konformacji chromosomów (3C, 4C, 5C, Hi-C). W ostatnim podrozdziale Wstępu przedstawione jest zagadnienie badawcze będące przedmiotem rozprawy doktorskiej – wykrywanie oraz interpretacja zmian w architekturze genomu kształtowane przez rearanżacje chromosomowe – oraz zwięźle ukazane są wyniki badawcze uzyskane w ramach doktoratu. Rozdział 1 kończy lista publikacji Autorki podzielona na dwie części: artykuły bezpośrednio związane z doktoratem (4 pozycje) oraz artykuły pozostałe (2 pozycje). W tym rozdziale brakuje według mnie przedstawienia motywacji podjętych badań. Przydałby się tutaj choćby krótki wyróżniony akapit nakreślający co możemy zyskać badając rearanżacje chromosomowe, dlaczego ich badanie jest ważne i dlaczego istnieje potrzeba aby angażować w nie nauki obliczeniowe.

Rozdział 2 ukazuje pierwsze z osiągnięć Doktorantki, jakim jest *PhaseDancer* - narzędzie do składania (asemblacji) regionów DNA zawierających duplikacje segmentowe. Autorka omawia krótko problem duplikacji, a następnie przechodzi do prezentacji algorytmu *PhaseDancer*. W kolejnych sekcjach charakteryzuje cztery fazy działania algorytmu – mapowanie, klastrowanie, składanie i rozbudowanie sekwencji – oraz podaje informacje dotyczące implementacji i dostępności metody. Następnie przedstawia dwa narzędzia wspomagające: *PhaseDancerViewer*, które pozwala wizualizować częściowe wyniki uzyskiwane podczas procesu asemblacji oraz *PhaseDancerSimulator*, które generuje kontigi na podstawie predefiniowanych scenariuszy historii ewolucyjnej i zestawu parametrów

opisujących sztuczne odczyty. Kolejny podrozdział poświęcony jest testom wydajnościowym, których wyniki ilustruje rysunek. W rozdziale 2.2 Autorka przedstawia materiały i metody, które zostały wykorzystane w analizie komparatywnej oraz eksperymentach obliczeniowych służących do ukazania przydatności narzędzia *PhaseDancer*. Samym eksperymentom i uzyskanym w nich wynikom poświęcono rozdział 2.3. Na szczególną uwagę zasługują tu wyniki porównania z innymi asemblerami, które pokazują imponującą przewagę metody *PhaseDancer* nad konkurencją w zakresie dokładności składania. Bardzo interesujące jest też zastosowanie *PhaseDancera* do analizy ewolucji regionów subtelomerowych małp człekokształtnych. Udzyskane tu wyniki doprowadziły do sformułowania cennej hipotezy dotyczącej wpływu fuzji przodków na ewolucję człowieka. Rozdział kończy się dyskusją oraz planami dalszych prac związanych z metodą *PhaseDancer*. W ogólności treść Rozdziału 2 jest bez zarzutu, mój jedyny komentarz jest następujący: w rozdziale pojawiły się pojęcia odległości Hamminga (str. 28) oraz odległości Levenshteina (str. 29). Uważam, że dobrze byłoby wyjaśnić czym są te miary odległości lub podać odnośniki do literatury, w której takie wyjaśnienie można znaleźć.

W Rozdziale 3 Autorka skupia się na datowaniu dużych zdarzeń ewolucyjnych. W szczególności odnosi się do metody określania momentu fuzji chromosomów bazującej na kwantyfikacji zdarzeń określanych jako tendencyjna konwersja genów (BCG, *biased gene conversion*) i proponuje udoskonalony algorytm szacowania zakresu czasowego głównych wydarzeń ewolucyjnych wykorzystujący statystyki sklastrowanych substytucji (UBCS, *unexpected bias clustered substitutions*). Rozdział rozpoczyna się przedstawieniem stanu wiedzy na temat datowania zdarzeń ewolucyjnych. Następnie opisane są zbiory danych genomowych wykorzystane w przeprowadzonych dalej badaniach, tj. sekwencje genomów małp człekokształtnych i ludzi współczesnych. W dalszej kolejności czytelnik może zapoznać się ze statystykami UBCS, ich wadami oraz wprowadzonymi przez Autorkę modyfikacjami i ich wpływie na oszacowanie czasu fuzji przodków. Treść uzupełniają liczne formuły matematyczne związane z analizą statystyczną oraz pseudokody trzech funkcji wykorzystanych w algorytmie do efektywnego wyznaczania oczekiwanej liczby zdarzeń BCS. Autorka przedstawia również obserwacje dotyczące zdarzeń ewolucyjnych związanych z mutacjami słabymi i silnymi. Na końcu rozdziału dyskutowane są możliwe ulepszenia, które można wprowadzić do przedstawionych scenariuszy analitycznych, zwłaszcza przy dostępności brakujących fragmentów chromosomów małp człekokształtnych. W kontekście

tego rozdziału zaintrygowały mnie różne definicje BCS. Być może przeoczyłam wyjaśnienie tego w rozprawie – jaka jest przyczyna tych rozbieżności i czy jest możliwe sformułowanie jednej kompromisowej definicji?

Rozdział 4 jest poświęcony poszukiwaniu wszystkich prawdopodobnych scenariuszy złożonych rearanżacji chromosomowych. Rozpoczyna się wyjaśnieniem mechanizmu powstawania złożonych rearanżacji, czyli zmian strukturalnych obejmujących więcej niż dwa punkty złamania w genomie. Autorka wskazuje na potrzebę listowania możliwych scenariuszy takich zmian w celu ich dalszych analiz wyjaśniających molekularne konsekwencji rearanżacji. Następnie proponuje własny algorytm listujący wszystkie możliwe scenariusze złożonych rearanżacji chromosomowych. Algorytm bazuje na grafie kariotypowym stanowiącym teoretyczny model złożonych rearanżacji. W kolejnych podrozdziałach Rozdziału 4 podane są więc podstawowe definicje dotyczące grafów w ogólności oraz grafu kariotypowego. Problem listowania scenariuszy złożonych rearanżacji jest przeformułowany do problemu wylistowania wszystkich minimalnych uporządkowanych Eulerowskich dekompozycji liniowodekomponowalnego grafu kariotypowego. Doktorantka przedstawia algorytm oraz analizuje jego złożoność obliczeniową. W części podsumowującej rozdział, pokazuje zastosowanie algorytmu do wyszukania prawdopodobnych scenariuszy złożonej rearanżacji chromosomowej dla przypadku klinicznego P5513_206 opisanego w pracy Nazaryana-Petersena i in. z 2018 roku. Według mnie w tym rozdziale warto byłoby umieścić pseudokod lub schemat algorytmu opracowanego przez Autorkę. W pewnym sensie lukę tę wypełnia drzewo rekursji, jednak w pracy informatycznej pseudokod lub schemat są równie pożądane.

Rozdział 5 przedstawia platformę obliczeniową *TADeus2*, która wspomaga diagnozy kliniczne na podstawie oceny patogenności zmian strukturalnych dezorganizujących strukturę trzeciorzędową chromatyny. Na początku rozdziału Autorka nawiązuje do wykorzystania sekwencjonowania nowej generacji w genetyce klinicznej oraz omawia przyczyny patologii – warianty strukturalne (SV) oraz zmienność liczby kopii (CNV). Następnie charakteryzuje istniejące narzędzia do klinicznej oceny wariantów strukturalnych. W dalszej części prezentuje metody zaimplementowane w aplikacji *TADeus2* oraz samą aplikację z jej wszystkimi funkcjonalnościami. W szczególności kładzie nacisk na ocenę przewidywalnej patogeniczności genu (wprowadzona własna miara do oceny i tworzenia rankingów) oraz ewaluację SV i CNV ze względu na ich patogenność. Podkreśla również – co jest warte szczególnej uwagi – że *TADeus2* jest pierwszym dostępnym narzędziem pozwalającym na

wizualizację wariantów strukturalnych z perspektywy punktów załamań rearanżacji. W kolejnym podrozdziale przedstawione są schematy zastosowane do przetestowania i oszacować poprawność działania aplikacji TADeus2. Przeanalizowane są również cztery przypadki pacjentów cierpiących na rożne schorzenia genetyczne, dla których z sukcesem wykorzystano system *TADeus2* przy ocenie patogeniczności wariantów strukturalnych. Rozdział kończy opis technicznej strony projektu oraz krótkie podsumowanie wraz z planami rozbudowy funkcjonalności systemu.

W Rozdziale 6 Doktorantka zawarła konkluzje oraz przedstawiła propozycje dalszych badań w tematyce związanej z pracą doktorską, w szczególności w zakresie algorytmiki oraz zastosowań klinicznych. W ramach kontynuacji proponuje m.in. wykorzystanie zaawansowanych modeli uczenia maszynowego, badania na bardzo dużych zbiorach danych genomowych oraz rozszerzenie analizy na inne gatunki. Ciekawą propozycją jest wykorzystanie opracowanych narzędzi do predykcji ryzyka pojawienia się chorób na podstawie wykrytych rearanżacji chromosomowych oraz włączenie narzędzi interpretacyjnych do badań klinicznych.

Ocena strony redakcyjnej

Rozprawa doktorska jest napisana w języku angielskim. Ma 136 stron maszynopisu i składa się z sześciu rozdziałów. Zawiera również krótkie streszczenie w języku polskim i angielskim, listę słów kluczowych, spis treści, wykaz rysunków i tabel oraz bibliografię. Tekst jest opatrzony 32 ilustracjami, z czego 30 jest kolorowych a 2 czarno-białe. W rozprawie znajdziemy również 5 tabel oraz pseudokody 3 algorytmów. Dokumentacja przekazana do recenzji zawiera dodatkowo rozszerzone streszczenia w języku polskim i angielskim, które nie stanowią integralnej części pracy.

Praca doktorska jest napisana poprawnym i zrozumiałym językiem. Autorka ma dar klarownego wyjaśniania nietrywialnych zagadnień naukowych oraz umiejętność przekazywania wiedzy w sposób treściwy, bez obarczania czytelnika szczegółami niezwiązanymi bezpośrednio z rozpatrywanymi problemami badawczymi. Struktura pracy jest prawidłowa, nie odbiega od powszechnie przyjętego schematu współczesnych rozpraw doktorskich. Podział treści jest odpowiedni, korzystny dla odbioru treści przez czytelnika. Bibliografia liczy 182 pozycje literaturowe wylistowane w układzie alfabetycznym według nazwiska pierwszego autora. Doktorantka oparła się na piśmiennictwie anglojęzycznym z lat

1962-2022 ukazującym się przede wszystkim w czasopismach oraz materiałach konferencyjnych z obszaru nauk o życiu (zwłaszcza biologii) oraz bioinformatyki. W większości są to oryginalne artykuły naukowe z listy Journal Citation Reports (JCR). Dobór bibliografii jest właściwy i wskazuje na dobre rozeznanie Autorki w tematyce, w której realizowane były badania. W opracowaniu zastosowano harwardzki system cytowań.

Rozprawa została przygotowana bardzo starannie i estetycznie, czyta się ją z dużą przyjemnością. Doceniam cytaty umieszczone na początku rozdziałów, dobrane w sposób nieoczywisty i świadczące o poczuciu humoru Autorki oraz zdrowym dystansie do pracy badacza. Błędy stylistyczne i interpunkcyjne należą do rzadkości (przykłady to nadmiarowe spacje – np. przed przed przecinkiem). Inne nieliczne usterki o charakterze redakcyjnym nie wpływają na zrozumiałość przekazu i nie obniżają mojej wysokiej oceny recenzowanej pracy doktorskiej, jako wartościowego opracowania naukowego. Poniżej wymieniam ważniejsze uchybienia redakcyjne:

- W pracy znajduje się kilkanaście formuł matematycznych, z czego tylko pięć zostało ponumerowanych. Rozumiem, że numery przydzielono tylko tym wzorom, do których Autorka odnosi się w wielu miejscach pracy. Wzory, do których nie ma odniesień nie posiadają numerów. Uważam, że praca zyskałaby gdyby wszystkie formuły potraktowano w ten sam sposób.
- W Bibliografii brakuje konsekwencji w pisowni nazw czasopism. Niektóre pisane są pełną nazwą, inne z wykorzystaniem skrótów – z kropkami na końcu lub bez nich. Wiele nazw własnych pisanych jest z małej litery. Nazwa tego samego czasopisma bywa pisana na różne sposoby (np. Genome research, Genome Research, Genome Res., Genome Res). W tytułach publikacji DNA pisane jest małymi literami (dna).
- Kolejność rozmieszczenia rysunków w Rozdziale 2 jest niezrozumiała i wydaje się przypadkowa. Rysunki powinny być umieszczone zgodnie z kolejnością w jakiej Autorka odnosi się do nich w treści rozdziału i najlepiej gdyby były wstawione blisko miejsca odniesienia – ułatwia to ich znalezienie. Odniesienia do rysunków występują w kolejności: Fig. 2.2, Fig. 2.1, Fig. 2.5, Fig. 2.8, Fig. 2.3, Fig. 2.6, Fig. 2.4, Fig. 2.7, Fig. 2.9, Fig. 2.10.
- W treści Rozdziału 3 brakuje odnośników do rysunków 3.2, 3.3 oraz 3.7. Ponadto pseudokod Algorytmu 3 powinien się znaleźć wcześniej w rozdziale (np. na str. 66).

Wnioski końcowe

Mgr Barbara Poszewiecka potrafi w sposób zrozumiały i przekonywujący prezentować wyniki badań naukowych, poprawnie formułuje hipotezy badawcze oraz wykazuje się trafnością wnioskowania. Posiada ogólną wiedzę teoretyczną w obszarach Informatyki, Bioinformatyki, Biomatematyki oraz Biologii i Genomiki Obliczeniowej. Wykazuje się głęboką znajomością aktualnego stanu wiedzy przedstawianego w literaturze światowej w zakresie zmienności genomowej i jej implikacji dla świata żywego, jak również pozyskiwania oraz badania danych genomowych metodami eksperymentalnymi i obliczeniowymi. Posiada umiejętność samodzielnego prowadzenia pracy naukowej oraz stosowania pozyskanej wiedzy w praktyce. Z sukcesem angażuje się także w prace interdyscyplinarnych zespołów badawczych.

Recenzowana praca doktorska stanowi oryginalne rozwiązanie problemu naukowego. Uzyskane przez Autorkę wyniki badań zostały opublikowane w czterech pierwszoautorskich artykułach, z czego trzy w wiodących, wysoko punktowanych czasopismach z dziedziny (ich sumaryczny współczynnik wpływu IF = 36,018; sumaryczna punktacja ministerialna = 540; wszystkie czasopisma znajdują się w I kwartylu wg Web of Science), a jedna – chronologicznie pierwsza – w materiałach międzynarodowej konferencji naukowej organizowanej pod patronatem IEEE. Doktorantka jest też współautorką dwóch innych publikacji, których tematyka dotyczy translokacji w genomie. Artykuły te zostały opublikowane w czasopismach z nauk medycznych i nie wchodzą w skład rozprawy. W dorobku mgr Poszewieckiej jest zatem łącznie 6 publikacji, które doczekały się 25/16 cytowań (Google Scholar/Web of Science).

Uważam, iż praca doktorska mgr Barbary Poszewieckiej pt. „Computational Methods for the Analysis of Chromosomal Rearrangements" spełnia wszelkie wymagania stawiane rozprawom doktorskim określone w art. 13 ust. 1 Ustawy o stopniach naukowych i tytule naukowym z dnia 14 marca 2003 roku oraz stanowi oryginalne rozwiązanie przez Autorkę zagadnienia naukowego. Autorka rozprawy uzyskała istotne wyniki badawcze, o czym świadczą publikacje tychże wyników w najlepszych czasopismach z dziedziny oraz praktyczne wykorzystanie stworzonych przez nią narzędzi do analiz genomowych. Wnoszę o dopuszczenie mgr Barbary Poszewieckiej do kolejnego etapu postępowania kwalifikacyjnego w celu uzyskania stopnia naukowego doktora w dyscyplinie Informatyka.

Marta Szachniuk

prof. dr hab. inż. Marta Szachniuk