Warszawa, 18 sierpnia 2023 r.

dr hab. inż. Robert Nowak, prof. uczelni
Instytut Informatyki
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska
ul. Nowowiejska 15/19
00-665 Warszawa

Recenzja rozprawy doktorskiej mgr Barbary Poszewieckiej pt „Computational Methods for the Analysis of Chromosomal Rearrangements (Metody obliczeniowe w analizie rearanżacji chromosomowych)"

Recenzja powstała na prośbę Rady Naukowej Dyscyplin Matematyka i Informatyka Uniwersytetu Warszawskiego, zgodnie z uchwałą Rady z dnia 29 czerwca 2023 roku, na podstawie: (1) rozprawy doktorskiej w języku angielskim, liczącej 136 stron, z czerwca 2023 r, (2) autoreferatu, (3) dorobku naukowego Kandydatki uwzględnionego w bazach Scoupus i Google Scholar, (4) kodów źródłowych prezentowanych rozwiązań dostępnych w repozytorium GitHub.

1 Tematyka badań

Rozprawa dotyczy metod analizy sekwencji DNA pozwalających na wykrywanie i interpretacje zmian genomu powodowane przez rearanżacje chromosomowe. Tytuł rozprawy odpowiada jej treści. Mgr Barbara Poszewiecka zaproponowała nowe metody analizy tego typu danych, przeprowadziła eksperymenty numeryczne na danych rzeczywistych, pochodzących z sekwenatorów nowej generacji. Wytworzone oprogramowanie zostało udostępnione w otwartych repozytoriach. Wyniki prac zostały opublikowane w międzynarodowych czasopismach naukowych z górnego decyla.

Cel badawczy jest postawiony właściwie, jest on interesujący i istotny. Przedstawione rozwiązania są poprawne, potwierdzone eksperymentami, rozwiązują one istotne zagadnienia biologiczne lub kliniczne.

2 Wyniki rozprawy

Jednym z mechanizmów ewolucji są rearanżacje chromosomowe, powstające na skutek pęknięć, a następnie łączenia się odcinków DNA. Obserwujemy wiele różnych rearanżacji chromosomowych, definiowanych jako zmiany na odcinku powyżej 10 000 par zasad, do najbardziej popularnych należą: duplikacje (powielenie fragmentu genomu), insercje, delecje, transpozycje. Zmiany takie nie są częste, ale zajmują ok. 10% genomu człowieka i są bardzo istotne ewolucyjnie, ponieważ dotyczą całych genów lub grup genów.

W rozprawie przedstawiono następujące, powiązane tematycznie, problemy badawcze:

1. asemblacja de-novo, z uwzględnieniem regionów zawierających duplikacje;

2. badanie sekwencji DNA pod kątem zmiany nukleotydów w procesie rekombinacji, na potrzeby datowania zdarzeń ewolucyjnych;

3. badania złożonych rearanżacji chromosomowych, uzyskiwanych przy więcej niż dwóch pęknięciach cząsteczki DNA,

4. wizualizacja wpływu rearanżacji chromosomalnych na choroby genetyczne, w tym tworzenie rankingu patologiczności genów.

Dla każdego z tych problemów w sposób właściwy przeprowadzono analizę źródeł z literatury światowej, co świadczy o dostatecznej wiedzy Autorki. Wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonujący. Doktorantka użyła właściwej metody do rozwiązania postawionych problemów.

2.1 Asemblacja *de-novo* z uwzględnieniem rejonów zawierających duplikacje

Asemblacja *de-novo* dostarcza sekwencję symboli reprezentującą cząsteczkę DNA na podstawie danych dostarczanych przez sekwenatory, nazywane odczytami, które są zbiorem sekwencji fragmentów cząsteczki. Asemblacja (znajdowanie nad-napisu) jest jednym z istotniejszych wyzwań w bioinformatyce, wynik asemblacji pozwala przeprowadzić dalsze analizy. Obecnie jest znanych kilkadziesiąt asemblerów, ale ze względu na istotność problemu, ciągle powstają nowe.

Mgr Poszewiecka zaproponowała nowy algorytm asemblacji, który opisano w rozdziale 2 rozprawy. Algorytm ten w kolejnych iteracjach tworzy sekwencje na podstawie grupowania podobnych odczytów, a następnie ją rozszerza. Pokazano, że takie podejście sprawdza się w sekwencjach zawierających powtarzające się, długie fragmenty, takie jak segmentalne duplikacje. Algorytm wykorzystano w nowym narzędziu PhaseDancer, zbadano na danych symulowanych, na danych ze sztucznych chromosomów bakteryjnych (BAC), oraz danych rzeczywistych z sekwencyjnych baz danych.

Elementy krytyczne i dyskusyjne

Porównano wyniki nowego asemblera z kilkoma innymi, pokazując ilość poprawnie odtworzonych klonów BAC. Moim zdaniem zabrakło tabeli, która porównuje wyniki w sposób standardowy dla asemblerów de-novo, pokazując statystyki N50, NA50, sumę długości kontigów $>$ 1000bp, itd. Badania takie wykonuje się na sekwencjach organizmów modelowych, dostępnych w bazach danych.

Pominięto przy porównaniu tzw. asemblery hybrydowe, wykorzystujące jednocześnie do asemblacji krótkie i długie odczyty. Asemblery hybrydowe powinny poprawnie odtwarzać sekwencje powtórzeń.

Podsumowanie

Powyższe uwagi nie podważają mojej wysokiej oceny przedstawionego rozwiązania.

2.2 Datowanie rearanżacji chromosomalnych

Drugim problemem, opisanym w rozdziale 3 rozprawy, jest badanie zjawiska zmiany pary nukleotydów A-T (słabo-stabilnej) na parę C-G (silnie-stabilną), podczas rekombinacji, w procesie naprawy DNA. Zjawisko to, nazywane Biased Gene Confersion, opisane w literaturze, pozwala na szacowanie daty zdarzeń ewolucyjnych. *Dreszer et all (2007)* opracował statystykę Unexpected Bias Clustered Substitution (UBCS), wykorzystaną m.in. do określenia czasu redukcji liczby chromosomów z 48 (u małp człekokształtnych) do 46 (u człowieka). Obliczanie UBCS polega na przetwarzaniu okien o szerokości 300 symboli, w oryginalnej pracy z przeskokiem 150 symboli, co oznacza, że każdy symbol jest analizowany dwukrotnie.

Mgr Poszewiecka zaproponowała ulepszony algorytm obliczania UBCS, gdzie przeskok jest parametrem algorytmu i może wynosić 1 symbol. Algorytm w kolejnych iteracjach analizuje okna, kompresując zmiany symboli do wektora liczb naturalnych i wykorzystując programowanie dynamiczne, co pozwala uzyskać zadowalającą złożoność obliczeniową. Na podstawie ulepszonych statystyk UBCS zaproponowano metodę datowania odległości ewolucyjnej pomiędzy gatunkami, wraz z szacowaniem ufności. Za pomocą tej metody zrekonstruowano odległości ewolucyjne pomiędzy gorylem, orangutanem, gibonem i człowiekiem. Wszystkie określone odległości pokrywają się z wynikami datowania uzyskanymi innymi metodami. Badania Doktorantki rzucają nowe światło na czas fuzji chromosomów, prowadzącej do powstania chromosomu 2 człowieka.

Dostarczono kody źródłowe rozwiązania na licencji MIT.

Elementy krytyczne i dyskusyjne

Miara UBCS jest zdefiniowana jako różnica pomiędzy występującą a spodziewaną liczbą zdarzeń podmiany nukleotydów. W rozprawie nie pokazano wartości spodziewanej liczby podmian. Nie jest dla mnie jasne, czy ta spodziewana wartość liczby podmian, określana na odcinku 1Mbp, była dla każdego odcinka inna, czy była taka sama dla całego genomu.

Jeżeli spodziewana wartość liczby podmian jest stała dla genomu, to chyba lepiej byłoby, przy definiowaniu ulepszonej UBCS, zrezygnować z różnicy i przedstawiać występującą liczbę podmian. Wtedy wartości byłyby liczbami naturalnymi, co upraszcza interpretacje, wykresy i algorytm.

Kolejna uwaga dotycząca ulepszenia: przeskok co 1 symbol oznacza wydłużony (patrząc na asymptotyczną złożoność) czas analizy. Z zamieszczonych wyników widać, że takie analizy udało się przeprowadzić. Badania badania dla przeskoku innego niż 1 (np 20, 150) są mniej użyteczne, dają wynik przybliżony. Ilość takich badań można byłoby zmniejszyć, nie zamieszczać wszystkich wykresów w pracy, wnioski są podejmowane na podstawie badań z przeskokiem 1.

Na rysunku 3.4 jest literówka, '(iii) 300 bp - one substitution can be contained in 300 windows', powinno być '(iii) 1 bp - one substitution can be contained in 300 windows'.

Kolejnym zagadnieniem, które moim zdaniem wymaga dyskusji jest ustalona szerokość okna analizy na 300, taka jak w oryginalnej pracy *Dreszer et all (2007)*. Czy to jest optymalna szerokość? Na stronie 68 pokazano, że

badano dla okna o długości 250 i 300. A dlaczego nie znacznie krótsze (np. 100, 50)? A dlaczego nie znacznie dłuższe (np. 1000)?

Z poprzedniej uwagi wynika kolejna: założono, że progowa wartość liczby podmian wynosi 5 lub 6. Skąd te liczby? Może warto byłoby wprowadzić dodatkowy parametr - szerokość okna do analizy i wtedy taki próg byłby obliczany, a nie dostarczany jako parametr?

Kod jest wysokiej jakości, niestety bez dokumentacji (komentarzy) i bez testów jednostkowych.

Podsumowanie

Powyższe uwagi nie umniejszają mojej wysokiej oceny rozwiązania, stanowią raczej wstęp do dyskusji.

3 Generowanie dekompozycji grafu kariotypowego

W ewolucji, oprócz prostych rearanżacji chromosomowych, które powstają, gdy DNA pęknie w dwóch miejscach, niekiedy występują rearanżacje złożone, uzyskiwanych przy więcej niż dwóch pęknięciach cząsteczki DNA. Dla rearanżacji złożonych, zmiany nie zawsze są jednoznacznie definiowane przez punkty złamań. Doktorantka dostarczyła algorytm, który wylicza wszystkie możliwe scenariusze złożonych rearanżacji dla danego zestawu punktów złamań, nazywane grafem kariotypowym. Takie rozwiązanie pozwala analizować konsekwencje molekularne rearanżacji. Rozdział 4 rozprawy prezentuje wyniki tych badań.

Problem po wykazaniu, że w analizie wpływu rearanżacji istotne są liniowe dekompozycje grafu kardiotypowego, został sprowadzony do równoważnego problemu dostarczana ścieżek Eulera. Analitycznie wyprowadzono złożoność czasową i pamięciową algorytmu. Algorytm zaimplementowano i badano jego własności na danych rzeczywistych, m.in. pokazano jego użyteczność w złożonej rearanżacji chromosomowej w genomie pacjenta.

Elementy krytyczne i dyskusyjne

Przede wszystkim brakuje odnośnika do strony projektu i kodu źródłowego oprogramowania. Liczba zastosowań dla rzeczywistych danych, gdzie opisano jeden przypadek, wydaje się niewielka.

4 Narzędzie TADeus2

W ramach rozprawy doktorskiej utworzono aplikację `TADeus2`, w architekturze aplikacji internetowej, która pozwala wizualizować wpływ rearanżacji chromosomowych (wariantów strukturalnych) na własności chromatyny. Wizualizacja dotyczy pełnego genomu (np. człowieka), także z perspektywy punktów złamań. Aplikacja zawiera algorytm szeregowania (tworzenia rankingu) istotności genów dla danego przypadku. Opis aplikacji `TADeus2` i udogodnień przez nią oferowanych jest w 5 rozdziale rozprawy. Rozwiązanie wykorzystano w praktyce klinicznej do analizy materiału genetycznego pochodzącego od pacjentów cierpiących na różne choroby genetyczne. Wyniki we wszystkich analizowanych przypadkach okazały się poprawne. Działający serwer jest dostępny na stronie UW, zaś kod źródłowy, na licencji GNU GPL, w repozytorium GitHub.

Elementy krytyczne i dyskusyjne

Utworzone narzędzie jest cennym zasobem dla badaczy, pozwalającym . Drobna uwaga dotyczy licencji, dlaczego użyto 'wirusowej' licencji GPL? Inne fragmenty oprogramowania, utworzone w pracy mają znacznie mniej restrykcyjną licencję MIT. Druga (drobna uwaga) dotyczy kodu źródłowego, jest tam trochę martwego kodu, brak komentarzy, brak testów.

5 Ocena dorobku naukowego

Dorobek mgr Barbary Poszewieckej zawiera 6 pozycji, cytowanych (w momencie pisania recenzji): 25 razy (Google Scholar), 16 razy (Scopus). Doktorantka publikuje w takich czasopismach jak (kolejność wynika z liczby cytowań): Journal of Medical Genetics (IF=5.9), Journal of Clinical Medicine (IF=3.9), Nucleic Acid Research (IF=19.2), BMC Genomics (IF=4.6), Journal of Thoracic Oncology (IF=20.1), czy materiałach IEEE Internatinal Conference on Bioinformatics and Biomedicine. Doktorantka jest pierwszym autorem trzech artykułów.

W mojej ocenie dorobek publikacyjny jest wyróżniający, z nadmiarem spełniający wymagania, aby dopuścić mgr Barbarę Poszewiecką do dalszych etapów przewodu doktorskiego.

6 Podsumowanie

Temat badawczy uważam za bardzo istotny, teza jest poprawna i oryginalna, wykazana w stopniu wyczerpującym. Opracowane rozwiązania są nowatorskie, potwierdzone eksperymentami i opublikowane w prestiżowych czasopismach naukowych.

Stwierdzam, że **recenzowana rozprawa doktorska mgr Barbary Poszewieckiej spełnia z nadmiarem warunki** określone w aktualnych przepisach i wnioskuję do Rady Naukowej Dyscyplin Matematyka i Informatyka Uniwersytetu Warszawskiego o dopuszczenie rozprawy doktorskiej do publicznej obrony.

Ponadto, ponieważ redakcja rozprawy jest zasadniczo bez zarzutu, osiągnięte przez Doktorantkę wyniki mają dużą wartość naukową, wnioskuję o **wyróżnienie rozprawy**.

KIEROWNIK
Zakładu Sztucznej Inteligencji
dr hab. inż. Robert Nowak
profesor uczelni