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

Poznań, 04.08.2023

Recenzja Rozprawy Doktorskiej

Tytuł: Computational Modelling and Analysis of the Three-Dimensional Structure of Human Genome at the Population Scale

Autor: mgr Michał Własnowolski

Promotor: prof. dr hab. Dariusz Plewczyński

Tematyka badawcza

Rozprawa doktorska Michała Własnowolskiego przedstawia badania przeprowadzone przez Autora w zakresie bioinformatyki oraz genomiki obliczeniowej. Skupia się na modelowaniu struktury przestrzennej genomu ludzkiego oraz analizie różnic, jakie występują w tej strukturze pomiędzy osobnikami w skali populacyjnej. Oba te problemy stanowią kluczowe zagadnienie współczesnej biologii molekularnej. Ich rozwiązanie jest możliwe dzięki zastosowaniu szeregu technik laboratoryjnych, modelowania problemów z wykorzystaniem narzędzi badań operacyjnych oraz zaawansowanych, wysokowydajnych metod obliczeniowych – m.in. symulacji komputerowych oraz analizy danych – operujących na bardzo dużych zbiorach danych.

W ramach badań przedstawionych w rozprawie, opracowano zestaw narzędzi obliczeniowych umożliwiających modelowanie i analizę zmian trójwymiarowej struktury chromatyny, która jest głównym składnikiem chromosomów. W badaniach wykorzystano dane o kontaktach chromatynowych uzyskane z eksperymentów ChIA-PET, informacje o wariantach strukturalnych takich jak duże delecje, insercje, inwersje i duplikacje, a także modelowanie trójwymiarowej struktury chromatyny za pomocą silnika symulacyjnego 3D-GNOME, opartego o algorytm symulowanego wyżarzania Monte Carlo. Platforma obliczeniowa 3D-GNOME jest jednym z najważniejszych wyników badawczych uzyskanych w pracy doktorskiej i stanowi istotny wkład Autora w rozwój genomiki obliczeniowej.

Układ rozprawy doktorskiej (w tym informacje o jej poszczególnych częściach składowych)

Rozprawa doktorska oparta jest na cyklu czterech artykułów naukowych, z czego trzy ukazały się już drukiem, a czwarty został zgłoszony do czasopisma i jest w fazie recenzji. Rozprawa rozpoczyna się od streszczenia, które wprowadza w tematykę pracy, przedstawia główne osiągnięcia Autora i wskazuje możliwe dalsze kierunki rozwoju badań. Następnie, we wstępie do rozprawy Autor prezentuje motywację stojącą za rozwojem narzędzi obliczeniowych dedykowanych do modelowania i analizy struktur 3D chromatyny na skalę populacyjną. Sekcja 1.1 zawiera opis działania algorytmu 3D-GNOME, wykorzystywanego do modelowania struktur 3D w toku badań opisanych w cyklu publikacji. W sekcji 1.2 Autor przedstawia biologiczny kontekst problemu badania regulacji ekspresji genetycznej wewnątrz jądra komórkowego oraz potrzebę syntetycznej analizy, która integruje informację o czynnikach genetycznych, epigenetycznych oraz elementach cis-regulatorowych, takich jak promotory i enhancery. Mapowanie tych czynników na modele 3D umożliwia bardziej informatywną wizualizację wybranego *locus* oraz analizę zmian dystansów pomiędzy czynnikami regulatorowymi. W kolejnych sekcjach – 1.3 oraz 1.4 – Doktorant przedstawia podstawowe cele badań opisanych w rozprawie oraz listę artykułów ujętych w cyklu.

Drugi rozdział rozprawy poświęcony jest głównym osiągnięciom opisanym w publikacjach z cyklu. W zwięzły sposób Autor przedstawia w nim swój wkład w prace podsumowane publikacjami [P1]-[P4]. Na szczególną uwagę zasługują wykonane z dużą dbałością ilustracje oraz pseudokody umieszczone w tym rozdziale. Rozdział jest podzielony na cztery podrozdziały, z których każdy dotyczy jednej publikacji z cyklu. Układ treści jest podobny w każdym podrozdziale i zawiera krótką motywację do podjęcia badań, opis wykorzystanych metod oraz uzyskanych wyników badawczych. Prace są przedstawione w kolejności chronologicznej, począwszy od najstarszej, która ukazała się w 2019 r.

W trzecim rozdziale zaprezentowane są dodatkowe osiągnięcia Autora. Obejmują one m.in. udział w grantach naukowo-badawczych, wizyty akademickie oraz współautorstwo publikacji nieuwzględnionych w cyklu. W rozdziale czwartym Autor przedstawił kolejne cele, m.in. badanie trójwymiarowej struktury chromatyny genomów archaicznych populacji ludzkich, takich jak Neandertalczycy i Denisowianie. Badania te są obecnie realizowane przez Autora we współpracy międzynarodowej z zespołem Dr. Guya Jacobsa z University of Cambridge.

Dodatkowo w pracy znajdziemy spis treści, spis ilustracji, bibliografię, kopie czterech publikacji stanowiących osiągnięcie naukowe opisane w jednotematycznym cyklu,

oświadczenia współautorów tych publikacji oraz cztery publikacje nieujęte w cyklu wraz ze wskazaniem wkładu Doktoranta.

Układ pracy mgr-a Michała Własnowolskiego jest prawidłowy, typowy dla powszechnie przyjętego schematu rozpraw doktorskich opartych na cyklu publikacji naukowych.

Zastosowane piśmiennictwo

Zastosowane piśmiennictwo jest ściśle związane z przedmiotem badań Autora. Bibliografia zawiera 104 pozycje literaturowe. Autor rozprawy stosuje tzw. vancouverski system cytowań. Odnosi się do artykułów publikowanych w najwyższej rangi czasopismach z dziedziny biologii, genomiki oraz bioinformatyki, m.in. *Nature*, *Nature Protocols*, *Nature Methods*, *Nature Communications*, *Nucleic Acids Research*, *Bioinformatics, Genome Biology, Proceedings of the National Academy of Sciences*. Znakomita większość tych publikacji to stosunkowo nowe (ukazały się w ciągu ostatnich 15 lat), wysoko cytowane prace. Przedstawiają one wyniki badań eksperymentalnych oraz obliczeniowych nad genomem, architekturą chromatyny lub opisują techniki eksperymentalne. Dobór bibliografii nie budzi zastrzeżeń. Wskazuje na bardzo dobre rozeznanie Doktoranta w tematyce podjętej w rozprawie doktorskiej oraz jego znajomość aktualnego stanu wiedzy w tym obszarze badawczym.

W spisie literatury znajdują się nieliczne usterki o charakterze redakcyjnym, na przykład nazwy czasopism przeważnie podawane są w formie skróconej lecz zdarzają się zapisy w formie pełnej mimo iż istnieje ogólnie przyjęty skrót (Proceedings of the National Academy of Sciences– poz. 2; Current Opinion in Genetics Development – poz. 4), nazwa tego samego czasopisma bywa pisana w różny sposób (Genome Biol. – poz. 48, 55; Genome biology – poz. 5, 16), zdarzają się nazwy własne pisane z małych liter (np. dna – poz 1, science – poz 8).

Cel pracy oraz zastosowane metody badawcze

Jak podaje Autor rozprawy, głównym celem pracy doktorskiej było opracowanie i wdrożenie bioinformatycznych narzędzi obliczeniowych do generowania i analizy trzeciorzędowych modeli chromatyny oraz zbadanie potencjalnego wpływu struktury przestrzennej chromatyny na aktywność genetyczną komórek. Przedmiotem prowadzonych badań był genom człowieka, jednak metody stworzone w ramach pracy można z powodzeniem zastosować do analizy innych genomów, w których tworzą się pętle chromatynowe.

Podczas prac badawczych Autor opierał się przede wszystkim na metodach szeroko stosowanych we współczesnej bioinformatyce łącząc przetwarzanie i modelowanie danych

biologicznych, analizy statystyczne, analizy dużych zbiorów danych, algorytmikę, algorytmy probabilistyczne, programowanie aplikacji internetowych, programowanie z wykorzystaniem kart graficznych, symulacje komputerowe.

Uważam, iż zastosowane metody badawcze są odpowiednie do rozwiązywanego problemu badawczego i wskazują na dobrą znajomość przez Autora rozprawy nowoczesnych i efektywnych metod oraz technologii stosowanych w naukach o życiu oraz naukach obliczeniowych. Założony przez Doktoranta cel pracy został osiągnięty.

Wyniki badań oraz ich praktyczne zastosowanie

Wyniki będące podstawą rozprawy doktorskiej zostały przedstawione w czterech publikacjach wieloautorskich [P1]-[P4]. Mgr Michał Własnowolski jest pierwszym autorem trzech spośród nich – [P2], [P3] i [P4]. Trzy publikacje – [P1], [P2], [P4] – ukazały się w wysoko punktowanych czasopismach naukowych z listy JCR w latach 2019-2023, czwarta została zgłoszona do czasopisma *Bioinformatics* i jest w trakcie recenzji.

Publikacja [P1], zamieszczona w czasopiśmie *Genome Biology* (IF_{2023} 18,01; 200 pkt MNiSW; kwartyl Q1), prezentuje narzędzie opracowane do przewidywania zmian kontaktów chromatynowych wprowadzanych na podstawie wariantów strukturalnych. Za pomocą tego narzędzia przeprowadzono kompleksową analizę trójwymiarowej struktury ludzkiego genomu na skalę populacyjną. [P1] jest najlepiej cytowaną pracą Doktoranta (22 cytowania według Web of Science).

Artykuł [P2], opublikowany w *Nucleic Acids Research* (IF_{2023} 19,16; 200 pkt MNiSW; kwartyl Q1), przedstawia implementację w serwisie internetowym 3D-GNOME (https://3dgnome.mini.pw.edu.pl/) metody opisanej w [P1] w ramach aktualizacji do wersji 2.0. Aktualizacja ta wprowadza narzędzia do porównywania zmian w kontaktach chromatynowych pomiędzy referencyjnym genomem GM12878 a alterowanym na podstawie wariantów strukturalnych. Umożliwia również porównywanie modeli trzeciorzędowej struktury referencyjnej i zmodyfikowanej, generowanych za pomocą silnika modelarskiego 3D-GNOME. Serwis został zintegrowany z zestawem danych wariantów strukturalnych 2504 genomów należących do 26 różnych populacji ludzkich z projektu 1000 Genome Project. Umożliwia on także wprowadzanie przez użytkowników własnych wariantów strukturalnych w formacie VCF. Publikacja [P2] ma 12 cytowań wg Web of Science.

Publikacja [P3] opisuje narzędzie cudaMMC, które jest rozszerzeniem silnika modelarskiego 3D-GNOME. Narzędzie to zwiększa wydajność obliczeń przez ich masowe zrównoleglenie na kartach GPU. Istotnie skraca to czas potrzebny do modelowania struktur 3D chromatyny – nawet do 25 razy dla największych chromosomów dla danych z *long-range* ChIA-PET CTCF. Największe przyspieszenie widać przy generowaniu całych kolekcji (*ensemble*) modeli 3D w oparciu o dane o znacznie większym rozmiarze, wygenerowane z eksperymentu *in situ* ChIA-PET, przy towarzyszącej temu większej stabilność czasu obliczeń. Wyniki te opisano w pracy, która została zgłoszona do czasopisma *Bioinformatics* (IF_{2023} 6,931; 200 pkt MNiSW; kwartyl Q1) i jest w trakcie recenzji.

Artykuł [P4] opublikowany w *Nucleic Acids Research* (IF_{2023} 19,16; 200 pkt MNiSW; kwartyl Q1), przedstawia narzędzie służące do analizy zmian rozkładu dystansów pomiędzy *loci* , które zawierają sekwencje promotorowe i enhancerowe. Analizator został dodany do serwisu internetowego 3D-GNOME w ramach aktualizacji do wersji 3.0, co wymagało wcześniejszej implementacji w tym serwisie narzędzia cudaMMC, opisanego w [P3]. Do obliczeń wykorzystywany jest klaster Eden, będący wewnętrznym heterogenicznym wysokowydajnym klastrem obliczeniowym HPC, wyposażonym w węzły Nvidia DGX A100. Prace te wymagały rozbudowy architektury serwisu internetowego i zarządzania obliczeniami przez oprogramowanie do kolejkowania zadań, *Slurm*. Dodatkowo baza danych z wariantami strukturalnymi została zaktualizowana do 3202 genomów z *1000 Genome Project*.

Na szczególną uwagę zasługuje fakt, iż wyniki badawcze uzyskane przez Doktoranta przyczyniły się do opracowania narzędzi analitycznych, które udostępniono poprzez serwis internetowy 3D-GNOME. Dzięki temu użytkownicy z całego świata mogą stosować je w praktyce w swoich badaniach naukowych. Serwis umożliwia badanie struktury 3D chromatyny przy wykorzystaniu zarówno z danych dostępnych na portalu (takich jak kontakty chromatynowe mediowane przez CTCF i RNAPII, warianty strukturalne z the 1000 Genome Project) jak i z własnych danych, które można wprowadzić do aplikacji. Dzięki przyjaznemu interfejsowi, platforma pozwala na przeprowadzanie skomplikowanych analiz nawet osobom nieposiadającym umiejętności programowania. 3D-GNOME umożliwia analizę wpływu zmian struktury 3D chromatyny na dystans między enhancerami a genami, a dzięki wykorzystaniu zrównoleglenia obliczeń na kartach GPU oraz infrastruktury klastra Eden, proces analizy jest wydajny. Autor rozprawy planuje rozbudowę bazy danych o kontakty chromatynowe dla kolejnych linii komórkowych (tj. H1ESC, HFFC6 i WTC11) oraz o dodatkowe warianty

strukturalne (the Simons Diversity Project), co pozwoli na badanie różnic między populacjami ludzkimi. Dodatkowo, planowane jest uwzględnienie danych dotyczących wymarłych ludzkich populacji, takich jak Neandertalczycy i Denisovianie, co poszerzy możliwości badawcze nad historią gatunku ludzkiego.

Uważam, iż wyniki badawcze uzyskane przez Doktoranta w rozprawie doktorskiej zasługują na wysoką ocenę. W szczególności doceniam fakt, iż mgr Własnowolski łączy umiejętności analityczne z algorytmicznymi, co pozwoliło na wdrożenie jego wyników badawczych w postaci ogólnodostępnych narzędzi obliczeniowych i umożliwienie efektywnego przetwarzania dużych wolumenów danych mających ogromne znaczenie we współczesnym świecie. Dużym atutem rozprawy są artykuły opublikowane w wysoko punktowanych czasopismach naukowych – publikacje [P1], [P2] i [P4] ukazały się w czasopismach z I kwartyla, ich sumaryczny współczynnik wpływu wynosi 56,33 a sumaryczna punktacja ministerialna to 600 pkt. Zgodnie z informacjami podawanymi przez Web of Science (na dzień 4.08.2023), mgr Michał Własnowolski posiada H-indeks = 4, a jego wszystkie publikacje były cytowane 48 razy (nie wliczając cytowań własnych).

Nieprawidłowości i braki w ocenianej rozprawie doktorskiej

Praca jest napisana ładnym i zrozumiałym językiem oraz przygotowana z dużą dbałością o szczegóły i zachowaniem wysokiej estetyki. Zauważyłam nieliczne błędy i nieścisłości, które nie wpływają na klarowność przekazu i nie zmieniają mojej wysokiej oceny rozprawy. Mają one często charakter błędów redakcyjnych, miejscami brakuje przedimków lub przecinków. Przykładowe usterki:

- str. 1: “To address the challenges” – lepiej brzmiałoby “To address these challenges”
- str. 5: “a singletons heatmap” – powinno być “a singleton heatmap”
- str. 5: “On each level of simulation Monte Carlo” – brak przecinka przed “Monte Carlo”
- str. 5: ”represented by the form” – powinno być “represented by the formula”
- str. 5: “simulation level the energy” – brak przecinka przed “the energy”
- str. 6: “Next term represent binding energy Eb, defined as:” – powinno być “Next term represents binding energy Eb and is defined as”
- str. 6: “In energy form, ws, wb, wo and wh are energy terms weights.” – to wyjaśnienie powinno znaleźć się bezpośrednio pod wzorem (1.5) lub należałoby napisać, że odnosi się ono to tego wzoru, np. „In formula (1.5), ws, wb, wo and wh denote weights of energy terms.”

- str. 7: “This necessitates specific regulation of gene expression, including at its initial stage – transcription” – nietypowy szyk wyrazów w zdaniu powoduje, że przekaz jest nieco niejasny
- str. 7: “These sequences operate by spatially interacting” – “These sequences operate by spatial interactions”

Uważam również, iż nie jest potrzebne zapowiadanie co Autor napisze w kolejnym rozdziale oraz umieszczanie w rozdziale wstępu, w którym Autor uprzedza, co znajdzie się w tym rozdziale. Takie powtórzenia według mnie osłabiają pracę i zmniejszają jej atrakcyjność.

W pracy (w rozdziale 3) zabrakło według mnie informacji o prezentacji wyników przez Doktoranta. Przypuszczam, że niejednokrotnie przedstawiał swoje wyniki badawcze na seminariach czy konferencjach naukowych. Udział w konferencjach jest istotną częścią pracy każdego naukowca, a wystąpienia konferencyjne to równie istotna forma przedstawiania swoich wyników badawczych jak ich publikowanie w artykułach naukowych.

Wnioski końcowe

Autor pracy wykazał się umiejętnością poprawnej i przekonywującej prezentacji wyników przeprowadzonych badań oraz trafnością wnioskowania. Dowiódł, iż w wysokim stopniu poznał dotychczasowy stan wiedzy o podejmowanym w pracy badawczej temacie, przedstawiany w przedmiotowej literaturze światowej. Posiada ogólną wiedzę teoretyczną w obszarze Informatyki, Bioinformatyki i Genomiki Obliczeniowej, wykazuje się umiejętnością samodzielnego prowadzenia pracy naukowej i zastosowania wiedzy w praktyce.

Recenzowana praca zawiera oryginalne rozwiązanie problemu naukowego. Uzyskane przez autora wyniki badań zostały opublikowane w wiodących, wysoko punktowanych czasopismach z dziedziny. Pracę oceniam bardzo wysoko. Ze względu na istotność uzyskanych wyników, szerokie podejście do rozwiązywanego problemu oraz bardzo dobre publikacje będące podstawą pracy, składam wniosek o jej wyróżnienie.

Stwierdzam, że praca mgr-a Michała Własnowolskiego pt. „Computational Modelling and Analysis of the Three-Dimensional Structure of Human Genome at the Population Scale” spełnia wymagania stawiane rozprawom doktorskim określone w art. 13.1 Ustawy o stopniach naukowych i tytule naukowym z dnia 14.03.2003 oraz stanowi oryginalne rozwiązanie przez autora zagadnienia naukowego.

……

prof. dr hab. inż. Marta Szachniuk