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

Poznań, 12.07.2024

Recenzja Rozprawy Doktorskiej

Tytuł: Spatial network model of sequence and structure diversity of Human genome at a population scale

Autor: mgr. inż. Mateusz Chiliński

Promotor: prof. dr hab. Dariusz Plewczyński

Tematyka badawcza

Rozprawa doktorska Mateusza Chilińskiego przedstawia interdyscyplinarne badania przeprowadzone przez Autora na styku biologii molekularnej i sztucznej inteligencji. Koncentruje się ona na wykorzystaniu narzędzi informatyki, w szczególności głębokiego uczenia w problematyce związanej z sekwencjonowaniem cząsteczek DNA. W ramach badań przedstawionych w rozprawie, przeanalizowano sekwencje DNA z punktu widzenia chorób genetycznych, opracowano algorytm do wykrywania wariantów genetycznych, stworzono modele do predykcji przestrzennej konformacji chromatyny, będącej głównym składnikiem chromosomów, a następnie wykazano zależności między strukturą 3D chromatyny w jądrze komórkowym a ekspresją genów ukazując możliwość przewidywania ekspresji genów w drodze symulacji konformacji chromatyny. Opracowane przez Autora modele predykcyjne zostały skonstruowane w oparciu o najbardziej obecnie popularne architektury sieci neuronowych. Ich działanie skonfrontowano z danymi pochodzącymi z eksperymentów służących do badania trójwymiarowej architektury genomów - ChIA-PET (Chromatin Interaction Analysis with Paired-End Tag) oraz Hi-C.

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

Rozprawa doktorska oparta jest na cyklu pięciu artykułów naukowych, z czego cztery ukazały się już drukiem. Piąty został zgłoszony do czasopisma w lutym tego roku, a jego przedruk jest dostępny w repozytorium BioRxiv. Rozprawa rozpoczyna się od streszczenia w języku

angielskim i jego odpowiednika po polsku. Streszczenie wprowadza w tematykę pracy oraz przedstawia główne osiągnięcia Autora. Dalej następuje wstęp do rozprawy, w którym Autor prezentuje motywację do podjęcia badań realizowanych w ramach doktoratu, nakreśla problematykę sekwencjonowania DNA (sekcja 1.1), opisuje modele uczenia maszynowego, z których korzystał w badaniach (konwolucyjne sieci neuronowe, transformery, architekturę BERT, modele dyfuzji oraz technikę wzmocnienia gradientowego) (sekcja 1.2), ukazuje ogólny cel pracy doktorskiej (sekcja 1.3) oraz przedstawia listę publikacji tworzących trzon rozprawy doktorskiej (sekcja 1.4). Lista publikacji zawiera dodatkowo informacje o wkładzie mgr. Mateusza Chilińskiego w każdą z prac, współczynnik wpływu (IF) czasopisma (czy są to wartości IF z roku 2022?), punktację ministerialną czasopisma oraz jego przypisanie do dyscypliny (wyjątek stanowią [P1] i [P4], gdzie nie podano takiego przypisania).

Drugi rozdział rozprawy poświęcony jest głównym osiągnięciom uzyskanym podczas pracy nad doktoratem. W zwięzły sposób Autor przedstawia w nim po kolei wyniki badawcze podsumowane w publikacjach [P1]-[P5]. Rozdział jest podzielony na pięć sekcji, z których każda dotyczy jednej publikacji z cyklu. Opisy są zwięzłe i opatrzone estetycznie przygotowanymi ilustracjami. Układ treści jest podobny w każdej sekcji i zawiera krótki wstęp, opis metod wykorzystanych w badaniach oraz uzyskanych wyników badawczych. Wyjątkiem jest pierwsza publikacja będąca pracą przeglądową. W sekcji opisującej tę pracę z oczywistych powodów nie umieszczono części prezentującej metody.

Trzeci rozdział zawiera podsumowanie i główne wnioski z przeprowadzonych badań, a także nakreśla kierunki przyszłych prac, które mogą być kontynuacją niniejszej pracy doktorskiej.

W czwartym rozdziale zaprezentowano dodatkowe osiągnięcia Autora. Obejmują one współautorstwo sześciu publikacji spoza cyklu, obowiązki dydaktyczne, udział w grantach naukowo-badawczych, staże i wizyty akademickie, uzyskane nagrody oraz prezentacje wyników badawczych na krajowych i zagranicznych konferencjach naukowych. Przyznaję, że jestem pod dużym wrażeniem wykazanej tu aktywności Autora dysertacji.

Kolejne rozdziały zawierają bibliografię (Rozdział 5), przedruki artykułów tworzących cykl rozprawy doktorskiej (Rozdział 6), deklaracje dotyczące wkładu mgr. Mateusza Chilińskiego w publikacje z cyklu podpisane przez wszystkich współautorów (Rozdział 7) oraz przedruki innych publikacji, których współautorem jest Doktorant (Rozdział 8).

Układ pracy mgr. inż. Mateusza Chilińskiego jest prawidłowy, typowy dla powszechnie przyjętego schematu rozpraw doktorskich opartych na cyklu publikacji naukowych. Dodam

jednak, że w dysertacji przydałaby się dodatkowa sekcja z listą algorytmów/programów opracowanych przez Autora w ramach pracy doktorskiej. Nowe metody są wartościowym (a przede wszystkim głównym) elementem doktoratu i dobrze byłoby to uwypuklić podając w jednym miejscu ich nazwy wraz z linkami do repozytorium, w którym je umieszczono.

Zastosowane piśmiennictwo

Zastosowane piśmiennictwo jest ściśle związane z tematyką badań. Bibliografia zawiera 110 pozycji literatury. 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, Science, Cell, Nature Methods, Nature Communications, Nucleic Acids Research, Bioinformatics, eLife, Genome Biology, Genome Research*. Znakomita większość tych publikacji to stosunkowo nowe (ukazały się w ciągu ostatnich 10 lat), dobrze cytowane prace. Przedstawiają one wyniki badań eksperymentalnych oraz obliczeniowych genomu, zastosowania różnych technik sekwencjonowania i analizy danych genomowych, badań chromatyny oraz wariantów genetycznych. 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 znalazłam nieliczne drobne usterki i niespójności, na przykład dla niektórych pozycji podana jest pełna lista autorów a dla innych tylko nazwisko pierwszego autora, przy czym nie widzę żadnej tu zależności od liczby współautorów artykułu, w pozycji 94 pojawił się artefakt między nazwiskami autorów a tytułem publikacji.

Cel pracy oraz zastosowane metody badawcze

Jak podaje Autor rozprawy, głównym celem pracy doktorskiej było opracowanie modeli uczenia maszynowego pozwalających na przejście z sekwencji DNA do danych eksperymentalnych wyższego poziomu. Cel jest sformułowany bardzo ogólnie i nieco enigmatycznie. O ile dobrze rozumiem, eksperymentem dającym dane wyższego poziomu Autor nazywa Hi-C, ChIA-PET, symulację komputerową i przewidywanie *in silico* struktury przestrzennej. W pracy założono, że badania będą wykonywane na sekwencjach DNA pochodzących z genomu człowieka, aczkolwiek opracowane modele mogą być również zastosowane do genomów innych organizmów. Rozdział zatytułowany „Cel rozprawy" zawiera streszczenia publikacji z cyklu. Wyjaśniają one w jaki sposób publikacje te realizują cel pracy doktorskiej. Metody badawcze zastosowane przez Autora są ujęte w samym

sformułowaniu celu. Są to przede wszystkim metody uczenia maszynowego takie jak konwolucyjne sieci neuronowe, transformery, architektura BERT, modele dyfuzji oraz technika wzmocnienia gradientowego. Wszystkie te modele zostały dość dobrze opisane w rozdziale wstępnym. W ogólności, podczas prac badawczych Autor opierał się na metodyce powszechnie stosowanej we współczesnej bioinformatyce łączącej przetwarzanie i modelowanie danych biologicznych, analizy statystyczne, analizy dużych zbiorów danych, algorytmikę, symulacje komputerowe oraz przeróżne techniki programistyczne.

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

Przy tak szeroko i ogólnikowo sformułowanym celu trudno jest uznać, że został on osiągnięty. Istnieje bowiem cały szereg danych eksperymentalnych wyższego poziomu, które również można symulować wychodząc z sekwencji DNA, a których niniejsza rozprawa doktorska nie dotyczy. Z tego względu uważam, że cel pracy mógłby być sformułowany bardziej szczegółowo, z ukierunkowaniem na konkretne dane. Nie mam jednak wątpliwości, że przedstawiona praca doktorska realizuje założony cel badawczy.

Wyniki badań oraz ich praktyczne zastosowanie

Wyniki będące podstawą rozprawy doktorskiej zostały przedstawione w pięciu publikacjach wieloautorskich [P1]-[P5]. Mgr Mateusz Chiliński jest pierwszym autorem wszystkich tych artykułów. Cztery publikacje – [P1], [P2], [P3] i [P5] – ukazały się w czasopismach naukowych z listy JCR w latach 2022-2023, piąta została wysłana do czasopisma w lutym tego roku.

Publikacja [P1], zamieszczona w czasopiśmie *Seminars in Cell & Developmental Biology* (IF_{2023} = 6,2; IF_{5Y} = 6,9; 140 pkt MNiSW; kwartyl Q1; Autor podaje IF = 7,3), to artykuł przeglądowy, który jest jednocześnie wprowadzeniem w biologiczną problematykę pracy doktorskiej. Ukazuje on związki między sekwencją DNA, genomiką przestrzenną oraz ekspresją genów. Przedstawiono w nim kwestię wariantów strukturalnych, metody ich wykrywania, ich wpływ na choroby oraz związek z trzeciorzędową strukturą genomu. [P1] jest najlepiej cytowaną pracą Doktoranta (11 cytowań według Web of Science).

Artykuł [P2], opublikowany w *Bioinformatics* (IF_{2023} = 4,4; IF_{5Y} = 7,6; 200 pkt MNiSW; kwartyl Q1; Autor podaje IF = 5,8), prezentuje oprogramowanie do odkrywania wariantów strukturalnych opracowane przez Doktoranta. Oprogramowanie składa się z dwóch części.

Pierwsza z nich, ConsensuSV-core, to metoda wykorzystująca DL do tworzenia kompromisowych wariantów strukturalnych z danych (wariantów) wygenerowanych przez kilka zewnętrznych algorytmów. Druga część to pakiet ConsensuSV-pipeline, który w kilku krokach przetwarza surowe dane z eksperymentów biologicznych, aby przygotować je do formatów przyjmowanych przez programy do wykrywania wariantów i – w ostatnim etapie – uruchomić ConsensuSV-core. Pipeline został udostępniony na platformie GitHub. Publikacja [P2] ma 3 cytowania wg Web of Science (w tym 2 cytowania obce). Chciałabym zapytać czy Autorowi wiadomo coś na temat wykorzystania oprogramowania ConsensuSV-pipeline w praktyce. Jedna z prac cytujących [P2] opisuje podobne oprogramowanie do generowania kompromisowych wariantów strukturalnych dla bakterii. Wynikałoby z tego, że ConsensuSV-core nie jest obsługuje sekwencji tych organizmów (lub autorzy drugiej pracy nie przetestowali ConsensuSV-core dla swoich danych). Czy ConsensuSV jest przeznaczony wyłącznie do analizy genomu człowieka?

Publikacja [P3] ukazała się w *Quantitative Biology* (IF_{2023} = 0,6; IF_{5Y} = 2,6; 200 pkt MNiSW; kwartyl Q4; Autor podaje IF = 3,1). Praca ma 2 cytowania odnotowane w Web of Science. Opisuje ona wykorzystanie hybrydowego głębokiego uczenia w przewidywaniu pętli chromatyny na podstawie sekwencji DNA. Autorzy wytrenowali kilka modeli ML (DNABERT, KNN, SVM, RF), a uzyskiwane z nich dane wyjściowe zintegrowali stosując mechanizm głosowania. Otrzymane predykcje mają dość dużą zgodność z danymi rzeczywistymi pochodzącymi z eksperymentu wskazując na duży potencjał hybrydowego podejścia do rozwiązania badanego zagadnienia.

Artykuł [P4] nie został jeszcze opublikowany. Przedstawia on model HiCDiffusion do predykcji macierzy Hi-C z sekwencji DNA. Co prawda istnieją już metody dedykowane do przewidywania macierzy Hi-C, jednak generują one niedokładne, rozmyte obrazy macierzy. Celem Autora było opracowanie modelu, który będzie radził sobie z problemem niskiej jakości obrazów wyjściowych. Aby ten cel osiągnąć opracował architekturę koder-dekoder wzmocnioną o transformator uwzględniający uczenie kontekstowe. Architektura ta została wykorzystana jako element modelu dyfuzji, który przechodzi od sekwencji DNA do macierzy kontaktów dając w wyniku macierze, które dla ludzkiego oka są niemal nieodróżnialne od rzeczywistych macierzy eksperymentalnych. Interesuje mnie kwestia walidacji modelu HiCDiffusion oraz porównania go z innymi narzędziami do predykcji macierzy kontaktów. W samym artykule wzmianka o porównaniu nie zawiera żadnych konkretów – np. napisano,

że zostały obliczone współczynniki korelacji Pearsona oraz FID score, ale nie podano ich wartości. Chętnie zapoznam się ze szczegółowymi wynikami analizy komparatywnej.

Publikacja [P5] ukazała się w czasopiśmie *Scientific Reports* IF_{2023} = 3,8; IF_{5Y} = 4,3; 140 pkt MNiSW; kwartyl Q1; Autor podaje IF = 4,996) i ma jedno cytowanie wg Web of Science. Przedstawiono w niej zmodyfikowaną wersję istniejącego algorytmu ExPecto do przewidywania wpływu wariantów na ekspresję i ryzyko chorób. Zmodyfikowany algorytm dodatkowo bierze pod uwagę wejściową sekwencje DNA. Autorzy porównali działanie oryginalnej i zmodyfikowanej wersji ExPecto pokazując poprawę w przewidywaniu ekspresji genów, jeśli uwzględnione w niej zostaną sekwencje położone blisko miejsca rozpoczęcia transkrypcji zarówno z punktu widzenia lokalizacji liniowej (blisko w obrębie sekwencji) jak i przestrzennej (blisko w strukturze 3D).

Uważam, iż wyniki badawcze uzyskane przez Doktoranta w rozprawie doktorskiej zasługują na pozytywną, wysoką ocenę. Dużym atutem rozprawy jest fakt, że mgr. Chiliński jest wiodącym autorem wszystkich publikacji w cyklu. Trzy artykuły z cyklu ukazały się w czasopismach z I kwartyla, a jeden w czasopiśmie z IV kwartyla. Sumaryczny współczynnik wpływu czterech opublikowanych prac wynosi 15 (wg IF_{2023})/ 21,4 (wg IF_{5Y})/ 21,196 (wg danych Autora). Sumaryczna punktacja ministerialna to 550 pkt. Dodatkowo Doktorant jest współautorem sześciu innych publikacji o łącznym współczynniku wpływu 39,5 (wg danych Autora) i sumarycznej punktacji ministerialnej równej 700. Zgodnie z informacjami podawanymi przez Web of Science (na dzień 12.07.2024), mgr Mateusz Chiliński posiada H-indeks = 3, a jego wszystkie publikacje były cytowane 27 razy, Google Scholar podaje H-indeks = 4 oraz liczbę cytowań równą 40. Biorąc pod uwagę fakt, że wszystkie prace, których współautorem jest Doktorant ukazały się w latach 2022-2024, uważam, że są to bardzo dobre wskaźniki i świadczą o tym, że publikacje adresują aktualne problemy badawcze, trafiają do właściwego grona odbiorców i mają wpływ na badania innych zespołów zajmujących się podobną tematyką.

Nieprawidłowości i braki w ocenianej rozprawie doktorskiej

Praca jest napisana raczej poprawnie, aczkolwiek w niektórych miejscach mało zrozumiale. Autor ma upodobanie do długich zdań z dygresjami, w których łatwo gubi się zarówno czytelnik jak i – prawdopodobnie – sam Doktorant. Angielski nie jest językiem fleksyjnym, w związku z czym unika się w nim długich zdań złożonych i wtrąceń, a jeśli są one konieczne,

należy umiejętnie stosować zasady interpunkcji. Tymczasem w pracy często zaznaczony jest początek wtrącenia, ale nie wiadomo, gdzie się ono kończy, gdyż brakuje odpowiedniego znaku. Autor w losowych miejscach wstawia myślniki – zarówno w części pisanej po angielsku jak i w polskim streszczeniu. Powoduje to chaos i negatywnie wpływa na odbiór i zrozumiałość przekazu. Zapoznając się z pracą czytałam niektóre zdania wielokrotnie próbując się zorientować czego kontynuacją jest fragment po trzecim czy czwartym myślniku/przecinku. Przykłady zbyt długich/skomplikowanych zdań, które z powodzeniem można zastąpić kilkoma krótszymi zdaniami:

- „To solve those limitations, the HiCDiffusion model was created – an algorithm that connects the modern approach of encoder-decoder architecture along with a transformer for context learning and enhances the results by applying conditional diffusion – to improve the quality of the final result so that it is indistinguishable from the original experimental Hi-C data."
- „The study was finished by showing relations between the spatial conformation of chromatin within the nucleus and gene expression – as models used for gene prediction very often include only local sequence (e.g. 20kbp around transcription starting site) – and are unaware of the sequences that are far away in the linear sense – though very close in 3D, thus interacting."

Inne błędy, nieścisłości czy powtórzenia nie wpływają na klarowność przekazu i jest ich stosunkowo mało, przykładowo:

- str. 11: „Rozprawa zaczyna się się od analizy sekwencji – i pokazania jej złożonego związku"
- str. 11: „Następnie zaprezentowano algorytm wykrywania wariantów – aby zapewnić"
- str. 11: „Algorytm (...) jest oparty na 8 najnowocześniejszych narzędzi"
- str. 11: "elementy regulatorowe są są położone"

Ponadto, jak wspomniałam wcześniej, w pracy przydałaby się lista programów stworzonych przez Autora – albo w oddzielnym rozdziale (zaraz po liście artykułów tworzących cykl), albo w oddzielnej sekcji w rozdziale przedstawiającym osiągnięcia.

Wnioski końcowe

Autor pracy wykazał się umiejętnością prowadzenia pracy badawczej, właściwej analizy danych oraz poprawnego wnioskowania. Dowiódł, iż w wystarczającym stopniu zapoznał się z aktualnym stanem wiedzy w zakresie problematyki podejmowanej w pracy doktorskiej.

Posiada ogólną wiedzę teoretyczną w obszarze Sztucznej Inteligencji, Bioinformatyki oraz Biologii Molekularnej, wykazuje się umiejętnością samodzielnego prowadzenia pracy naukowej, a także pracy w zespole badawczym oraz zastosowania uzyskanej wiedzy do rozwiązywania problemów praktycznych.

Recenzowana praca zawiera oryginalne rozwiązanie problemu naukowego. Uzyskane przez Autora wyniki badań stanowiące trzon rozprawy doktorskiej zostały opublikowane w pierwszoautorskich artykułach w dobrych, w większości wysokopunktowanych czasopismach z dziedziny.

Pracę mgr. inż. Mateusza Chilińskiego pt. „Spatial Network Model of Sequence and Structure Diversity of Human Genome at a Population Scale" oceniam pozytywnie i stwierdzam, że spełnia ona 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. Wnoszę o dopuszczenie mgr. inż. Mateusza Chilińskiego do kolejnego etapu postępowania kwalifikacyjnego w celu nadania stopnia naukowego doktora w dziedzinie nauk inżynieryjno-technicznych w dyscyplinie informatyka techniczna i telekomunikacja.

Marta Szachniuk

prof. dr hab. inż. Marta Szachniuk