Uniwersytet Wrocławski

WYDZIAŁ BIOTECHNOLOGII
prof. dr hab. **Paweł Mackiewicz**
ZAKŁAD BIOINFORMATYKI I GENOMIKI
ul. F. Joliot-Curie 14a
50-383 Wrocław
tel. +48 71 375 63 03
pamac@smorfland.uni.wroc.pl

Wrocław, 15.09.2024

Recenzja rozprawy doktorskiej Pana mgra Mateusza Chilińskiego pt. "Spatial Network Model of Sequence and Structure Diversity of Human Genome at a Population Scale"

Znaczne zmniejszenie kosztów sekwencjonowania umożliwia bardzo szybkie uzyskanie sekwencji wielu genomów człowieka. Jednakże, aby zrozumieć organizację i odczyt zakodowanej informacji genetycznej konieczne jest przeprowadzenie wielu kosztownych eksperymentów dotyczących ekspresji genów (RNA-Seqs), dostępności chromatyny (ATAC-Seqs), wiązania białek do DNA (ChIP-Seqs), określenia miejsc kontaktowych chromatyny (Hi-C) i formowania pętli przy pomocy różnych białek (ChIA-PET/HiChIP). Ze względu na złożoność tych zjawisk, analizy tych eksperymentów są skomplikowane i czasochłonne. Z pomocą przychodzą różne techniki uczenia maszynowego, które umożliwiają odpowiednie interpretowanie wyników i wgląd w procesy ekspresji genów związane z organizacją chromatyny. Ponieważ metody te nie są wciąż doskonałe istnieje potrzeba ich rozwijania. Dlatego bardzo słusznie ambitnym przedmiotem pracy doktorskiej Pana mgra Mateusza Chilińskiego stało się opracowanie zaawansowanych modeli uczenia maszynowego, które umożliwiają analizę sekwencji DNA oraz bardziej wymagających danych eksperymentalnych dotyczących przestrzennej organizacji chromatyny i ekspresji genów. Analizy wykonano głównie na danych dotyczących genomu człowieka.

Rozprawa doktorska została napisana w języku angielskim ze streszczeniem w języku polskim. Zawiera ona: abstract i streszczenie, spis treści, wstęp z celami i wykazem publikacji związanych z rozprawą doktorską, wyniki przedstawiające najważniejsze osiągnięcia naukowe zawarte w załączonych pięciu pracach, wnioski, rozdział zawierający pozostałe osiągnięcia doktoranta, bibliografię, kopie publikacji będące przedmiotem rozprawy, oświadczenia współautorów i kopie dodatkowych publikacji.

Wstęp stanowi zwięzłe, ale dobre wprowadzenie do rozwoju technik biologii molekularnej związanych z badaniami genomu, metod sekwencjonowania oraz zróżnicowania genomów na poziomie sekwencyjnym. Osobny podrozdział poświęcono algorytmom uczenia maszynowego: konwolucyjnej sieci neuronowej (splotowej), architekturze transformerów i BERT, modelom dyfuzyjnym oraz wzmacniaczom (boosting) gradientowym. Opisy są jasne, jednakże przy wzorach dobrze byłoby wytłumaczyć wszystkie oznaczenia symboli. Mam również pytanie związane z sieciami konwolucyjnymi. We Wstępie doktorant napisał, że w jej przypadku często stosuje się rektyfikowaną jednostkę liniową (ReLu) jako funkcję aktywacji. Czy inne funkcje aktywacji też mają zastosowanie i w jakich przypadkach?

Cele pracy doktorskiej zostały jasno sformułowano i w pełni zrealizowane. Było nimi opracowanie i zastosowanie narzędzi obliczeniowych do analizy genomowych danych sekwencyjnych oraz wyników eksperymentów związanych z ekspresji genów i przestrzenną organizacją chromatyny. Wyniki zostały również przejrzyście streszczone i zaprezentowane. Rozdział Wnioski jest napisany przejrzyście i zawiera najważniejsze informacje uzyskane z przeprowadzonych badań i zamieszczonych prac.

Osiągnięcia związane z rozprawą doktorską zostały przedstawione w pięciu pracach. We wszystkich z nich doktorant jest pierwszym autorem. Wyniki trzech prac zostały opublikowane w prestiżowych czasopismach (*Seminars in Cell & Developmental Biology*, *Bioinformatics* i *Scientific Reports*). Jedna praca wysłana do redakcji jest zamieszczona na stronie bioRxiv. Wyniki zaprezentowane w tych pracach są spójne i układają się w jedną całość, ponieważ dotyczą zastosowania metod uczenia maszynowego do badania danych genomowych.

Załączone oświadczenia autorów nie pozostawiają wątpliwości, że Pan Mateusz Chiliński miał wiodący i istotny wpływ na koncepcję prac i uzyskanie oraz opisanie wyników w publikacjach będących przedmiotem pracy doktorskiej. Przedstawiona rozprawa opisująca wyniki prac jest oryginalnym wkładem doktoranta.

Pierwsza publikacja jest artykułem przeglądowym, który stanowi dobre wprowadzenie do przedmiotu pracy doktorskiej. Opisano w niej organizację materiału genetycznego od sekwencji genomu do przestrzennej organizacji chromatyny wyższego rzędu. Artykuł przedstawia związki między sekwencją DNA, genomiką przestrzenną i ekspresją informacji genetycznej. Przedstawiono w niej organizację chromatyny we wszystkich skalach począwszy od podwójnej helisy DNA, przez nukleosomy, poziom 10 nm włókna, pętle chromatyny,

topologicznie powiązane domeny, domeny kontaktowe chromatyny i regiony chromosomowe do samych chromosomów i kompartymentalizacji A/B. Organizację chromosomu zaprezentowano na przejrzystych rysunkach obrazujących wyniki eksperymentów Hi-C i ChIA-PET: diagramach łukowych, mapach ciepła i sieciach oddziaływań. W opisie skupiono się głównie na pętlach chromatynowych i domenach chromatyny, ponieważ istnieje wiele chorób związanych z ich zaburzeniami. Szczegółowiej omówiono różne rodzaje mutacji genomowych: polimorfizmy pojedynczego nukleotydu (SNP), indele (krótkie delecje/insercje) i warianty strukturalne o długości ponad 50 pz (SV). Przedstawiono wiele sposobów wykrywania SV w tym read-pair, read-depth, split-read i local-assembly. Wyjaśniono także mechanizmy powstawania chorób, jak autyzm, schizofrenia, łuszczyca, hemofilia A i zaburzenia autoimmunologiczne, w oparciu o rearanżacje genomowe zakłócające działanie domen i mające swoje odzwierciedlenie w wariantach strukturalnych. Przeglądówka jest dobrze zorganizowana pod względem formalnym i graficznym. Uwzględniono w niej najbardziej istotne fakty i nowe odkrycia związane z przedstawianym zagadnieniem. Pokazuje ona, że doktorant bardzo dobrze zna i rozumie podłoże molekularne zjawisk badanych komputerowo.

W drugiej publikacji został przedstawiony pakiet oprogramowania ConsensuSV do odkrywania genomowych wariantów strukturalnych. Istnieje konieczność opracowywania takich programów, ponieważ identyfikacja tych wariantów jest bardzo trudna z powodu ograniczeń eksperymentalnych, a dotychczasowe algorytmy mają wiele ograniczeń. Dlatego doktorant wraz z promotorem opracowali uniwersalny program oparty na wielu algorytmach i sztucznej inteligencji. Oprogramowanie składa się z dwóch części. Pierwsza część algorytmu wykorzystuje głęboką sieć neuronową do uzyskiwania wariantów konsensusowych w oparciu o wyniki wielu wcześniej znanych algorytmów, ocenia każde narzędzie i na podstawie wspólnych wyników podaje pozycje danego wariantu. Druga częścią programu jest potokiem (pipeline), który automatyzuje analizy po dostarczeniu surowych danych sekwencyjnych. Ważne jest to, że ta część programu przy pomocy pierwszej części wykonuje kompleksowe analizy niewymagające innych indywidualnych narzędzi, ponieważ program wykonuje kontrolę jakości, przyrównanie, konwertowanie do formatu BAM, sortowanie, indeksowanie, eliminowanie duplikatów oraz identyfikowanie wariantów. Wyniki zapisywane są w formie standardowych plików VCF (Variant Call Format), które są pomocne w innych analizach. Warto dodać, że program podaje także wyniki dla krótszych wariantów, czyli indeli i polimorfizmów pojedynczego nukleotydu, co świadczy o jego uniwersalności. Dzięki znacznemu polepszeniu program jest w stanie wykryć precyzyjniej więcej wariantów

strukturalnych niż inne programy. W wykonanym teście znalazł on prawie 2500 wariantów nieodkrytych przez programy MetaSV i FusorSV, które łącznie zidentyfikowały około 300 wariantów nieznalezionych przez ConsensusSV. Zaletą programu jest to, że jest przyjazny dla biologów nieobeznanych w programowaniu. W czasie obrony prosił bym doktoranta o przedstawienie w jaki sposób surowe dane są konwertowane i podawane do sieci neuronowej w tym programie. Jak program radzi sobie w przypadku niejednoznacznych nukleotydów, np. R, Y, W, S M, K i N?

Trzecia publikacja opisuje pierwsze wykorzystanie głębokiego hybrydowego uczenia w przewidywaniu pętli chromatyny tylko na podstawie sekwencji genomowej bez kosztownych i czasochłonnych wyników eksperymentów badających przestrzenną organizację chromatyny. Mimo to, program wykazał się dobrą jakością przewidywań. Uważam, że jest to duża zaleta tego programu. W szczególności zastosowano wiele wytrenowanych modeli DNABERT, a także klasyczne metody uczenia maszynowego: maszyna wektorów nośnych (SVM), lasy losowe (RF) i algorytm najbliższych sąsiadów (KNN). Model DNABERT został odpowiednio dostosowany do potrzeb tego zagadnienia. W końcowym etapie wyniki modeli były łączone. Dzięki tym zmianom uzyskane wyniki okazały się bardzo wiarygodne, ponieważ uzyskano około 84% dokładności w testowanych zestawach danych. W programie na etapie uczenia użyto pętli jako zbioru pozytywnego, a losowych części genomu jako zestawu negatywnego. Jak stwierdzono może to wprowadzać pewne odchylenie w przewidywaniach. Prosiłbym doktoranta o wyjaśnienie, dlaczego tak może być i jaki lepszy byłby zbiór negatywny, aby uniknąć tych problemów.

Czwarta publikacja prezentuje wykorzystanie zaawansowanego modelu HiCDiffusion złożonego z wielu modułów. Jego celem jest uzyskiwanie lepszej jakości obrazów macierzy kontaktów przestrzennej chromatyny Hi-C przedstawianych jako mapy ciepła. Dotychczas stosowane były metody oparte na architekturze transformera kodera-dekodera uwzględniającego uczenie kontekstowe, w której jednowymiarowa sekwencja zakodowana do ukrytej reprezentacji była dekodowana za pomocą splotów dwuwymiarowych do macierzy kontaktów. Metody te, przy jednoczesnym uzyskiwaniu wysokich współczynników korelacji i ulepszonych metryk, dostarczały macierze Hi-C, które były rozmyte z powodu architektury modelu głębokiego uczenia, a przewidywania dostarczały zbyt wielu możliwych pętli. W nowym podejściu rozwiązano te problemy. Najpierw wyuczono sieć neuronową kodera-dekodera, a następnie użyto jej do modelu dyfuzyjnego, w którym kierowano dyfuzją za pomocą ukrytej reprezentacji sekwencji, a także końcowego wyniku z kodera-dekodera. W ten

sposób uzyskano macierze Hi-C o wysokiej rozdzielczości, które lepiej przypominały wyniki eksperymentalne poprawiając parametr Fréchet Inception Distance (FID) średnio 12 razy, przy największej poprawie nawet 35 razy. Ponadto macierze te charakteryzowały się podobnymi wartościami klasycznych metryk w porównaniu do najnowocześniejszych architektur kodera-dekodera. Ponieważ w analizach do porównania macierzy sosowany jest współczynnik korelacji Pearsona, mam pytanie do doktoranta czy jest ona zasadny, ponieważ dane nie są niezależne i mogą nie spełniać normalności rozkładu. Czy nie lepiej zastosować współczynnik korelacji Spearmana? Jak wtedy by wychodziły wyniki?

Piąta publikacja w umiejętny sposób łączy analizy interakcji przestrzennych chromatyny z przewidywaniami ekspresji genów. Przeprowadzone analizy wyraźnie pokazują, że jakość tych przewidywań jest dużo lepsza, jeśli uwzględni się nie tylko sekwencje, które są blisko promotora transkrypcji, ale także te, które są położone blisko w przestrzeni. Dotyczy to np. wzmacniaczy, które są w sekwencji położone daleko od miejsca promotorowego, a są kluczowe dla prawidłowego funkcjonowania danego genu. W tym celu zastosowano odpowiedni model sztucznej inteligencji, już istniejący algorytm, ExPecto, który został zmodyfikowany tak, aby przetworzyć dodatkowo podaną sekwencję DNA, która jest przestrzennie blisko danego miejsca rozpoczęcia transkrypcji, ale bardzo daleko pod względem odległości liniowej. Opracowany model porównano z modelem podstawowym ExPecto używając 32 zestawów danych dotyczących przestrzennej organizacji chromatyny z eksperymentów ChIA-PET. Uzyskano znaczący średni wzrost w większości modeli. Wyniki zależały od tego jakie białko pośredniczyło w tworzeniu pętli. Uważam, że uwzględnianie danych przestrzennych jest ważne w polepszeniu przewidywań ekspresji genów, jednakże tworzenie się pętli może bardzo zależeć od wielu czynników.

Reasumując mogę stwierdzić, że doktorant włożył dużo trudu w opracowanie nowatorskich programów i przeprowadzone analizy, a przedstawione opisy wyników świadczą o jego dużej dojrzałości naukowej, rzetelności i umiejętności wydobywania najważniejszych informacji z uzyskanych rezultatów. Nie mam zastrzeżeń do metodyki przeprowadzonych analiz. Mam jednak ogólne pytanie: jak dobierano liczbę warstw sieci i unikano przeuczenia modeli? Opisy są dobrze przedstawione pod względem formalnym. Praca i artykuły są napisane poprawnym językiem i stylem.

Tematyka pracy doktorskiej jest bardzo zasadna, ponieważ istnieje potrzeba tworzenia narzędzi i analizy struktury przestrzennej chromatyny z uwzględnieniem zmian strukturalnych i elementów regulujących transkrypcję, co stało się przedmiotem rozprawy. Należy podkreślić,

że modele wykorzystane w pracy są zaawansowane i bardzo nowoczesne, a przedstawione wyniki pokazują, że wykorzystanie nowoczesnych metod uczenia maszynowego, a w szczególności głębokich sieci neuronowych, jest bardzo owocne w biologii molekularnej i przynosi intersujące odkrycia. Stwierdzam, że recenzowana rozprawa z publikacjami stanowi istotny wkład w przewidywanie i analizę struktury przestrzennej chromatyny.

Na uwagę zasługuje dodatkowy dorobek publikacyjny doktoranta obejmujący sześć prac, trzy wizyty naukowe oraz udział w 17 prezentacjach konferencyjnych, uczestnictwo w czterech projektach badawczych i sześciu kursach. Pan Chiliński uzyskał także dwie nagrody. Wszystkie osiągnięcia świadczą o wszechstronności i dojrzałości doktoranta.

Uważam, więc, że przedstawiona do recenzji rozprawa doktorska spełnia wszystkie wymogi Ustawy o Stopniach Naukowych. Zgłaszam, zatem wniosek do Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej o uznanie rozprawy Pana mgra Mateusza Chilińskiego za odpowiadającą wymogom stawianym rozprawom doktorskim i o dopuszczenie doktoranta do dalszych etapów przewodu doktorskiego. W związku z tym, że doktorant miał postawiony trudny cel badawczy i go efektywnie rozwiązał, a wyniki zostały przedstawione w pięciu bardzo dobrych pracach proponuję wyróżnić rozprawę.

Prof. dr hab. Paweł Mackiewicz