Wydział Automatyki, Elektroniki i Informatyki
Katedra Sieci i Systemów Komputerowych

dr hab. inż.
Prof. Politechniki Śląskiej
Aleksandra Gruca

Gliwice, 02.08.2023

Recenzja rozprawy doktorskiej mgr inż Mateusza Chilińskiego

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

Ukończonej na Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej

Pod opieką promotora Profesora Dariusza Plewczyńskiego

Tematyka i cel pracy, problem badawczy i jego znaczenie

Przedstawiona mi do recenzji rozprawa doktorska powstała na Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej pod kierunkiem Profesora Dariusza Plewczyńskiego.

Tematyka rozprawy obejmuje zastosowanie metod sztucznej inteligencji, a w szczególności głębokich sieci neuronowych do przewidywania wyników eksperymentów biologicznych. W swoich badaniach Doktorant koncentruje się na analizie informacji na temat przestrzennej organizacji struktury 3D genomu. Celem prac jest stworzenie algorytmów, które na podstawie sekwencji DNA przewidywałyby występowanie wariantów strukturalnych oraz analizę trójwymiarowej struktury chromatyny.

Analiza trójwymiarowej struktury chromatyny to dynamicznie rozwijająca się dyscyplina, a wiedza na temat przestrzennej organizacji chromatyny w jądrze komórkowym jest istotna z punktu poznania mechanizmów regulacji ekspresji genów, co bezpośrednio przekłada się na zrozumienie funkcji genomu. Analizowane w niniejszej rozprawie technologie eksperymentalne ChIA-PET and Hi-C pozwalające na badanie interakcji chromatyny powstały dopiero pod koniec pierwszej dekady tego wieku, w związku z tym są relatywnie dość nowe oraz kosztowne. Z kolei technologie sekwencjonowania genomu są już na tyle ugruntowane, że informacje o sekwencji DNA są łatwo dostępne, a koszt ich uzyskania relatywnie niski. W związku z tym prace badawcze mające na celu stworzenie metod pozwalających na uzyskanie informacji na temat trójwymiarowej struktury chromatyny na podstawie sekwencji DNA wydają się bardzo aktualne i mają

Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
Katedra Sieci i Systemów Komputerowych

ul. Akademicka 16, pok. 414, 44-100 Gliwice
+48 32 237 21 51 / +48 32 237 27 33
Aleksandra.gruca@polsl.pl

NIP 631 020 07 36
ING Bank Śląski S.A. o/Gliwice 60 1050 1230 1000 0002 0211 3056

potencjalnie duże możliwości praktycznego wykorzystania wyników. Obserwowany w ostatnich latach rozwój głębokich sieci neuronowych otwiera nowe możliwości analizy danych pozwalając na zastąpienie kosztownych eksperymentów laboratoryjnych wynikami analiz obliczeniowych.

W związku z tym należy stwierdzić, że problem, który podjął rozwiązać się Doktorant w przedstawionej pracy jest istotny i aktualny i bez wątpienia wpisuje się w aktualne trendy w dziedzinie bioinformatyki oraz biologii obliczeniowej.

Charakterystyka rozprawy

Przedstawiona praca doktorska została złożona w postaci zbioru opublikowanych i powiązanych tematycznie artykułów naukowych. Na cykl publikacji składa się pięć artykułów naukowych, cztery z nich zostały opublikowane w bardzo dobrych czasopismach naukowych, piąty został opublikowany w formie preprintu na platformie *bioRxiv*. W wszystkich artykułach cyklu Doktorant jest pierwszym autorem.

Pierwszy z artykułów [**P1**] pt. *„From DNA human sequence to the chromatin higher order organisation and its biological meaning: Using biomolecular interaction networks to understand the influence of structural variation on spatial genome organisation and its functional effect”* opublikowany został w czasopiśmie Seminars in Cell & Developmental Biology (IF 7.3, 140 pkt. MNiSW). Praca ta stanowi artykuł przeglądowy i koncentruje się na opisaniu przestrzennej architektury chromatyny ludzkiego genomu, w tym terytoriów chromosomalnych, przedziałów A/B, topologicznie asocjujących domen (TAD) i pętli. Przedstawiono tu również aktualne technologie eksperymentalne stosowane do wykrywania interakcji chromatynowych, w tym Hi-C i ChIA-PET. Dodatkowo w pracy omówiono najpopularniejsze metody wykrywania wariantów strukturalnych i przedstawiono narzędzi oraz programów przewidujących ich wystąpienie. Wraz z opisem algorytmów dokonano ich porównania przedstawiając ich mocne i słabe strony. W pracy omówiono też najnowsze odkrycia dotyczące wpływu wariantów strukturalnych na strukturę 3D genomu oraz potencjalne konsekwencje tego wpływu w odniesieniu do ekspresji genów. Ostatnia część artykułu koncentruje się na omówieniu "omicznych" sieci i możliwości ich wykorzystania do modelowania interakcji biomolekularnych.

Artykuł [**P2**] pt. “*ConsensuSV - from the whole-genome sequencing data to the complete variant list*” opublikowany został w czasopiśmie Bioinformatics (IF 5.8, 200 pkt. MNiSW). W pracy tej przedstawiono narzędzie wykorzystujące konsensusowy algorytm oparty o głębokie sieci neuronowe do wykrywania wariantów strukturalnych. Przestawiony w publikacji pakiet oprogramowania jest podzielony na dwa moduły: pierwszy z nich: ConsensuSV-core służy do uzyskiwania konsensusu z już wywołanych wariantów strukturalnych. Dane wejściowe modułu to pliki VCF (ang. *Variant Call Format*), a rezultat to scalony plik meta-VCF zawierający wyniki konsensusowe. Drugi moduł narzędzia, ConsensuSV-pipeline, jest kompletnym rozwiązaniem zdolnym do przewidywania wszystkich typów wariantów, wykorzystując jako dane wejściowe surowe pliki FASTQ Illumina. Wynikiem są tu listy przewidzianych wariantów strukturalnych, Indel-i oraz SNPs (ang. Single Nucleotide Polymorphism). Zaprojektowane narzędzie pozwala na pełną automatyzację analizy.

Kolejny artykuł [**P3**] zatytułowany *„Prediction of chromatin looping using deep hybrid learning (DHL)”* opublikowany został w czasopiśmie Quantitative Biology (IF 3.1, 70 pkt. MNISW). W pracy tej przedstawiono metodę do przewidywania pętli chromatyny 3D na podstawie sekwencji DNA. W tym celu zaproponowano

algorytm wykorzystujący strategię głębokiego uczenia hybrydowego, polegającej na połączeniu wyników predykcji z pretrenowanej sieci DNABERT wykorzystującej architekturę transformerów oraz trzech algorytmów klasycznego uczenia maszynowego. W porównaniu do metod opartych o klasyczne algorytmy maszynowego uczenia, a także podejścia opartego tylko o sieć DNABERT, podejście hybrydowe pozwoliło na uzyskanie lepszych wartości miar oceny jakości algorytmu.

Czwarty artykuł cyklu [**P4**] pt. „*HiCDiffusion - diffusion-enhanced, transformer-based prediction of chromatin interactions from DNA sequences*" opublikowany został w formie preprintu na platformie bioRxiv, a także , zgodnie z informacją podaną w streszczeniu pracy, znajduje się w recenzji. Artykuł prezentuje model *HiCDiffusion* do przewidywania interakcji chromatyny na bazie sekwencji DNA. Wejściem metody jest sekwencja DNA, a wynikiem przewidziana macierz Hi-C kontaktów chromatynowych. Opisany algorytm w pierwszym kroku trenuje sieć neuronową o architekturze enkoder-dekoder. Następnie, sieć ta jest wykorzystywana do dalszego przetwarzania w paradygmacie uczenia transferowego (ang. *transfer learning*), gdzie wstępnie wytrenowana sieć enkoder-dekoder jest częścią finalnej architektury opartej na modelach dyfuzyjnych. W wyniku zaproponowanego podejścia udało się stworzyć algorytm generujący bardziej realistyczne macierze Hi-C przy zachowanej wysokiej jakości współczynniku korelacji Pearsona, porównywalnej z narzędziami *state-of-the-art.*

Ostania praca cyklu [**P5**] to artykuł pt. "*Enhanced performance of gene expression predictive models with protein-mediated spatial chromatin interactions*", opublikowany w czasopiśmie Scientific Reports (IF 4.996, 140 pkt. MNiSW). W pracy tej zaproponowano algorytm SpEx, który powstał poprzez zmodyfikowanie algorytmu ExPecto. Algorytm ExPecto działa w oparciu o konwolucyjne sieci neuronowe i przewiduje ekspresję genów na podstawie sekwencji DNA. Zaproponowana w pracy modyfikacja dodaje do procesu przetwarzania analizę przestrzenne *heatmapy* interakcji chromatynowych z wyników eksperymentów ChIA-PET. Dołączenie na wejściu metody informacji przestrzennej pozwoliło uzyskać wyższe wartości współczynnika korelacji Spearmana w porównaniu do modelu bazowego.

Opinia o rozprawie

Należy podkreślić, że problem, który podjął się rozwiązać Doktorant jest niewątpliwie ważny i trudny. Wykorzystanie metod opartych o głębokie sieci neuronowe do analizy przestrzennej struktury genomu wpisuje się w trendy najnowszych badań zarówno w dziedzinie maszynowego uczenia, sztucznej inteligencji jak i biologii obliczeniowej.

Po analizie artykułów (patrz moja uwaga w części „Uwagi krytyczne i dyskusyjne") nie mam wątpliwości, że przedstawione w ramach pracy doktorskiej prace stanowią spójny cykl powiązanych ze sobą publikacji w tematyce rozprawy doktorskiej. Poza jednym z artkułów, który znajduje się w recenzji, pozostałe prace zostały opublikowane w bardzo dobrych czasopismach naukowych. Należy tez podkreślić i bardzo wysoko ocenić fakt, że we wszystkich tych pracach Doktorant jest pierwszym autorem.

Dobór artykułów stanowiących cykl jest spójny i logiczny. Pierwsze praca jest pracą przeglądową i omawia zależności pomiędzy sekwencją DNA, a strukturą 3D genomu. Omawia również metody detekcji wariantów strukturalnych i ich wpływ na przestrzenną strukturę chromatyny oraz metody analizy struktury 3D genomu

oparte o modele sieciowe. Kolejna praca omawia narzędzie do detekcji wariantów strukturalnych. Następne trzy prace związane są z analizą przestrzennej struktury chromatyny. Praca [**P3**] opisuje metodę przewidywania pętli chromatyny 3D na podstawie sekwencji DNA, a praca [**P4**] opisuje model do przewidywania interakcji chromatyny na bazie sekwencji DNA. Ostatnia praca cyklu przedstawia algorytm, gdzie modyfikacja polegająca na dołożeniu informacji na temat interakcji chromatynowych do informacji o sekwencji DNA pozwala na poprawienie jakości predykcji ekspresji genów.

Spis literatury składa się ze 110 pozycji i oddaje aktualny stan wiedzy w zakresie, którego dotyczy rozprawa, chociaż spośród cytowanych pozycji tylko 22 są artykułami cytującymi prace opublikowane na przestrzeni ostatnich 5-ciu lat. Pozostałe artykuły cytują prace starsze. Biorąc pod uwagę, iż tematyka, która zajmuje się doktorant, a w szczególności wykorzystanie głębokich sieci neuronowych do analizy danych biologicznych i biomedycznych to dziedzina bardzo nowa, być może Doktorant powinien nieco odświeżyć swoją wiedzę w zakresie literatury przedmiotu rozprawy.

Pozytywnie też oceniam fakt, iż kod źródłowy metod opisanych w pracach [**P2**], [**P4**] oraz [**P5**] został udostępniony środowisku naukowemu na platformie GitHub. Niestety, nie udało mi się znaleźć żadnych informacji odnośnie dostępności kodu opisanego w pracy [**P3**].

Podsumowując, uważam iż zastosowane metody badawcze są odpowiednie do rozwiązywanego problemu badawczego. Wykorzystane w pracy podejścia do analizy danych wskazują na dobrą znajomość przez Doktoranta nowoczesnych i efektywnych metod stosowanych w dziedzinie informatyki.

Przedstawiony w rozprawie cel:

stworzenie modeli uczenia maszynowego, które umożliwiają przejście od prostszych i łatwiejszych do uzyskania danych - a mianowicie sekwencji DNA, do bardziej wymagających danych eksperymentalnych - w tym informacji przestrzennych o jądrach i ekspresji genów.

został zrealizowany.

Uwagi krytyczne i dyskusyjne

Na początku tej części chciałbym podkreślić, że nie znalazłam w przedstawionych wynikach żadnych zasadniczych błędów merytorycznych. Wszystkie poniższe uwagi wynikają z chęci podjęcia dyskusji i dialogu na temat niektórych aspektów pracy. Uwagi te nie obniżają mojej pozytywnej oceny pracy.

- W ocenie przedstawionego cyklu pubilkacji pewną trudność stanowiło zrozumienie całościowego obrazu zawartego w artykułach składających się na powiązany tematycznie cykl. Spowodowane to jest faktem, że opracowane metody oraz uzyskane wyniki są w wielu przypadkach przedstawione w Autoreferacie w sposób skrótowy i dopiero przeczytanie artykułów pozwoliło na pełne zrozumienie kontekstu wykonanych badań oraz wyników. Przykładowo Doktorant w Autoreferacie umieszcza rysunki z publikacji z minimalnym komentarzem co się na nich znajduje (np. *Fig. 5* czy *Fig. 8*

z minimalnym komentarzem: „wyniki przedstawiono na rysunku X"), podczas gdy w opublikowanych artykułach pod tymi samymi rycinami znajduje się ich szczegółowy opis.

- Temat przedstawionej rozprawy brzmi: *Spatial network model of sequence and structure diversity of Human genome at a population scale*. W związku z tym interesuje mnie w jakim aspekcie przeprowadzone analizy prowadzone były w skali populacyjnej i jakie udało się uzyskać wyniki przedstawionych badań w tym kontekście?

- W opisie pracy [**P3**] Doktorant wspomina, że zbiorem negatywnym dla algorytmu przewidywania pętli były losowo wybrane fragmenty genomu. Słusznie też zauważa, że może to doprowadzić do błędnego działania algorytmu, który do podejmowania decyzji o przypisaniu do klasy pozytywnej wykorzystywać będzie nie cechy sekwencji związane z tworzeniem się pętli, ale regiony zawierające fragmenty kodujące białka. W jaki sposób można by lepiej skonstruować zbiór negatywy, aby wykluczyć tego rodzaju sytuacje.

- W pracy [**P3**] nie znalazłam porównania wyników uzyskanych za pomocą zaproponowanej metody hybrydowej z innymi metodami predykcji pętli chromatynowych na bazie sekwencji DNA. Jak zaproponowana metoda wypada w porównaniu z innymi opublikowanymi metodami (przykładowo, *Lv et al., 2021* [PMID: 33634313], *Zhang et. al., 2022* [PMID: 35997565]).

- W pracy [**P4**] do oceny jakości zmodyfikowanego algorytmu wykorzystano współczynnik korelacji Pearsona, a w pracy [**P5**] współczynniki korelacji Pearsona oraz Spearmana. Czy uzyskane różnice pomiędzy algorytmami podstawowym a nowymi podejściami są istotne statycznie? W pracy [P4] do oceny jakości wykorzystano współczynnik jedynie korelacji Pearsona. Czy wykonano testy sprawdzające czy analizowane dane spełniają założenia pozwalające na zastosowanie tej metody?

Podsumowanie

Pan mgr inż. Mateusz Chyliński przedstawił zbiór opublikowanych i powiązanych tematycznie artykułów naukowych stanowiących oryginalne rozwiązanie problemu naukowego. Opublikowanie uzyskanych wyników w postaci czterech artykułów naukowych, w których Doktorant jest pierwszym autorem wskazuje, iż posiada on ogólną wiedzę teoretyczną w dyscyplinie informatyka techniczna i telekomunikacja oraz umiejętność prowadzenia pracy naukowej.

Poza publikacjami składającymi się na pracę doktorską, Doktorant jest również współautorem sześciu innych publikacji. Cztery z nich zostały opublikowane w bardzo dobrych czasopismach naukowych (powyżej 100 punktów MNiSW). Dodatkowo Doktorant brał udział w czterech projektach badawczych, w tym w trzech z nich był stypendystą. Brał też udział w trzech wizytach naukowych, przy czym czas trwania ostatniej z nich to sześć miesięcy. Dodatkowe publikacje oraz udział w wymianach naukowych wskazują na bardzo duże zaangażowanie doktoranta w prace naukowe.

Biorąc pod uwagę powyższą ocenę, stwierdzam, że przedstawiona do oceny praca doktorska w pełni odpowiada warunkom określonym w Art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (tekst jednolity Dz. U. z 2023 r. poz. 742 z późn. zm.) i na tej podstawie wnoszę do Wysokiej Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej o dopuszczenie mgr inż. Mateusz Chylińskiego do dalszych etapów przewodu doktorskiego. Równocześnie biorąc pod uwagę, iż wyniki prac opublikowane zostały w bardzo dobrych czasopismach naukowych, a Doktorant we wszystkich tych artykułach jest w nich pierwszym autorem, wnoszę o wyróżnienie rozprawy.

Signed by /
Podpisano przez:

Aleksandra Helena Gruca
Politechnika Śląska

Date / Data:
2024-08-02 22:00

dr inż. hab. Aleksandra Gruca, prof. PŚ.