prof. dr hab. inż. Marta Kasprzak
Instytut Informatyki
Politechnika Poznańska

Poznań, 24.08.2023

Recenzja rozprawy doktorskiej mgr. Kaustava Sengupty pt. "The Meta-Network Model of the Complete Human Biomolecular Interactome at the Whole Cell Scale"

1. Problematyka naukowa rozprawy

DNA w komórce występuje w postaci upakowanej hierarchicznie, którą osiąga w kolejnych krokach zwijania z udziałem białek. Stopień upakowania DNA wpływa na uaktywnienie się lub wyciszenie procesu transkrypcji genu, koncentrującego się na miejscu chwilowego rozkurczenia chromatyny. Obserwacja struktury przestrzennej chromatyny, np. z użyciem technik Hi-C lub ChIA-PET, ma zatem znaczenie kluczowe dla zrozumienia procesu ekspresji genów. Uzupełnienie tej informacji o dane pochodzące z innych eksperymentów badających przestrzenne interakcje cząsteczek w komórce pozwala uzyskać bardziej kompleksowy obraz procesów zachodzących w organizmie.

Tematyka rozprawy, przedłożonej w formie jednotematycznego cyklu publikacji, wpisuje się w obszar bioinformatyki. Koncentruje się na zintegrowaniu danych różnego pochodzenia, opisujących własności i relacje struktur białkowych, chromatynowych i DNA, oraz na wykorzystaniu ich do zamodelowania problemów związanych z przestrzenną organizacją DNA w komórce i przewidywaniem funkcji białek. Ze względu na fundamentalną rolę badanych procesów w funkcjonowaniu organizmów, ich zrozumienie i analiza stanowi istotny wkład w stan wiedzy. Prace ukierunkowane zostały na zgłębianie mechanizmów zachodzących w organizmie człowieka i pozwalają wyciągnąć wnioski, które można bezpośrednio przenieść na grunt badań medycznych i farmaceutycznych. Problemy te zostały rozwiązane przez mgr. Kaustava Senguptę na drodze zamodelowania ich na gruncie teorii grafów. Jako dane wejściowe wykorzystał rezultaty zróżnicowanych eksperymentów biologicznych, uzyskanych głównie za pomocą najnowszych technik analizy molekularnej. Zaproponował nowe modele grafowe uwzględniające takie heterogeniczne dane w celu uzyskania jak najbardziej kompleksowej informacji, pogłębiającej opis złożonych systemów biologicznych. Do przetwarzania tej informacji wykorzystał rozmaite podejścia statystyczne i algorytmiczne. Tematyka ta może być zatem przedmiotem rozprawy doktorskiej w dyscyplinie informatyka techniczna i telekomunikacja.

2. Opinia o rozprawie

Przedłożony do oceny cykl publikacji składa się z pięciu artykułów opublikowanych w latach 2020–2023. Artykuły są wieloautorskie, z liczbą autorów od trzech do dziewięciu, w trzech z nich mgr Kaustav Sengupta jest wymieniony na pierwszej pozycji. Cztery z nich opublikowane zostały w czasopismach bioinformatycznych lub biologicznych o liczbie punktów MEiN 140 [P1, P3, P5] lub 100 [P4]. Deklarowany wkład Doktoranta w algorytmy i analizy przedstawione w tych artykułach jest następujący: jeden z sześciu autorów w [P1], jedyny w [P2], jeden z dwóch w [P3], jeden z pięciu w [P4] i jeden z czterech w [P5]. Deklarowany wkład w całość prac nad artykułami: 40% w [P1], 70% w [P2], 45% w [P3], 25% w [P4], 55% w [P5].

W odniesieniu do pracy [P1], Doktorant miał wkład w opracowanie metody odwołującej się do teorii grafów, która symuluje tworzenie się pętli chromatynowych i zwijanie włókna chromatyny. W tym modelu wierzchołki w grafie odpowiadają miejscom zakotwiczenia białka CTCF, krawędzie odpowiadają pętlom chromatynowym, a cały graf oddaje powiązania przestrzenne obecne w chromosomie. Proces zwijania chromatyny został zasymulowany algorytmami stopniowo uzupełniającymi krawędzie w grafie w różnej kolejności, która determinuje kolejność podlegania poszczególnych fragmentów chromatyny mechanizmowi ustalania konformacji. Dwa spośród siedmiu tych algorytmów opracował Doktorant. Dodawane krawędzie odzwierciedlają rzeczywiście istniejące pętle chromatynowe, badania jednak wykazały, że wystarczy wprowadzić około 80% ich całkowitej liczby i uzyskać trafnie przewidzianą strukturę przestrzenną. Symulacje dokonane różnymi algorytmami, wykorzystującymi wyniki eksperymentu ChIA-PET osobno lub uzupełnione o dodatkowe dane, wykazały wysoką skuteczność modelu uwzględniającego szeroko ujęte cechy epigenomiczne. Finalnie, z użyciem oprogramowania do modelowania molekularnego, uzupełnionego o podejście do formowania pętli na podstawie obecności motywów w sekwencji genomowej współopracowane przez Doktoranta, został sporządzony model 3D cząsteczki. W porównaniu do wcześniejszych metod, nowość podejścia polegała na uwzględnieniu dynamicznego aspektu tworzenia struktury chromatyny i analizie całego genomu naraz.

Praca [P2] zawiera wyniki pracy Doktoranta nad nowym modelem sieci powiązań biologicznych. Standardowa sieć PPI oddziaływań funkcjonalnych pomiędzy białkami, która używana jest celem uwzględnienia informacji o przestrzennej interakcji białek w rozmaitych problemach bioinformatycznych, jest nieskierowana. Doktorant skupił się w swoich badaniach na skierowanej wersji takiego grafu, która ma na celu oddanie kierunku przepływu sygnałów w systemach biologicznych. Powstaje ona przez uwzględnienie, która cząsteczka z pary zainicjowała daną relację. Skorzystał w tym celu z danych pochodzących z eksperymentów testowania leków. Podszedł do przetwarzania takiej informacji kompleksowo, zestawiając pojedyncze sieci PPI dostępne w wielu różnych źródłach w jedną sieć konsensusową, zawierającą najbardziej uwiarygodnione połączenia. W testach wykazał, że sieć konsensusowa ułatwia poprawne rozpoznanie kierunku interakcji pomiędzy białkami, które z kolei, reprezentowane skierowaną siecią PPI, lepiej oddają rzeczywiste zależności w modelowanych systemach biologicznych.

Celem kolejnych prac [P3] i [P4] było zintegrowanie informacji zawartej w sieciach PPI, z uwzględnieniem zaproponowanego podejścia konsensusowego, z danymi pochodzącymi z eksperymentu ChIA-PET, z analizy wariantów strukturalnych i polimorfizmów pojedynczych nukleotydów w genomie oraz z innych źródeł. W sumie dają one bardziej kompleksowy obraz procesów zachodzących w organizmach żywych, w tym przypadku w organizmie człowieka. Doktorant wziął udział w tworzeniu takiej sieci powiązań zbudowanej na drodze badania struktur białek, chromatyny i sekwencji genomowej. To podejście wymagało opracowania nowych zasad umożliwiających integrację sieci, w których wierzchołkami są białka, z sieciami, w których wierzchołkami są motywy w sekwencji DNA. Uczyniono to na drodze powiązania miejsc w sekwencji DNA z genami, a następnie z białkami. Tak utworzoną sieć wykorzystano do zbadania wpływu zmienności genetycznej powiązanej z chorobami na przestrzenną strukturę chromatyny i wykrycia regionów ulegających zmianom funkcjonalnym.

W pracy [P5] również wykorzystane zostały dane różnego rodzaju, mianowicie sieci PPI, sekwencje białkowe i baza ontologii genów, tym razem do przewidywania funkcji białek. Graf powiązań białek uzupełniony został o wartości opisujące ich fizykochemiczne własności, a następnie formowano w nim klastry po ich podobieństwie. Białko poddawane analizie miało wstępnie przypisywaną funkcję na podstawie białek z najbliższego mu klastra. Finalny wynik wyprowadzany był na zasadzie konsensusu po uwzględnieniu wszystkich źródeł danych. Testy wykazały przewagę tego podejścia nad innymi, które pomijały niektóre z uwzględnionych danych, np. odwoływały się tylko do podobieństwa sekwencji białek z pominięciem ich fizykochemicznych własności. Ale nawet po pominięciu jednego lub dwóch źródeł danych w metodzie współopracowanej przez Doktoranta, przewyższała ona inne metody korzystające z tych samych okrojonych danych.

Formułując swoją ocenę rozprawy, pragnę na wstępie podkreślić dużą staranność Doktoranta w uwzględnieniu przetwarzanych zbiorów biologicznych jak najbardziej wiarygodnych, aktualnych i obszernych. Takie podejście pozwoliło nie tylko na rzetelne przetestowanie proponowanych modeli i metod, ale przede wszystkim na wyprowadzenie z badań wniosków istotnych dla środowiska biologicznego i medycznego. Potwierdzeniem tego są czasopisma, w których rezultaty tych badań zostały przyjęte.

Do najważniejszych osiągnięć badawczych przedstawionych w rozprawie zaliczyłabym:
– Opracowanie podejścia konsensusowego, wykorzystującego trzy metody składowe, do przewidywania funkcji białek. Na uwagę zasługuje m.in. algorytm wstępnego przetwarzania sieci PPI, który usuwa struktury potencjalnie niekorzystnie wpływające na dalsze wnioskowanie. Obszerne testy odwołujące się do danych o różnym zakresie potwierdziły przewagę nowego podejścia we wszystkich testowanych przypadkach.
– Opracowanie skierowanej sieci PPI na podstawie wielu źródeł i z użyciem algorytmów przetwarzania informacji pochodzącej z eksperymentów testowania leków.
– Uzupełnienie metody modelowania molekularnego 3D cząsteczek chromatyny o metodę formowania pętli chromatynowych na podstawie obecności motywów w sekwencji genomowej.

W trakcie czytania rozprawy nasunęły mi się następujące uwagi krytyczne. W [P1] zabrakło mi zbadania, jak przewidziana struktura przestrzenna chromatyny ma się do rzeczywistego kształtu cząsteczki. W [P3] nie jest dla mnie jasne, dlaczego informacja nt. ontologii genów i szlaków oddziaływań molekularnych wplatana jest w strukturę grafu w postaci osobnych wierzchołków i stowarzyszonych krawędzi, jak takie wierzchołki, o odmiennym znaczeniu niż pozostałe, są interpretowane przez algorytmy. Jednak największe moje wątpliwości wzbudziły wszystkie algorytmy zapisane w rozprawie w postaci pseudokodu. Błędy obecne w algorytmach wykraczają poza zwykłą nieuwagę przy ich zapisie i sprawiają, że niektóre fragmenty są niezrozumiałe.

– Algorytm 2.1: P_{row} jest uporządkowaną listą, a P_{to}, P_{from}, $P_{confidence}$ są nieuporządkowanymi zbiorami, jednak dostęp do ich elementów założony jest jakby występowały w tej samej kolejności w każdym ze zbiorów. Zamiast $G_{ppi}(Egde)$ powinno być E_p, zamiast $pair(x,y,w)$ powinno być $pair((x,y),w)$, zamiast *index* powinno być używane w tej pętli *i*, jak również *index* powinien pobierać kolejne wartości z jakiegoś zbioru lub zakresu, podczas gdy użyta tam funkcja *len*() zwraca pojedynczą liczbę. Zapis $P \leftarrow len(P_{unique})$ miał chyba inaczej wyglądać, gdyż *P* zgodnie z konwencją powinno być zbiorem albo listą, poza tym nie jest dalej używane. Pozostała część algorytmu jest całkowicie niezrozumiała, gdyż używana jest zmienna niewprowadzona *t*, zmienna *sym* nie jest dalej używana, niejasne jest znaczenie argumentów i działania funkcji *sps*() i zwracana jest zmienna P_sG wcześniej ani razu nieużyta. Literówki: *V* zamiast V_p, $len(_{from})$ zamiast $len(P_{from})$, wiersze czasem kończą się średnikiem, czasem nie, a raz dwukropkiem oraz jest o jedno słowo kluczowe *end* za dużo.

– Algorytm 2.2: Zostały użyte dwie zmienne *prior* i *Prior* tak, że trudno dociec, czy to ta sama zmienna, dodatkowo na wejściu jest inna wersja niż potem wykorzystana. Zmienna z wejścia $threshold_{convergence}$ nie została użyta, typy zmiennych w operacjach są pomieszane, dwa wiersze są albo sklejone razem, albo jest to jeden niewłaściwie zapisany wiersz.

– Algorytm 2.3: Użyta została niewprowadzona zmienna *Drug_Info*, nie zgadza się liczba parametrów funkcji *prior_Smooth*, prawdopodobnie w trzeciej i czwartej pętli *for* miało być V_{target} zamiast V_{source}. Tym razem nie wiadomo, co miało być podstawione pod wynikową sieć $P_{diffused}G$ zamiast użytej tu niewprowadzonej zmiennej $G_{ppi}(Egde)$, gdyż raczej nie zbiór krawędzi grafu. Tutaj $pair(x,y)$ powinna oznaczać parę (krawędź, waga), oznacza jednak raczej parę wierzchołków. Zapis $prior_{c\_u}$ miał prawdopodobnie oznaczać $prior_c[u]$ wg przyjętej wcześniej konwencji. Argumenty funkcji *score_edge* nie pasują do opisu działania tej funkcji na dole algorytmu, a wykorzystane tu zmienne $prior_c$ i $prior_e$ nie zostały wcześniej wprowadzone. Literówki: *V* zamiast V_p, ostatnia pętla *foreach* zaczyna się w środku wiersza, niepotrzebna definicja funkcji *len*(), która nie jest tu używana.

– Algorytm 2.4: Litera Φ miała chyba oznaczać symbol zbioru pustego, wtedy jednak błędnie jest przypisana zmiennej liczbowej w_p. Instrukcja warunkowa *if* jest niepotrzebna. Zmienna e_p oznacza pojedynczą krawędź, nie można więc do niej dodać ścieżki złożonej z wielu wierzchołków. Zbiór V_p, ustawiony początkowo na pusty, nie jest w algorytmie uzupełniany, więc zwracany graf P_sG również jest pusty.

Powyższe uwagi nie podważają mojej pozytywnej oceny pracy naukowej Doktoranta. Układ rozprawy i wykorzystanie źródeł literaturowych nie budzą zastrzeżeń.

3. Podsumowanie

Pan mgr Kaustav Sengupta w swojej pracy doktorskiej wykazał, na przykładzie rozwiązywanych problemów bioinformatycznych, że integracja wielu biologicznych źródeł danych sprzyja lepszemu zamodelowaniu świata rzeczywistego i znacznie poprawia uzyskiwane wyniki i wnioski w porównaniu do standardowo interpretowanych danych w tych problemach. Do rozwiązania problemów zastosował nowe modele grafowe i szereg metod statystycznych i algorytmicznych. Wykazał się zatem wiedzą i umiejętnościami z zakresu informatyki, ale także dodatkową głęboką znajomością problemów i literatury biologicznej. Jego osiągnięcia uznane zostały w środowisku naukowym wysoko punktowanymi publikacjami ujętymi w przedłożonym cyklu. Należy też wspomnieć o pozostałych artykułach Doktoranta, niewłączonych do recenzowanego cyklu publikacji. Jest on współautorem artykułów w czasopismach *Nucleic Acids Research* (200 pkt. MEiN), *Briefings in Functional Genomics* (140 pkt.) oraz w dwóch innych publikacjach spoza wykazu MEiN, a także czterech w niepunktowanych materiałach konferencyjnych.

Na podstawie wyrażonych powyżej opinii stwierdzam, że rozprawa pt. “The Meta-Network Model of the Complete Human Biomolecular Interactome at the Whole Cell Scale” autorstwa mgr. Kaustava Sengupty spełnia warunki stawiane rozprawom doktorskim przez obowiązującą ustawę o stopniach naukowych i tytule naukowym. Wnoszę o dopuszczenie tej rozprawy do publicznej obrony.