

Katedra Grafiki, Wizji Komputerowej i Systemów Cyfrowych
Politechniki Śląskiej

Recenzja rozprawy doktorskiej

Autor: Mgr Kaustav Sengupta

Tytuł: The meta-network model of the complete human biomolecular interactome at the whole cell scale

Promotor: Prof. dr hab. Dariusz Plewczyński

Prof. dr hab. Anna Gambin

Ogólna charakterystyka rozprawy

Przedłożona do recenzji rozprawa doktorska jest napisana w języku angielskim. Liczy 167 stron tekstu. Tematyka pracy bazuje na pięciu oryginalnych, wieloautorskich artykułach naukowych, których współautorem jest Doktorant. Artykuły te poświęcone są modelowaniu przestrzennej struktury chromatyny z wykorzystaniem teorii wybuchowej perkolacji, zastosowaniu symulacji procesu dyfuzyjnego do ustalania (oceny) kierunkowości w sieciach oddziaływań pomiędzy białkami, wykorzystaniu modelu sieci oddziaływań biomolekularnych do badania wpływu wariantów genomowych DNA na strukturę i funkcjonalność chromatyny, konstrukcji sieci oddziaływań genomowych chromatyny i jej związek/wpływ na sieci oddziaływań białkowych, wykorzystanie metodologii integracji struktury sieci oddziaływań białkowych oraz terminów ontologicznych do przewidywania funkcji białek. Zawartość pięciu publikacji będących podstawą recenzowanej pracy/rozprawy doktorskiej jest streszczona w jej rozdziale 2. W rozdziale 3 rozprawy przedstawione są dokładnie udziały Doktoranta w każdej z pięciu publikacji. Ponadto, w rozdziale 3 Doktorant zamieszcza informacje o swoim udziale w grantach, wykaz odbytych wizyt naukowych oraz listę publikacji, których jest współautorem, niewłączonych do prac będących podstawą doktoratu. W rozdziale 4 zatytułowanym „Summary” przedstawiony jest wykaz i charakterystyka wszystkich osiągnięć naukowych Doktoranta uzyskanych w jego pracy doktorskiej. Na końcu opracowania zamieszczone są kopie wszystkich pięciu artykułów wchodzących w skład pracy doktorskiej.

Artykuły naukowe, na których bazuje praca są interesujące, opublikowano je w wysoko punktowanych czasopismach naukowych, Computational and Structural Biotechnology Journal - 140pkt, IF=6.0, Seminars in Cell and Developmental Biology - 140 pkt, IF=7.3, Methods - 100pkt, IF=4.8, Frontiers in Genetics - 140pkt, IF=3.4.

W pracy P1, (Kaustav Sengupta, Michał Denkiewicz, Mateusz Chiliński, Teresa Szczepińska, Ayatullah Faruk Mollah, Sevastianos Korsak, Raissa D'Souza, Yijun Ruan, Dariusz Plewczynski, Multi-scale phase separation by explosive percolation with single-chromatin loop resolution, Computational and Structural Biotechnology Journal, Volume 20, 2022, Pages 3591-3603, ISSN-2001-0370) studiowano problem opisu procesu kompresji przestrzennej chromosomów wewnątrz jąder komórkowych. Proces ten od strony konstrukcji modeli matematycznych z jednej strony rozumiano jak proces przemiany fazowej a z drugiej jako proces (model) perkolacji. Przy takim podejściu uważa się, że zaszła przemiana fazowa z formy nieskompresowanej na formę skompresowaną chromosomu jeśli liczba gałęzi w grafie oddziaływań pomiędzy fragmentami DNA jest na tyle duża, że powstaje jeden spójny graf połączeń. Krawędzie w grafie oddziaływań modelującym proces kompresji DNA są pętle chromatynowe. W procesie modelowania kompresji DNA stopniowo dodawane są nowe krawędzie do grafu oddziaływań pomiędzy fragmentami DNA i jednocześnie obserwowana i charakteryzowana jest liczba rozłącznych podgrafów. Autorzy pracy P1 nawiązują także do (hipotetycznego) dynamicznego procesu zwijania polimeru DNA w formę skompresowaną, realizowanego przez silniki molekularne, nazywanego mechanizmem wyrzucania pętli (loop extrusion), studiowanego obecnie w wielu pracach w literaturze. Rozpatrują kilka metod (scenariuszy) dodawania połączeń do polimeru DNA, metody topologiczne, metody w których pętle chromatynowe opisywane są dodatkowo parametrem częstości, metody w których występują kompartmenty w strukturze chromatyny, metody w których pętlom przypisuje się zmienne skalarne używane do charakteryzacji trajektorii procesu perkolacji oraz metody w których do charakteryzacji procesu perkolacji używa się modelu dwuparametrowego – jeden parametr dotyczy pętli chromatynowej a drugi kotwic reprezentujących węzły grafu perkolacji. Swoje modele matematyczne odnoszą do danych eksperymentalnych, wykorzystują publicznie dostępne wyniki eksperymentów prowadzących do odkrywania miejsc oddziaływań pomiędzy fragmentami DNA (eksperymenty sekwencjonowania ChIA-PET, Chromatin Interaction Analysis with Paired-End Tag).

W pracy P2, (Sengupta, K., Gambin, A., Basu, S., & Plewczynski, D. (2022, June) MultiNet: A Diffusion-Based Approach to Assign Directionality in Protein Interactions Using a Consensus of Eight Protein Interaction Datasets. In Proceedings of International Conference on Frontiers in Computing and Systems: COMSYS 2021 (pp. 13-20). Singapore: Springer Nature Singapore) autorzy podejmują problem oceny kierunku oddziaływań pomiędzy białkami w dużych sieciach oddziaływań białko-białko. Jest to interesujący problem oraz kierunek badań dlatego, że obecnie istnieje wiele dostępnych danych, w postaci bardzo wysoko wymiarowych sieci oddziaływań pomiędzy białkami, w których nie są znane kierunki

tych oddziaływań. Istnieją metody eksperymentalne ustalania kierunkowości oddziaływań, jednak ich możliwości, w obliczu wielkich wymiarowości sieci oddziaływań, są niewystarczające. Dlatego metody bioinformatyczne, bazujące na modelowaniu matematycznym są bardzo atrakcyjną i interesującą alternatywą. We wcześniejszych publikacjach, w literaturze zaproponowano formułowanie problem oceny kierunkowości oddziaływań pomiędzy białkami jako problemu orientacji grafu. Problem orientacji grafu jest formułowany tak, aby jego rozwiązanie było użyteczne dla oceny kierunkowości badanych oddziaływań białko - białko. W sieci oddziaływań część kierunków jest znana dzięki posiadanym danym – głównie wynikom badań eksperymentalnych. Z kolei problem orientacji grafu jest problemem optymalizacji kombinatorycznej, o którym wiadomo, że należy do grupy problemów NP-trudnych. W literaturze, w której badano ten problem jako model dla oceny kierunkowości oddziaływań białko – białko zaproponowano heurystyki do jego rozwiązywania bazujące na symulacji sieci dyfuzyjnych. W sieciach tych ocenia się intensywność oddziaływania pomiędzy parami węzłów, z wykorzystaniem iteracyjnego algorytmu modyfikacji siły /intensywności tych oddziaływań. Otrzymane ustalone rozwiązanie wykorzystywane jest do oceny kierunków oddziaływań białko – białko. Autorzy publikacji P2 wykorzystują to podejście literaturowe, dla konstrukcji sieci oddziaływań i symulacji intensywności dyfuzyjnych wykorzystują bardzo szerokie bazy danych literaturowych (por. tabela 1 w pracy P2). Wykazują także efektywność uzyskanego podejścia. Prezentują także opracowane w ramach tego podejścia narzędzie bioinformatyczne, MultiNet.

W pracy P3 (Chiliński, M., Sengupta, K., & Plewczynski, D. (2022, January). 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. In *Seminars in Cell & Developmental Biology* (Vol. 121, pp. 171-185) Academic Press) autorzy wprowadzają koncepcję sieci oddziaływań biomolekularnych (biomolecular interaction network, BIN) którą nazywają meta – siecią. Jest to model możliwie jak najbardziej kompletnej sieci oddziaływań występujących w komórkach, który otrzymuje się przez integrację różnych typów oddziaływań komórkowych. Oddziaływania zintegrowane w stworzonej przez autorów meta – sieci obejmują oddziaływania DNA – białka, oddziaływania pomiędzy białkami i oddziaływania chromatyny. Dodatkowo struktura sieci BIN została zintegrowana ze strukturami oddziaływań wynikającymi z danych genowych sieci sygnałowych oraz metabolomicznych, a także z danymi z drzewa ontologii genowych. Wykorzystując stworzoną przez siebie meta – sieć autorzy potrafili odkrywać istotne biologiczne oraz fizjologiczne zależności. Pokazali istnienie statystycznie istotnych związków pomiędzy wariantami strukturalnymi DNA w 6 chromosomie w lokacjach związanych z kompleksem HLA, chr6:32014923-33217929, z chorobami autoimmunologicznymi. Odkryli także związek modyfikacji wariantów strukturalnych na chromosomie 14, w lokacjach chr14:35138733-36158360, z ekspresją czynników transkrypcyjnych.

W pracy P4 (Halder, A. K., Denkiewicz, M., Sengupta, K., Basu, S., & Plewczynski, D. (2020). Aggregated network centrality shows non-random structure of genomic and proteomic networks. *Methods*, 181, 5-14) autorzy wykorzystują metodologie i dane ChIA-PET (analizę interakcji chromatyny za pomocą danych sekwencjonowania znaczników ze sparowanymi końcami) z ludzkiego genomu w celu skonstruowania zestawu sieci interakcji genomowych. Ich model uwzględnia podział chromatyny na domeny kontaktowe z chromatyną (CCD, chromatin contact domains). Sieci genomowe następnie odwzorowują na sieci interakcji proteomicznych. W efekcie uzyskują informacje o interakcjach białko-białko (PPI, protein protein interactions). Analizują podsieci oddziaływań białko – białko, wykorzystując sieciowe cechy topologiczne. Wykazują statystycznie istotne różnice pomiędzy stworzonymi/odkrytymi podsieciami oddziaływań białko – białko, a sieciami budowanymi losowo. Jako narzędzia statystycznego używają testu t budowanego na wartościach cech topologicznych porównywanych sieci (tabela 1). Wykorzystując swoją metodologię identyfikują region na chromosomie 6, chr6:32014923-33217929 (ten sam, który był wskazywany w publikacji P3), który posiada dużą (ponad przeciętną) koncentrację strukturalnych wariantów genomowych związanych z chorobami autoimmunologicznymi.

W pracy P5 (Sengupta, K., Saha, S., Halder, A. K., Chatterjee, P., Nasipuri, M., Basu, S., & Plewczynski, D. (2022). PFP-GO: Integrating protein sequence, domain and protein-protein interaction information for protein function prediction using ranked GO terms. *Frontiers in Genetics*, 13, 969915) autorzy przedstawiają metodologię integracji struktury sieci oddziaływań białkowych oraz terminów ontologicznych z zastosowaniem do przewidywania funkcji białek. Rezultatem zaprezentowanym w tej publikacji jest także narzędzie obliczeniowe, PFP-GO do przewidywania funkcji białek. W pracy dokonuje się omówienia znanych obecnie podejść do przewidywania funkcji białek, w tym podejść bazujących na sekwencjach aminokwasów w białkach, bazujących na domenach białkowych, bazujących na znanych sieciach oddziaływań białko – białko oraz podejść łączących / integrujących różne informacje. Na tym tle opisana jest metodologia opracowana w publikacji. Aplikacja PFP-GO używa zasobów z baz danych String, Uniprot, Pfam, Domine oraz z bazy danych ontologii genowych. W pracy przedstawia się metodologię integracji oraz filtrowania informacji uzyskanych z połączenia baz danych, prowadzących do hipotez dotyczących funkcjonalności analizowanych białek. Idea funkcjonowania aplikacji PFP-GO przedstawiona jest graficznie na szeregu rysunkach/schematach (rysunek 1, rysunek 2, rysunek 3, rysunek 4, rysunek 5, rysunek 6). Formuła filtracji dla gałęzi tworzonego grafu białek wykorzystuje ideę opisaną wcześniej w literaturze. Autorzy przedstawiają porównania efektywności swojego narzędzia z innymi narzędziami publikowanymi w literaturze (tabela 2, tabela 3, tabela 4) które wykazują przewagę PFP-GO nad innymi metodami.

Ocena rozprawy

Moja ocena recenzowanej pracy doktorskiej jest zdecydowanie pozytywna. Recenzowana praca jest bardzo ściśle związana z oryginalnymi publikacjami, wydanymi w renomowanych czasopismach naukowych.

Doktorant dokumentuje opanowanie warsztatu naukowego obejmującego modelowanie matematyczne procesów komórkowych, techniki informatyczne, metody sztucznej inteligencji oraz uczenia maszynowego, konstrukcję i implementację potoków analiz bioinformatycznych i elementy wiedzy biologicznej. Praca bazuje na bardzo współczesnej wiedzy dotyczącej procesów molekularnych, biologicznych, komórkowych. Wyniki zastosowania nowych technik eksperymentalnych prowadzą do uzyskiwania wiedzy na temat tych procesów. Ich zrozumienie, proponowanie modeli matematycznych, które je tłumaczą jest obecnie bardzo intensywnie rozwijanym kierunkiem badań naukowych. Doktorant jako członek grup badawczych potrafił włączyć się w ten nurt badań i uzyskać oryginalne wyniki naukowe. Wyniki artykułów zamieszczonych w pracy dokumentują posiadanie przez Doktoranta kompetencji w zakresie modelowania matematycznego, technik informatycznych, programowania, konstrukcji algorytmów, a także wiedzy biologicznej. Doktorant wykazuje się zaangażowaniem, pomysłowością i motywacją do pracy naukowej.

Praca bazuje na publikacjach współautorskich. W trzech spośród pięciu prac, Doktorant jest pierwszym autorem. Ponadto w pracy zamieszczono oświadczenia współautorskie, a także, dodatkowo bardzo dokładne wyjaśnienia udziału Doktoranta w każdej z publikacji. Na podstawie tych informacji można stwierdzić istotny udział Doktoranta w powstaniu cyklu przedłożonych do recenzji publikacji.

Obok cyklu publikacji, który Doktorant przedłożył do recenzji, Doktorant posiada także wiele dodatkowych, wartościowych osiągnięć, publikacji naukowych, udział w grantach, wizyty naukowe.

Uwagi dyskusyjne i krytyczne

W publikacjach P1, P2, P3, P4, P5 zrealizowano oryginalny plan badań naukowych, sformułowano i opublikowano oryginalne wnioski. Tematyka tego cyklu prac jest interesująca i szeroka. Jednak opisy cyklu prac wchodzących w skład doktoratu, zamieszczone w rozdziale 1 i 2 pozostawiają pewien niedosyt. Byłoby bardzo interesujące, gdyby w tych opisach zamieścić jakieś uzupełnienia, dodatkowe informacje, szczegóły, uzupełnienia do prac P1, P2, P3, P4, P5. Zwłaszcza dotyczy to aspektów metodologicznych. Przy czytaniu tych prac nasuwa się wiele pytań dotyczących szczegółów konstrukcji modeli matematycznych, ich implementacji komputerowych, porównań z innymi wynikami. Można by także uwypuklić związki pomiędzy publikacjami cyklu, rozwój metodologii badań

naukowych w tych pracach. Jednak rozdziały 1 i 2, doktoratu/opracowania są tylko streszczeniem/skrótem cyklu publikacji. Nie ma w nich raczej niestety dodatkowych szczegółów.

Praca P1 ma w swoich suplementach filmy / animacje pokazujące wyniki symulacji procesu zwijania się chromatyny w różnych skalach. Widać, że uzyskano możliwości obliczeniowe pozwalające na wyznaczanie dowolnych wskaźników dynamiki, kształtu czy też stopnia kompresji chromatyny. Z drugiej strony autorzy cytują pracę (Banigan, E. J., & Mirny, L. A. (2019). Limits of chromosome compaction by loop-extruding motors. *Physical Review X*, 9(3), 031007), w której przedstawia się modele matematyczne (w postaci układów równań różniczkowych) oraz wyniki symulacji prowadzące do ocen stopnia kompresji chromatyny. Czy autorzy porównywali swoje wyniki z powyżej cytowaną pracą, lub czy ew. planują to zrobić?

Zarówno praca P3 jak i P4 wskazuje region chromosomu 6, chr6:32014923-33217929 który posiada dużą (ponad przeciętną) koncentrację strukturalnych wariantów genomowych związanych z chorobami autoimmunologicznymi. Czy to oznacza, że ten sam region / zakres został odkryty dwoma różnymi metodami? Jaką metodą (jakimi metodami) szacowano długość tych regionów w chromosomach?

Konkluzja

Uwagi dyskusyjne i krytyczne nie podważają oryginalnych osiągnięć uzyskanych w pracy.

Praca stanowi podsumowanie oryginalnych publikacji naukowych, w których formułowano i weryfikowano oryginalne hipotezy badawcze. Osiągnięcia i oryginalne elementy rozprawy są na pewno wystarczające do jej ogólnej pozytywnej oceny. Stwierdzam, że rozprawa spełnia warunki stawiane pracom doktorskim i wnioskuję o jej dopuszczenie do publicznej obrony.

