UNIWERSYTET WARSZAWSKI

Instytut Informatyki
Uniwersytet Warszawski
ul. Banacha 2
02–097 Warszawa
POLSKA

dr hab. Bartosz Wilczyński
profesor uczelni
Phone: +(48 22) 5544 577
Fax: +(48 22) 5544 400
e-mail: bartek@mimuw.edu.pl

Warszawa, 4. października 2024 r.

Recenzja rozprawy doktorskiej pt. „Modelowanie komputerowe struktury trójwymiarowej chromatyny na podstawie danych genomicznych i mikroskopii wysokorozdzielczej" przedstawionej przez mgr Zofię Tojek

Recenzja niniejsza została sporządzona na zlecenie Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej zgodnie z wymogami ustawy dotyczącej procedur nadawania stopnia doktora. Recenzja ta składa się najpierw ze skrótowego opisu merytorycznej zawartości rozprawy, następnie zawiera moje uwagi krytyczne dotyczące cech rozprawy, które stanowią pewne usterki czy niedociągnięcia i zakończona jest podsumowaniem.

Opis rozprawy

Praca mgr Tojek złożona jest nominalnie z 8 rozdziałów, przy czym pierwsze 4 stanowią moim zdaniem faktyczną rozprawę, podczas gdy rozdziały 5-8 stanowią różnego rodzaju dodatki, typowo umieszczone na końcu rozprawy , tj. deklaracje wkładu autorki, spis osiągnięć naukowych czy też dwa manuskrypty, których współautorką jest autorka rozprawy, umieszczone w rozdziale 8. *verbatim*. Osobiście wolałbym, aby (może poza rozdziałem 7., czyli spisem literatury) ta część pracy stanowiła raczej dodatki, niż rozdziały pracy, ale nie zmienia to zasadniczego kształtu rozprawy.

Rozdział 1. poświęcony jest wprowadzeniu pojęć potrzebnych w dalszej części rozprawy. Jako że zakres tematyczny rezprawy jest bardzo szeroki i dotyczący wielu zagadnień z wielu dziedzin nauki, autorka stała tu przed dość trudnym zadaniem. Opisywane w tej części pojęcia dotyczą zarówno informatyki (tj. problem komiwojażera) jak i zagadnień matematycznych (tj. interpolacja funkcjami sklejanymi - splajnami), statystycznych (tj. analiza składowych głównych) ale też wiele tematów związanych z biologią molekularną

(tj. podstawy biologii komórkowej i funkcji chromatyny). Mimo tego, że rozdział ten liczy niemal 30 stron, nieuchronnie pozostawia niedosyt ze względu na to, że każde opisywane pojęcie jest potraktowane dość zdawkowo. Autorka nie ustrzegła się w tej części rozprawy różnych - typowych dla rozpraw doktorskich - usterek stylistycznych czy narracyjnych, które przypisywałbym tu głównie szczególnie trudnej materii związanej z interdyscyplinarnością projektu, niemniej pewne decyzje autorki dają do myślenia. Wydaje mi się, że decyzja, aby część wstępu dotycząca „eksperymentalnej" części pracy jest dłuższa niż część dotycząca metod obliczeniowych, które stanowią rdzeń wyników. W mojej ocenie, część wprowadzająca nie jest najmocniejszą stroną rozprawy i sugeruje, że autorka - być może w obliczu zbyt dużego nagromadzenia różnorodnych technik wykorzystywanych w pracy - nie do końca była w stanie dobrze zredagować wprowadzenia do metodologii.

Rozdział 2. jest, w moim odczuciu, zdecydowanie najważniejszym elementem rozprawy. Autorka opisuje tu narzędzie ChromoLooping oraz jego zastosowanie do wyników analizy jednej z pętli chromatynowych na podstawie obrazów uzyskanych techniką iPALM. Dodatkowo, opisuje adaptację tej techniki do danych EMISH. Duża część tych wyników jest opublikowana w postaci pracy w Scientific Reports, której mgr Tojek jest jedną z głównych autorów. Praca ta jest, moim zdaniem, głównym osiągnięciem autorki, które bez wątpienia zawiera jej duży wkład i jednocześnie jest już opublikowane w uznanym czasopiśmie naukowym. Warto zauważyć, że implementacja narzędzia ChromoLooping jest nieodzowną częścią wzmiankowanej pracy i niewątpliwie niemal w całości jest osiągnięciem autorki (w repozutorium github jedyną autorką kodu jest mgr Tojek). Część rozdziału poświęcona adaptacji metody ChromoLooping do obrazów z mikroskopu elektronowego (EMISH) nie są już tak jednoznaczne w sensie osiągnięć naukowych (nie jest dla mnie jasne, czy te wyniki są już gdzieś opublikowane, czy też nie), ale tutaj także można zauważyć niewątpliwie główny wkład autorki w implementację metody densitydots, która jest dostępna w repozytorium github.

Rozdział 3. przedstawia głównie wyniki zawarte w manuskrypcie (o ile mi wiadomo jeszcze nie recenzowanym, a jedynie umieszczonym na serwerze biorxiv) pt: „Improved cohesin HiChIP protocol and bioinformatic analysis for robust detection of chromatin loops and stripes". Mgr Tojek jest jedną z trojga głównych autorów tego manuskryptu. Praca ta zawiera zarówno elementy eksperymentalne (związane z poprawioną wersją protokołu cohesin HiChIP), jak i nowe narzędzia bioinformatyczne (nf-hichip) oraz zastosowania tego pakietu do analiz nowo powstałego zbioru danych hiChIP. Wg informacji zawartych w tym manuskrypcie, za przygotowanie narzędzia nf-hiChIP i wykonanie analiz odpowiedzialna była mgr Tojek i Abhishek Agarwal. W samym doktoracie w rozdziale 5. autorka próbuje uściślić swój wkład w tę część pracy, jednak nie jest w pełni czytelne, gdyż miejscami nie jest jasne co autorka ma na myśli. Po zapoznaniu się z repozytoriami kodu źródłowego, do których autorka podaje w pracy odnośniki, wszystko wskazuje na to, że autorka napisała pierwszą wersję tego pakietu, nazwaną luigi_seq, natomiast manuskrypt wykorzystują nowszą implementację nf-hichip-pipeline, w której tworzeniu brało udział dwóch innych współautorów wspomnianej wyżej pracy. Główną różnicą wydaje się byc wykorzystanie

pakietu nextflow zamiast Luigi, jednak bardzo trudno jest mi to jednoznacznie ocenić, a szkoda, że autorka nie przeprowadza głębszej analizy porównawczej tych pakietów.

Wnosząc z opisu w części 5. pracy, autorka jest też odpowiedzialna za dużą część analiz, które zostały dokonane w tym manuskrypcie (m. in. korelacje między próbkami, wizualizacje macierzy interakcji itp.), jednak wg opisu "authors' contribution" w artykule, jest to praca wspólna z innymi współautorami. Mimo tych, dość istotnych, problemów z opisem wkładu autorki, trzeba przyznać, że rzeczywiście narzędzia przygotowane przez zespół, którego niewątpliwie była częścią, sprawiają wrażenie dość przekrojowego zestawu narzędzi do analizy danych hiChIP. Przygotowywanie tego rodzaju pakietów jest bardzo skomplikowane i często wymaga pracy zespołowej, co czyni ją trudną do oceny w postępowaniach o nadanie stopnia doktora. Nie powinno to, moim zdaniem, nadmiernie negatywnie wpływać na ocenę pracy mgr Tojek, choć wolałbym, aby lepiej opisała swój wkład w te zespołowe prace.

Kolejny, 4. rozdział, który zawiera wnioski i kierunki dalszych badań jest niezwykle krótki - zajmuje około półtorej strony. Moim zdaniem jest to zdecydowany problem tej pracy, gdyż można byłoby wywnioskować, że autorka nie widzi zbyt wielu wniosków płynących z jej pracy. Biorąc pod uwagę, że duża część wniosków jest jednak zawarta w podsumowaniach rozdziałów 3 i 4, skłaniałbym się do opinii, że jest to raczej niedociągnięcie redakcyjne, niż faktyczny brak świadomości ze strony autorki, nt. wniosków płynących z jej pracy. Tak jak pisałem wcześniej, dalsze części pracy (rozdziały 5-8) stanowią raczej dodatki i nie wnoszą wiele nowego do rozprawy.

Ogólnie, uważam, że praca mgr Tojek zawiera dużo wartościowych wyników, zwłaszcza opisanych w rozdziałach 3 i 4 oraz w napisanym przez nią w trakcie pracy nad doktoratem kodzie. Niestety, na poziomie redakcyjnym, jest wiele mniejszych lub większych niedociągnięć, które mają niestety wpływ na moją ocenę całości pracy, choć nie zmniejszają mojego przekonania, że praca zasadniczo spełnia wymogi stawiane pracom doktorskim.

Uwagi krytyczne

Moje uwagi krytyczne dotyczą głównie dwóch aspektów: kwestii redakcyjnych pracy i niezbyt precyzyjnego opisu wkładu autorki w narzędzie nf-hichip. Dodatkowo mam dwie wątpliwości związaną z rozdziałem 2 dotyczącym zrekonstruowanych struktur chromatyny.

Jeśli chodzi o uwagi redakcyjne, to uważam, że praca ta, niestety, nie jest zredagowana zbyt dobrze. Autorka zamieszcza dwa manuskrypty (jeden opublikowany, jeden jeszcze nie) na końcu pracy, jakby praca miała opierać się na tych wieloautorskich tekstach, jednak jednocześnie powtarza dużą część wyników w rozdziałach 2 i 3. SPrawia to wrażenie, jakby autorka nie mogła się zdecydować, czy praca jest "złożeniem" artykułów, czy też jednak ma naturę monograficznej rozprawy. Ma to zapewne też wpływ, na fakt, że materia wstępna jest bardzo pobieżna - niemniej część materiału wstępnego jest zawarta w dalszych rozdziałach, pochodzących z artykułów, gdzie wyjaśniona jest duża część niezbędnej

w publikacji metodologii. Nie ułatwia to lektury tej, i tak bardzo skomplikowanej, pracy. Praca zawiera też sporo drobnych błędów stylistcznych (np. “szybkim czasie”, “pracy doktoranckiej”, “podstawowy podział metody” itp.), które w większości nie przeszkadzają w lekturze.

Jeśli chodzi o tę drugą kwestię, to oczywiście zrozumiałym jest, że w tego typu badaniach interdyscyplinarnych, bardzo często nie jest łatwo dokładnie przypisać wkładu autorskiego dla każdego z autorów. Jest to jeden z powodów, dla którego tak trudno czasami docenić pracę bioinformatyków przy nadawaniu stopniu doktorskich. Niezależnie od obiektywnych trudności związanych z określeniem wkłądu autorki, uważam, że w przypadku opisu jej roli w projektowaniu i implementacji pakietów Luigi-Seq i nf-hichip nie zostało nakreślone wystarczająco czytelnie.

Merytorycznie, opis wyników w pracy i metod opracowanych przez doktorantkę jest zasadniczo dobry i nie pozostawia wątpliwości co zostało zrobione (nawet jeżeli nie do końca jest jasne kto to zrobił). Jest jednak kilka miejsc, których nie jestem w stanie w pełni zrozumieć.

Drobniejsza wątpliwość, być może związana z lietrówką dotyczy pętli chromatynowej rozważanej na stronie 42. Nie jest jasne, czy pętla ta ma 33kb, czy 23kb. Sądząc z liczby sond, które są używane do modelowania, wskazywałoby to na 33kb, jednak autorka określa jej długość jako 23kbp.

Istotniejsza wątpliwość dotyczy kwestii porównywania zrekonstruowanych pętli. Na tyle, na ile rozumiem metodę chromolooper, nie znamy oryginalnej kolejności sond na podstawie obrazu mikroskopowego - stąd konieczność wykorzystania metody TSP. Jednak konsekwencją tego faktu, wg mnie, jest to, że zwróconą przez algorytm krzywą przestrzenną możemy “przebiegać” w dwie różne strony. Jednak w opisie porównań zrekonstruowanych struktur nie dopatrzyłem się opisu rozważania tego problemu. Wydaje się, że możliwe są różne sposoby (choćby maksimum lub średnia z obu wersji orientacji), ale nie wiem jak autorka ostatecznie rozwiązuje ten problem. Bardzo chętnie usłyszę o tym podczas obrony.

Podsumowanie

Podsumowując, praca mgr Zofii Tojek podejmuje istotne, nowe problemy w dziedzinie bioinformatyki. Autorka niewątpliwie włożyła istotny wkład w stworzenie dwóch nowych narzędzi bioinformatycznych (ChromoLooper i Luigi-Seq) oraz wykazała się umiejętnością zastosowania tych narzędzi w praktyce do realnej analizy danych pochądzących z różnego rodzaju eksperymentów mających na celu zbadanie struktury chromatyny. Przedstawiona praca została częściowo opublikowana (część dotycząca narzędzia chromLooper) w czasopiśmie Scientific Reports, a częściowo upubliczniona w postaci preprintu w systemie biorxiv (część dotycząca analiz hiChip). Zawiera także nieopublikowane jeszcze analizy z wykorzystaniem pakietu nf-hichip. Mimo pewnych niedociągnięć i usterek, o

których pisałem wcześniej, praca stanowi niewątpliwie wkład do dyscypliny informatyka w dziedzinie nauk technicznych i świadczy o tym, że autorka uzyskała poziom wiedzy i samodzielności naukowej oczekiwany na etapie doktoratu. W związku z tym uważam, że **rozprawa doktorska spełnia ustawowe wymagania wobec prac doktorskich** i może zostać skierowana do kolejnych etapów przewodu doktorskiego.

Z poważaniem,

Bartosz Wilczyński