CENTRUM NOWYCH TECHNOLOGII

UNIWERSYTET WARSZAWSKI

ul. S. Banacha 2c, 02-097 Warszawa
TEL.: + 48 22 55 43 600, **FAX:** +48 22 55 43 606,
E-MAIL: d.plewczynski@cent.uw.edu.pl
www: **http://www**.cent.uw.edu.pl

Warszawa, 6.12.2020

Warszawa, 6/12/2020

Prof. dr hab. Dariusz Plewczyński
Laboratorium Genomiki Funkcjonalnej i Strukturalnej
Centrum Nowych Technologii
Uniwersytet Warszawski
ul. Banacha 2c, 02-097 Warszawa, Polska

RECENZJA

rozprawy doktorskiej magistra Rafała Zaborowskiego

COMPUTATIONAL METHODS FOR DIFFERENTIAL ANALYSIS OF CHROMATIN CONTACT MATRICES

wykonanej w Instytucie Informatyki

Wydziału Matematyki, Informatyki i Mechaniki

Uniwersytetu Warszawskiego

pod kierunkiem promotora

dr hab. Bartosza Wilczyńskiego

Przedstawiona mi do recenzji praca jest owocem udanej analizy teoretycznej w paradygmacie genomiki obliczeniowej. W pracy udało się twórczo połączyć nowatorską metodologię informatyczną z istotnym biologicznie problemem badawczym skupiającym się na głębszym poznaniu struktury przestrzennej genomu ludzkiego. Genomika obliczeniowa, a w szczególności genomika trójwymiarowa to unikalna, wysoce interdyscyplinarna dziedzina badawcza w której w celu uzyskania opisu zjawiska biologicznego musimy użyć zaawansowanych metod obliczeniowych do odrzucenia błędnych odczytów wynikających z

wysokiego poziomu szumu doświadczalnego. Tylko dzięki zaproponowaniu złożonych modeli statystycznych budowanych na podstawie rozległych (pełno-genomowych) danych NGS możemy zidentyfikować istotne cechy opisywanego zjawiska. W szczególności praca w ciekawy sposób poddaje krytyce paradygmat naukowy obecny od roku 2012 (publikacja Dixon et al.), ugruntowany w 2015 roku kolejną publikacją (Rudan et al.). Autorowi udało się sfalsyfikować powtarzaną w wielu aktualnie cytowanych publikacjach tezę o podobieństwie struktury domenowej między typami komórek, tkankami, organizmami czy całymi gatunkami (np. człowiek vs. mysz). Uważam to za najistotniejsze dokonanie pracy, razem z solidnie przemyślanym i ugruntowanym formalizmem matematycznym.

Przedmiotem mojej oceny, w myśl wymagań Ustawy o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki z dnia 14 marca 2003 r. (Dz.U. 2017 poz. 1789, z późn. zm.) oraz Rozporządzenia Ministra Nauki i Szkolnictwa Wyższego z dnia 19 stycznia 2018 r. w sprawie szczegółowego trybu i warunków przeprowadzania czynności w przewodzie doktorskim, w postępowaniu habilitacyjnym oraz w postępowaniu o nadanie tytułu profesora (Dz.U. 2018 poz. 261), jest oryginalność rozwiązanego problemu naukowego, ogólna wiedza teoretyczna Kandydata w dziedzinie informatyki, a także umiejętność samodzielnego prowadzenia pracy naukowej.

Rozprawa doktorska Pana mgr Rafała Zaborowskiego została przygotowana w Laboratorium dr hab. Bartosza Wilczyńskiego na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego w Warszawie pod kierownictwem kierownika laboratorium. Autor skupia się na sformułowaniu nowej teorii matematycznej (czy raczej statystycznej) związanej z porównywaniem macierzy interakcji między różnymi replikatami w ramach pojedynczego doświadczenia biologicznego, między różnymi stanami komórki (np. zmieniającej się pod wpływem stresu), różnymi typami komórek, czy wreszcie różnymi gatunkami. Praca proponuje nową metodę, jednocześnie odnosząc się do dotychczasowo używanych algorytmów i narzędzi, w moim przekonaniu adresując bardzo ważny problem badawczy. Motywacją pracy jest potrzeba środowiska badaczy trójwymiarowej genomiki związana ze statystycznie istotnymi porównywaniami różnych segmentacji chromatyny na domeny. Wynika to z bardzo złożonego problemu, dużego poziomu szumu w danych doświadczalnych jak i samą iluzorycznością domen genomicznych. Niektórzy badacze uważają

nawet, że domeny TAD nie są realne, a wynikają tylko ze statystycznej agregacji danych w doświadczeniach 3C typu populacyjnego. Jest to ciekawy argument, który wymaga dalszej pracy. Zaproponowane przez Autora narzędzia wpisują się w ten nowy nurt badawczy umożliwiając rygorystyczne i statystycznie istotne porównywanie różnych metod dzielenia nici DNA na bliskie przestrzennie regiony.

Autor w jednostronicowym streszczeniu pracy przedstawia główne tezy rozprawy. Podkreśla znaczenie opracowanych algorytmów, choć połączenie między ekspresją i strukturą 3D nici chromatynowej jest dość pobieżnie naszkicowane. Opisana jest po pierwsze miara BP do porównywania macierzy kontaktów Hi-C, oraz algorytm DiADeM służący do wykrywania statystycznie unikalnych oddziaływań dla każdego z porównywanych dwóch zbiorów danych Hi-C.

W rozdziale pierwszym autor skupia się na wprowadzeniu podstawowych pojęć biologicznych używanych w pracy. Robi to sprawnie i bez przesadnie rozbudowanego opisu, mimo wszystko podając kluczowe referencje do literatury przedmiotu.

W rozdziale drugim autor referuje metody komputerowe używane w analizie danych z sekwencjonowania następnej generacji (NGS). Tą część pracy można potraktować jako bardziej matematyczny wstęp do problematyki rozprawy, który pozwolił autorowi bardziej przejrzyście przedstawić wyzwania związane z analizą danych genomicznych (jednowymiarowych) jak i zaprezentować modele statystyczne ekspresji genów. Pokazuje to erudycję autora, choć nie ma to bezpośredniego związku z dokonaniami opisanymi w kolejnych rozdziałach. Być może początkową intencją pracy było powiązanie struktury trójwymiarowej z informację o ekspresji, jeśli tak było w istocie to zabrakło ostatniego rozdziału który łączyłby formalizm ekspresji genów ze strukturą 3D genomu ludzkiego.

Trzeci rozdział poświęcony jest nadal w formie wprowadzenia (wstępu) zagadnieniu porównywania macierzy kontaktów w eksperymentach Hi-C. Zaczyna się od opisania metod normalizacji danych doświadczalnych – co recenzentowi wydaje się absolutnie kluczowym i często robionym zbyt pospiesznie etapem prac genomiki 3D. Następnie Autor referuje wieloskalową strukturę chromatyny, zaczynając od homopolimerowego zaniku sygnału

strukturalnego wraz z oddalaniem się od diagonali macierzy kontaktów. Następnie opisany jest poziom kompartmentalizacji chromatyny (podziału na aktywną i pasywną), później poziom domen genomicznych (TADów) a na końcu wysoko-rozdzielczy poziom pętli chromatynowych. Dalej zamieszczony jest wstęp do zagadnienia porównywania macierzy kontaktów w powyżej wymienionych skalach przestrzennych.

W kolejnym rozdziale Autor referuje istniejące metody porównywania domen genomicznych, prezentuje solidnie przygotowany i sformalizowany opis swojej nowej metody BP Score, oraz porównuje swoje osiągnięcie z literaturą i dostępnymi narzędziami. Miara BP zapewnia nową metrykę do porównywania segmentacji chromatyny, szczególnie przydatną do porównywania różnych stanów komórki (np. z oraz bez szoku termicznego), oceny jakości replikatów w doświadczeniach 3C, porównywania różnych typów komórek, oraz na koniec dla różnicowego porównania różnych algorytmów identyfikujących TADy.

W ostatnim rozdziale Autor najpierw opisuje różne metody identyfikacji istotnych statystycznie różnic między macierzami kontaktów (metody diffHiC, FIND, HiCcompare, SELFISH), a następnie wprowadza rygorystycznie nową metodę DiADeM, która następnie benchmarkuje na danych syntetycznych i rzeczywistych. Metoda ta ma służyć wg. zamierzeń autora identyfikacji oddziaływań dalekiego zasięgu, którymi mają się różnić macierze Hi-C dla np. dwóch stanów komórkowych. Autor adresuje dwa postawowe problemy spotykane w tego typu zagadnieniach: (i) zanik siły sygnału przy oddalaniu się od diagonali, oraz (ii) wpływ obciążeń związanych z istotnymi różnicami w pokryciu odczytami krótkich fragmentów w trakcie sekwencjonowania DNA.

Podsumowując – autor wykazał się znajomością genomiki, zarówno od strony zagadnień biologicznych, ale również bioinformatycznych. Zaprezentował swoje możliwości i zdolności zaproponowania, implementacji algorytmów matematycznych i komputerowych, świetnymi umiejętnościami statystycznej analizy danych. Opracowane przeze niego programy nie zostały niestety udostępnione jako pakiety i serwery sieciowe, co może dziwić i we współczesnych czasach jest to już niespotykane. Wyniki opracowane w trakcie doktoratu zostały opublikowane w dwóch pracach w punktowanych czasopismach o zasięgu międzynarodowym, oraz jednym pre-princie BioRXiv.

Poniżej postaram się krótko wypunktować kluczowe wg. autora i recenzenta osiągnięcia badawcze doktoranta, w tym zaproponowane algorytmy i narzędzia, oraz formalizmy metodologiczne związane z zagadnieniem porównywania danych trójwymiarowych we współczesnej genomice bazującej na technologii 3C i pochodnych.

- opracowanie miary odległości do porównań segmentacji chromosomowych, co jest istotne dla właściwego opisu danych doświadczalnych;
- pokazanie, że ww. miara odległości spełnia własności metryki, co jest niezwykle istotne przy porównywaniu wielu zbiorów danych eksperymentalnych;
- przedstawienie przydatności zaproponowanej metryki do analiz rzeczywistych danych Hi-C, co podbudowuje przedstawione przez Autora tezy;
- wskazanie przy użyciu metryki BP par linii komórkowych, w których segmentacje chromosomowe są znacząco różne mimo tego, że do tej pory wiele publikacji podkreśla, że zasadniczo granice TADów są w dużym stopniu zachowane pomiędzy komórkami;
- zaproponowanie modelu statystycznego do wykrywania istotnych oddziaływań w porównaniach macierzy kontaktów Hi-C:
    - zaproponowanie nowej definicji kontaktów różnicowych opartej na podobieństwie przekątnych macierzy Hi-C,
    - zaproponowanie nowej metody agregacji zbiorów kontaktów (przekątnych), która do tej pory nie była stosowana
    - zaproponowanie nowej metody agregacji wykrywanych kontaktów różnicowych co jest niezwykle istotne w interpretacji doświadczeń;
- potwierdzenie skuteczności opracowanej metody w wykrywaniu kontaktów różnicowych na symulowanych zbiorach danych oraz jej konkurencyjności w stosunku do istniejących metod;
- zasugerowanie istotności tak opracowanych nowych metodologii bioinformatycznych do poszukiwania nowych interakcji regulatorowych między izolatorami, wyciszaczami, enhancerami oraz regionem promotorowym genów.

Ocena końcowa

W podsumowaniu mojej oceny rozprawy doktorskiej Pana magistra Rafała Zaborowskiego stwierdzam, że zaprezentowaną pracę oceniam wysoko. Biorąc pod uwagę czytelność i walory naukowe rozprawy doktorskiej, udane połączenie starannie opisanych i matematycznie sformalizowanych algorytmów, oraz istotnych biologicznie danych i pytań badawczych, oceniam rozprawę doktorską mgr Rafała Zaborowskiego jako ważny wkład do informatyki w zakresie genomiki obliczeniowej i bioinformatyki. Oceniam, że rozprawa ta spełnia zwyczajowe i ustawowe wymogi, stawiane rozprawom doktorskim, stanowi oryginalne rozwiązanie problemu naukowego, unaocznia ogólną wiedzę teoretyczną kandydata w informatyce oraz pokazuje umiejętność samodzielnego prowadzenia pracy naukowej. Wnoszę zatem do Rady Wydziału Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego o dopuszczenie Pana magistra Rafała Zaborowskiego do dalszych etapów przewodu doktorskiego.

Dodatkowo biorąc pod uwagę wysoki poziom merytoryczny rozprawy, jej staranne matematyczne przygotowanie, przejrzysty sposób prezentacji tematyki badawczej oraz wyników wnoszę o wyróżnienie rozprawy stosowną nagrodą.

Profesor dr hab. Dariusz Plewczynski
Laboratorium Genomiki Funkcjonalnej i Strukturalnej
Centrum Nowych Technologii
Uniwersytet Warszawski
ul. Banacha 2c, 02-097 Warszawa, Polska