Prof. dr hab. inż. Andrzej Polański

Katedra Grafiki, Wizji Komputerowej i Systemów Cyfrowych

Politechniki Śląskiej

Gliwice 3.03.2024

Recenzja rozprawy doktorskiej

Autor: Mgr Senbai Kang

Tytuł: Probabilistic graphical models for inferring tumor phylogeny and genomic variants from single cell DNA sequencing data

Promotor: Prof. dr hab. Ewa Szczurek

Ogólna charakterystyka rozprawy

Przedłożona do recenzji rozprawa doktorska jest napisana w języku angielskim. Liczy 150 stron tekstu, składa się z wstępu, dwóch rozdziałów opisujących narzędzia analizy danych sekwencjonowania pojedynczych komórek, dodatków oraz obszernej bibliografii liczącej 145 pozycji. Praca poświęcona jest opisowi oryginalnej metodologii wywoływania mutacji oraz oceny filogenezy komórek nowotworowych w danych pochodzących z sekwencjonowania DNA pojedynczych komórek. Problem wywoływania mutacji w danych pochodzących z eksperymentów sekwencjonowania pojedynczych komórek jest bardzo intensywnie opracowywany w literaturze. Algorytmy jego rozwiązywania mają bardzo duże znaczenie dla współczesnej biologii molekularnej, genetyki, genomiki, bioinformatyki. Modelowanie genealogii / filogenezy komórek nowotworowych jest bardzo często elementem badań nad rozwojem nowotworów. Poświęca mu się w literaturze wiele uwagi. Uzyskanie oryginalnych wyników naukowych w intensywnie rozwijanych obszarach badawczych należy uznać za istotne osiągnięcie naukowe. Opisane w pracy metody mają bezpośrednie odniesienie do opublikowanych artykułów oraz do tekstów dostępnych w Internecie na platformie arxiv, których współautorem (pierwszym autorem) jest Doktorant.

Pierwszy rozdział pracy stanowi wstęp. Otwierają go krótkie opisy mechanizmów biologicznych nowotworzenia oraz technik eksperymentalnych sekwencjonowania DNA pojedynczych komórek. Kolejnym elementem rozdziału wstępnego pracy są opisy modeli matematycznych wykorzystywanych w konstrukcji algorytmu wywoływania mutacji. Aparat

matematyczny wykorzystywany w pracy jest zaawansowany i różnorodny. Opisywane (wymieniane) są kolejne jego elementy, probabilistyczne modele grafowe definiowane głównie jako sieci Bayesowskie, techniki odtwarzania drzew filogenetycznych oraz metody próbkowania / aproksymacji stochastycznej MCMC (Markov Chain Monte Carlo). Ostatnią składową rozdziału wstępnego jest charakterystyka motywacji do podjęcia badań opisanych w pracy, omówienie rozkładu pracy oraz wymienienie oryginalnych elementów pracy. Motywacją do badań w pracy są wyzwania stojące przed rozwijanymi obecnie algorytmami analizy danych sekwencjonowania pojedynczych komórek. Wśród nich stosunkowo niskie oraz niejednorodne pokrycie oraz występowanie w procesie nowotworzenia wariantów mutacji o różnorodnym charakterze, substytucji, delecji i insercji, mutacji nawracających i rekurencyjnych, utraty heterozygotyczności, wariantów liczby kopii. Oryginalne elementy pracy to przede wszystkim opracowanie adekwatnych modeli matematycznych ujmujących genetykę i genomikę mutowania DNA w ewolucji nowotworów, aspekty techniczne sekwencjonowania DNA, możliwość wywoływania różnych typów mutacji, powiązanie procedur wywoływania mutacji w DNA komórek nowotworowych z oceną ich filogenetyki, opracowanie odpowiednich algorytmów i oprogramowania (w środowisku JAVA), wreszcie wykazanie efektywności i przydatności rozwiniętych podejść na tle innych narzędzi dostępnych już w literaturze. W kontekście oceny filogenetyki komórek nowotworowych interesującym, nowatorskim podejściem jest ocena długości pnia drzewa filogenetycznego. Jest to możliwe i uzasadnione dlatego, że dostępne są referencyjne stany sekwencji DNA w stosunku do wyznaczonych mutacji.

Kolejne dwa rozdziały pracy są poświęcone opisowi dwóch narzędzi do wywoływania mutacji w DNA pojedynczych komórek. Rozdział drugi poświęcony jest opisowi opracowanego algorytmu analizy danych sekwencjonowania DNA pojedynczych komórek oraz związanego z nim narzędzia „SIEVE". Przedstawia się model matematyczny narzędzia „SIEVE", który składa się z modelu w postaci sieci Bayesowskiej przedstawionej na rysunku 2.1 oraz modelu drzewa filogenetycznego dla wywoływanych mutacji. Typy wykrywanych / wywoływanych mutacji wypisane są w tabeli 2.1. Funkcja wiarygodności (logarytmiczna funkcja wiarygodności), która łączy w sobie wiarygodność sieci Bayesowskiej z rysunku 2.1 oraz drzewa filogenetycznego zapisana jest w ogólny sposób we wzorze 2.13. Relacje budujące zależność 2.13 są szczegółowo omówione w części „Methods" rozdziału 2.

W kolejnej części rozdziału 2 („Results") dokonywana jest analiza efektywności narzędzia „SIEVE" oraz porównań do innych dostępnych narzędzi wywoływania / wykrywania mutacji w danych sekwencjonowania DNA pojedynczych komórek. Dokonuje się porównań czułości i specyficzności wykrywanych mutacji, a także innych adekwatnych indeksów, porównuje się także topologie oraz metryki oceniony przez narzędzie „SIEVE" drzew filogenetycznych ewolucji mutacji z analogicznymi parametrami ocenionymi z użyciem innych narzędzi. Podsumowaniem tych porównań są wnioski prowadzące do uznania przewagi opracowanego narzędzia „SIEVE" nad innymi podejściami.

W rozdziale 3 przedstawia się drugi algorytm oraz powiązane z nim narzędzie programowe „DELSIEVE". Jest ono rozwinięciem opisanego w poprzednim rozdziale narzędzia „SIEVE". Lista możliwych mutacji DNA jest tu rozbudowana o delecje i insercje. Konstrukcja algorytmu „DELSIEVE" jest całkowicie analogiczna do konstrukcji algorytmu „SIEVE". Sieć Bayesowska modelująca zależności pomiędzy ocenianymi genotypami a dostępnymi danymi, z użyciem założonych postaci rozkładów prawdopodobieństwa jest przestawiona na rysunku 3.1. Jak widać podstawowa różnica pomiędzy siecią z rysunku 3.1 a siecią z rysunku 2.1 jest w macierzy przejść pomiędzy genotypami. Lista analizowanych genotypów jest tu szersza (tabela 3.1) przez co macierz przejść ma tu także więcej wierszy i kolumn. Podobnie jak w poprzednim rozdziale genotypy ocenia się przez maksymalizację (logarytmicznej) funkcji wiarygodności zapisanej wzorem 3.14. Ma ona dwie składowe, pierwsza odpowiada za wiarygodność zapisaną w sieci Bayesowskiej a druga jest wiarygodnością wnoszoną przez drzewo filogenetyczne związane z ewolucją komórek nowotworowych.

Podobnie jak w poprzednim rozdziale, kolejnym etapem analizy jest ocena jakości opracowanego narzędzia „DELSIEVE". Wykorzystuje się wskaźniki analogiczne do tych w poprzednim rozdziale. Istnieje różnica, polegająca na tym, że z uwagi na nowatorski charakter narzędzia „DELSIEVE" nie można dobrać referencyjnych narzędzi do porównań.

Rozdział czwarty stanowi podsumowanie wyników pracy.

Rozdział piąty jest to bardzo obszerny suplement, zawierający szczegółowo opisane wyniki wielu dodatkowych analiz potwierdzających skuteczność zaproponowanych podejść.

Ocena rozprawy

Recenzowana praca jest bardzo ściśle związana z oryginalną publikacją:

Kang, S., Borgsmüller, N., Valecha, M., Kuipers, J., Alves, J. M., Prado-López, S., ... & Szczurek, E. (2022). SIEVE: joint inference of single-nucleotide variants and cell phylogeny from single-cell DNA sequencing data. *Genome Biology*, *23*(1), 248.,

oraz z manuskryptem opublikowanym w bazie danych biorxiv:

Kang, S., Borgsmüller, N., Valecha, M., Markowska, M., Kuipers, J., Beerenwinkel, N., ... & Szczurek, E. (2023). DelSIEVE: joint inference of single-nucleotide variants, somatic deletions, and cell phylogeny from single-cell DNA sequencing data. *bioRxiv*, 2023-09.

Czasopismo naukowe Genome Biology jest bardzo renomowane (punktacja ministerialna 200, IF=12.3, 99 centyl). Obie prace są wieloautorskie. Doktorant jest pierwszym autorem w obu pracach. W podpunkcie 1.7.1 doktoratu opisany jest dokładnie wkład naukowy Doktoranta w treść obu prac. Obejmuje on przede wszystkim zaprojektowanie potoku obliczeń bioinformatycznych, wdrożenie obliczeń oraz badań

symulacyjnych oraz porównawczych, a także napisanie obu prac. Należy zatem ocenić, że wkład Doktoranta w obie te prace jest oryginalny i bardzo istotny.

Doktorant posiada także inne bardzo ciekawe artykuły naukowe na swoim profilu Google Scholar.

Dużym walorem pracy jest jej logiczna konstrukcja, bardzo jasny język, precyzja formalizmu matematycznego. Praca jest napisana bardzo starannie. Należy podkreślić, że mimo silnego powiązania treści doktoratu z dwoma powyższymi publikacjami, nie jest on streszczeniem czy omówieniem tych artykułów. Metody matematyczne stosowane w tych artykułach są omawiane w doktoracie systematycznie, znacznie szerzej i dokładniej niż ma to miejsce w artykułach. Wszystkie zagadnienia są w doktoracie przedstawione bardziej wyczerpująco niż ich odpowiedniki w dwóch artykułach naukowych.

W ramach pracy zostało opracowane oprogramowanie dostępne na platformie Github. Dostępny jest zarówno kod jak i wyniki obliczeń benczmarkowych. Staranność opracowania tej dokumentacji wykazuje doświadczenie programistyczne Doktoranta.

Doktorant wykazuje się bardzo szerokimi kompetencjami w zakresie modelowania matematycznego i technik informatycznych. Potrafi także prowadzić badania o charakterze interdyscyplinarnym, potrafi współpracować w grupie naukowej o interdyscyplinarnym.

Konkluzja

Praca stanowi podsumowanie oryginalnych, interdyscyplinarnych badań 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.