UNIWERSYTET
WARSZAWSKI

Wydział Matematyki, Informatyki i Mechaniki
Instytut Informatyki
dr hab. Norbert Dojer

Warszawa, 20.03.2021.

Recenzja rozprawy doktorskiej *mgr inż. Wiktora Kuśmirka*
zatytułowanej
Szacowanie liczby powtórzeń fragmentu DNA

Strona formalna rozprawy

Rozprawa doktorska mgr inż. Wiktora Kusmirka ma charakter cyklu pięciu artykułów. Wszystkie artykuły zostały opublikowane w renomowanych czasopismach naukowych (70-200 punktów ministerialnych, impact factor 2.213-7.267): dwa w BMC Bioinformatics oraz po jednym w GigaScience, Scientific Data i BioMed Research International. Wszystkie artykuły są wieloautorskie, w każdym przypadku wkład doktoranta jest starannie opisany i potwierdzony oświadczeniami współautorów. W publikacjach [P1], [P2] i [P3] wkład doktoranta został oceniony na 70-80% i obejmował m.in. zaprojektowanie badań/opracowanie algorytmów, wykonanie eksperymentów oraz przygotowanie tekstu manuskryptów i przeprowadzenie procesu publikacji, uzasadnione jest więc uznanie mgr inż. Wiktora Kuśmirka za głównego autora tych prac. W artykule [P4] wkład doktoranta został oceniony na 10%, natomiast w pracy [P5] na 25%. Należy jednak zaznaczyć, że ta ostatnia publikacja jest efektem interdyscyplinarnej współpracy kilku grup badawczych, a wkład doktoranta w obliczeniową część projektu można uznać za wiodący. Dołączony autoreferat w przejrzysty sposób prezentuje główne wyniki wymienionych artykułów i wyjaśnia ich znaczenie.

Tematyka badań

Tematyka rozprawy skupiona jest wokół dwóch zagadnień:

- wyznaczanie występującej w genomie liczby kopii fragmentów DNA z danych

ul. Banacha 2, 02-097 Warszawa, Polska
+48 22 55 44 577
e-mail: dojer@mimuw.edu.pl

z sekwencjonowania pełnoeksomowego,

- asemblacja *de novo* sekwencji genomowych ze szczególnym uwzględnieniem obszarów repetytywnych.

Tytuł rozprawy, czyli *Szacowanie liczby powtórzeń fragmentu DNA*, łączy te zagadnienia podkreślając jej spójność tematyczną. Warto jednak zaznaczyć, że wyniki uzyskane w publikacjach poświęconych drugiemu z wymienionych zagadnień istotnie wykraczają poza zasugerowany w tytule zakres.

Problem składania sekwencji repetytywnych to obecnie jedno najważniejszych wyzwań w dziedzinie asemblacji sekwencji genomowych. O ile większość genomu człowieka zrekonstruowana została już 20 lat temu, poznanie obszarów centro- i telomerowych przez wiele lat wydawało się nieosiągalne ze względu na ich złożoną strukturę. Istotny postęp w tej dziedzinie dokonał się dopiero w ostatnich latach za sprawą doskonalenia technologii sekwencjonowania trzeciej generacji oraz, co nie mniej ważne, opracowania metod analizy zdolnych do wykorzystania informacji zawartej w otrzymanych tymi technologiami danych. Wiele prac dotyczących tej tematyki ukazało się w prestiżowych czasopismach w ciągu ostatnich trzech lat, czyli równolegle lub już po ukazaniu się publikacji wchodzących w skład rozprawy. Dlatego uważam, że tematyka rozprawy bardzo dobrze wpisuje się w aktualne badania w obszarze problemu asemblacji sekwencji DNA.

Główne wyniki rozprawy

Pierwszemu z wymienionych zagadnień, czyli wyznaczaniu liczby kopii fragmentów DNA, poświęcone są dwie prace z cyklu: [P1] oraz [P4]. Praca [P1] dotyczy doboru próbek do modelowania tła podczas szacowania liczby kopii. Zaproponowano zastosowanie w tym celu grupowania próbek metodą k-means. Pokazano, że k-means daje podobnie dobre wyniki jak najlepsze z dotychczas stosowanych podejść, czyli algorytm k najbliższych sąsiadów, ale jest znacznie szybszy.

Artykuł [P4] poświęcony jest wcześniejszemu etapowi wyznaczania liczby kopii wariantów, tzn. obliczaniu głębokości pokrycia odczytami z sekwencjonowania analizowanych fragmentów DNA. W pracy opracowano i zaimplementowano efektywnie zrównoleglonąaplikację obliczającą głębokość pokrycia.

Pozostałe prace cyklu dotyczą drugiego ze wspomnianych zagadnień, czyli asemblacji sekwencji genomowych. Artykuł [P2] prezentuje aplikację dnaasm do asemblacji *de novo* odczytów z sekwencjonowania technologią drugiej generacji. Algorytm asemblacji opiera się na powszechnie stosowanych przy tym problemie grafach de Bruijna, ale w nowatorski sposób wykorzystuje głębokość pokrycia grafu

ul. Banacha 2, 02-097 Warszawa, Polska
+48 22 55 44 577
e-mail: dojer@mimuw.edu.pl

odczytami do zrekonstruowania sekwencji repetytywnych.

Z kolei praca [P3] dotyczy asemblacji heterogenicznych zbiorów danych, tzn. łączących odczyty z drugiej i trzeciej generacji sekwencjonowania. W pracy zaprezentowano aplikację dnaasm-link, służącą do łączenia contigów otrzymanych w wyniku asemblacji krótkich odczytów z drugiej generacji sekwencjonowania w oparciu o długie odczyty z trzeciej generacji. Aplikacja pozwala ponadto na wykorzystanie długich sekwencji do wypełnienia przerw pomiędzy contigami.

Aplikacje dnaasm i dnaasm-link zostały wykorzystane przez doktoranta do złożenia genomu tasiemca szczurzego. Opisany w artykule [P5] rezultat asemblacji pozwolił znacząco poprawić jakość genomu referencyjnego tasiemca, np. parametr N50 został zwiększony ponad czterdziestokrotnie. Odtworzenie w całości sekwencji genomu mitochondrialnego potwierdza skuteczność opracowanych aplikacji w asemblacji obszarów repetytywnych.

Podsumowując, narzędzia opracowane w pracach [P1] i [P4] pozwalają istotnie przyspieszyć wyznaczanie liczby kopii fragmentów DNA przy zachowaniu wysokiej jakości wyników, natomiast aplikacje opracowane w pracach [P2] i [P3] umożliwiają poprawę jakości asemblacji repetytywnych obszarów sekwencji DNA. Na podkreślenie zasługuje wartość praktyczna uzyskanych wyników, w przypadku narzędzi dnaasm i dnaasm-links potwierdzona zastowaniem w opisanym w pracy [P5] rzeczywistym projekcie sekwencjonowania genomu eukariotycznego.

Uwagi krytyczne i dyskusyjne

Poniższe uwagi nie podważają oceny uzyskanych w rozprawie wyników.

1. W pracach [P2] i [P3] sprawdzenie efektywności odtwarzania obszarów repetytywnych zostało oparte na klasyfikacji wynikowych sekwencji programem Tandem Reapeat Finder. Intencja stojąca za przyjęciem takiej miary jest dla mnie niezrozumiała. Celem nie jest bowiem zwrócenie przez program sekwencji wykazującej cechy repetytywności, ale wierne odtworzenie oryginalnej sekwencji z obszaru repetytywnego. To ostatnie mogło być sprawdzone bezpośrednio, poprzez porównanie wynikowych sekwencji z odpowiednimi fragmentami genomu referencyjnego.
2. Zarazem narzędzie Tandem Reapeat Finder znakomicie nadaje się do wykrywania fałszywych pozytywów w rekonstrukcji obszarów repetytywnych, czyli sekwencji błędnie uznanych przez algorytm asemblacji za powtórzenia. W rozprawie brakuje dyskusji tego problemu, choć podejście zastosowane w aplikacji dnaasm (czyli szacowanie liczby kopii na podstawie głębokości

ul. Banacha 2, 02-097 Warszawa, Polska
+48 22 55 44 577
e-mail: dojer@mimuw.edu.pl

pokrycia) stwarza niebezpieczeństwo wystąpienia tego rodzaju błędów. Co więcej, wyniki zamieszczone w tabelach 5 i 6 pracy [P2] świadczą o tym, że algorytmowi zdarza się przeszacować liczbę kopii sekwencji repetytywnej.

3. W pracy [P3] przedstawiono klasyfikację metod łącznego wykorzystania do asemblacji odczytów z drugiej i trzeciej generacji sekwencjonowania na cztery różne podejścia. Tymczasem ewaluacja aplikacji dnaasm-link została ograniczona do porównania z narzędziami reprezentującymi to samo podejście. Dla uzyskania pełnego obrazu możliwości zaproponowanej metody należałoby dołączyć do porównania narzędzia należące do pozostałych kategorii, ewentualnie scharakteryzować mocne i słabe strony poszczególnych podejść.
4. W pracy [P5] obok aplikacji dnaasm do asemblacji genomu został użyty także program ABYSS. Podobnie do łączenia contigów obok dnaasm-link wykorzystano też program LINKS. Celowość zastosowania różnych programów do wykonania tych samych zadań została uzasadniona tylko częściowo – wspomniano, że wykorzystanie dnaasm-link było podyktowane niewystarczającą wydajnością pamięciową programu LINKS. Dla właściwej oceny faktycznego wkładu aplikacji dnaasm i dnaasm-links w asemblację genomu tasiemca szczurzego należałoby scharakteryzować obszary zastosowania poszczególnych narzędzi bądź opisać zasady agregowania ich wyników, jeśli zostały zaaplikowane do tych samych danych.

Konkluzja

Uważam, ze przedstawiona rozprawa spełnia zwyczajowe i ustawowe wymogi stawiane rozprawom doktorskim, stanowi oryginalne rozwiązanie problemu naukowego, unaocznia ogólną wiedzę i umiejetnosci techniczne doktoranta w informatyce oraz świadczy o umiejętności samodzielnego prowadzenia pracy naukowej. Wnoszę zatem o dopuszczenie Pana magistra inżyniera Wiktora Kuśmirka do dalszych etapów przewodu doktorskiego. Ponadto, biorąc pod uwagę wysoki poziom merytoryczny rozprawy, znaczenie podjętych problemów badawczych oraz walory praktyczne uzyskanych wyników, wnioskuję o wyróżnienie rozprawy.

Norbert Dojer

ul. Banacha 2, 02-097 Warszawa, Polska
+48 22 55 44 577
e-mail: dojer@mimuw.edu.pl