Prof. dr hab. inż. Sebastian Deorowicz
Katedra Algorytmiki i Oprogramowania
Wydział Automatyki, Elektroniki i Informatyki
Politechnika Śląska
44-100 Gliwice, ul. Akademicka 16

Gliwice, 16.09.2022 r.

Recenzja rozprawy doktorskiej

Tytuł rozprawy:
Inference of Credible Associations between Genes and Genomes
(tytuł polski: Rekonstrukcja wiarygodnych relacji między genami a genomami)

Autor:
mgr Agnieszka Mykowiecka

Promotor:
Dr hab. Paweł Górecki

Problematyka rozprawy

Uzyskanie wiarygodnych informacji o pochodzeniu gatunków, które można reprezentować w postaci drzewa filogenetycznego jest kluczowe dla naszego zrozumienia tego jak przebiegała ewolucja i jakie zdarzenia ewolucyjne miały miejsce w bliższej bądź odleglejszej przeszłości. Tworzenie drzewa filogenetycznego jest zadaniem trudnym, gdyż musimy tu polegać często na niekompletnych informacjach a także uwzględniać fakt, że zdarzenia ewolucyjne zachodzą mniej lub bardziej przypadkowo. Nie mamy także dostępu do informacji genetycznej gatunków dawno wymarłych. Wszystko to sprawia, że problematyka ta jest dalej bardzo aktualna i pozostaje jeszcze wiele do zrobienia a istniejące drzewa filogenetyczne należy traktować raczej jako hipotezę na temat tego jak przebiegała ewolucja.

Jednym z głównych narzędzi, którym dysponujemy jest wykorzystanie wielu drzew genów. Drzewa te reprezentują zależności pomiędzy genami występującymi w różnych gatunkach a pełniącymi podobne funkcje. Wiedząc, z których gatunków pochodzą dane geny możliwe jest podjęciu próby uzgodnienia wielu drzew genów w celu uzyskania drzewa filogenetycznego. Zadanie to zdecydowanie nie należy do łatwych.

Można także wykorzystać drzewo filogenetyczne jako pewnego rodzaju wzorzec, który choć jest tylko hipotezą, to jednak bez wątpienia lepiej opisuje pokrewieństwo gatunków niż pojedyncze drzewo genów. Z wzorcem tym możemy zestawić pojedyncze drzewo genów i na podstawie różnic w topologii możemy szukać sygnałów świadczących o wystąpieniu w przeszłości zdarzeń ewolucyjnych takich jak duplikacja, straty czy horyzontalny transfer genów (HGT). Pozwala to lepiej zrozumieć to jak ewoluowały różne gatunki. Właśnie problematyka uzgadniania drzew genów z drzewami filogenetycznymi jest jednym z głównych problemów, których dotyczy recenzowana praca.

Analiza treści rozprawy oraz uzyskanych wyników

Treść rozprawy

Doktorantka postawiła sobie za cel stworzenie nowych algorytmów, które będą umożliwiały wykrywanie zdarzeń ewolucyjnych m.in. na podstawie analizy podobieństwa drzew genów i drzew gatunków.

Rozprawa napisana jest w języku angielskim i składa się z 7 rozdziałów. W pierwszym rozdziale zawarte jest krótkie wprowadzenie w tematykę pracy, nakreślenie rysu historycznego oraz określenie problemów podjętych w rozprawie.

Rozdział drugi zawiera podstawowe definicje, w tym dotyczące kluczowych pojęć: drzewa gatunków i drzewa genów. Przedstawiono tu także istniejące metody uzgadniania drzew, zarówno ukorzenionych jak i nieukorzenionych.

Rozdział trzeci podejmuje problematykę oceny wiarygodności zdarzeń duplikacji i specjacji. Doktorantka proponuje tutaj nową miarę wsparcia dla tych zdarzeń. Następnie proponuje liniowy algorytm pozwalający na jej wyznaczanie. Algorytm ten oparty jest na uzgadnianiu nieukorzenionych drzew i nieparametrycznym bootstrapie. Metody te zostały zaimplementowane i poddane ocenie eksperymentalnej na zestawach zarówno wygenerowanych sztucznie jak i na danych rzeczywistych.

W rozdziale czwartym Doktorantka skupiła się na problematyce wykrywania zdarzeń horyzontalnego transferu genów, często spotykanego pomiędzy różnymi gatunkami bakterii. W tym celu opracowano miarę opartą się na nieparametrycznym bootstrapie, nazwaną wsparciem transferu. Pokazano także algorytm pozwalający na wyznaczanie najlepszych kandydatów na zdarzenia HGT. W części eksperymentalnej pokazano jak zachowuje się ta metoda zarówno dla blisko, jak i daleko, spokrewnionych gatunków.

Rozdział piąty podejmuje problem wykrywania przyporządkowania genów zidentyfikowanych w danych z sekwencjonowania do gatunków. W przypadku kiedy prowadzimy badania metagenomowe, w sekwencjonowanej próbce mogą się znajdować fragmenty genów rozmaitych gatunków i rozpoznanie tego z jakimi gatunkami mamy do czynienia może nie być oczywiste. Doktorantka proponuje tutaj dwie metody oparte na uzgadnianiu z transferami. Skuteczność tego podejścia oceniona jest na podstawie wyników eksperymentalnych.

Rozdział szósty podejmuje problemy, z którymi spotykamy się w przypadku, w którym duże podobieństwo sekwencji bardzo utrudnia albo wręcz uniemożliwia skonstruowanie drzew filogenetycznych, które opisywałyby złożone relacje ewolucyjne. W takim przypadku możliwe jest skonstruowanie sieci opisujących te zależności. Przykładem, na którym oparto się w pracy są sekwencje receptora BCR pochodzące z limfocytów B pacjentów chorych na chłoniaka pęcherzykowego.

Ostatni rozdział podsumowuje osiągnięte w pracy wyniki i stanowi także pewnego rodzaju przewodnik po wcześniejszych rozdziałach.

Bibliografia rozprawy składa się z ponad 110, dobrze dobranych, pozycji.

Najważniejsze wyniki przedstawione w rozprawie

Doktorantka opracowała kilka interesujących miar oraz algorytmów ich wyznaczania, które mogą służyć do lepszego uzgadniania drzew filogenetycznych i drzew genów, identyfikowania zdarzeń horyzontalnego transferu genów czy też określania pochodzenia genów w eksperymentach metagenomowych. Wyniki te zostały zweryfikowane eksperymentalnie. Warto podkreślić, że znalazły one też odzwierciedlenie w publikacjach naukowych w liczących się czasopismach i materiałach konferencji naukowych. Prawdopodobnie najciekawsze są tutaj metody oparte na nieparametrycznym bootstrapie.

Uwagi merytoryczne

Praca zawiera dość dobre wprowadzenie w tematykę wykrywania i analizy relacji między genami i gatunkami. Zaproponowane metody omówione są precyzyjnie. Nie zabrakło także formalnego udowodnienia właściwości proponowanych metod. Badania eksperymentalne są dość dobrze zaprojektowane a ich wyniki zanalizowane.

Mam jednak pewne uwagi krytyczne i prosiłbym o odniesienie się do nich w trakcie publicznej obrony.

1. Badania eksperymentalne zostały przeprowadzone na stosunkowo niedużych zestawach danych. Przykładowo w podrozdziale 3.2 mowa o 9 genomach, w podrozdziale 4.7 jest to 29 gatunków. Z jednej strony pozwala to na ręczną analizę uzyskanych wyników. Z drugiej zaś strony rodzi pytania o skalowalność proponowanych metod dla znacznie większych zestawów danych oraz o możliwość jakiegoś zautomatyzowanego wykorzystania uzyskiwanych wyników. Prosiłbym o odniesienie się do tej kwestii skalowalności.
2. W podrozdziale 4.7.1 wskazano, że sekwencje genowe zostały dopasowane za pomocą narzędzia MUSCLE. Jest to jednak dość stary algorytm a w międzyczasie opublikowano wiele nowszych, które wydają się być znacznie dokładniejsze. Prosiłbym o uzasadnienie wyboru MUSCLE i/lub próbę odpowiedzi na pytanie czy wykorzystanie nowszych narzędzi mogłoby się jakoś przełożyć na uzyskane w rozprawie wyniki.
3. Opracowane algorytmy podane są w postaci pseudokodów, co jest cenne, ale jednak w środowisku bioinformatycznym częste jest publikowanie opracowanych metod w postaci gotowych do użycia kodów. Nie znalazłem w pracy informacji o tym czy taka publikacja miała miejsce. Jeśli tak, to prosiłbym o jej wskazanie. Jeśli nie, to prosiłbym o odpowiedź czy Doktorantka planuje udostępnienie tych kodów, co z pewnością przyczyniłoby się do większego zainteresowania wynikami opisywanych badań.

Uwagi redakcyjne

Ogólna kompozycja rozprawy jest poprawna. Sposób składu tekstu także jest poprawny. Strona formalna także nie budzi moich większych zastrzeżeń. W pracy pojawiają się jednak drobne błędy redakcyjne. Ich liczba nie jest duża i nie utrudnia lektury, więc ograniczę się tylko do wskazania niektórych:

- [str. 17] "is a agglomerative".
- [str. 17] "returns and unrooted tree".
- [str. 22] "((a, b), c)" -> "(a, (b, c))".
- [str. 25] "the the graph".
- [str. 30] "A commonly used method..." - w tym zdaniu brak orzeczenia.

Podsumowanie

Przedstawiona do oceny rozprawa zawiera interesujące wyniki. Tematyka jest bez wątpienia ciekawa i ważna. Uzyskane wyniki mogą być podstawą do dalszych prac a możliwe kierunki Doktorantka sama wskazuje. Drobne niedociągnięcia redakcyjne nie umniejszają mojej pozytywnej oceny przedłożonej rozprawy.

Konkluzja

Podsumowując opinie zawarte w mojej recenzji, mogę stwierdzić, że zgodnie z „Ustawą o Stopniach naukowych..." recenzowana rozprawa mgr Agnieszki Mykowieckiej nt. „Inference of Credible Associations between Genes and Genomes" stanowi oryginalne rozwiązanie problemu naukowego oraz wykazuje ogólną wiedzę teoretyczną kandydatki w dyscyplinie informatyka oraz umiejętność prowadzenia pracy naukowej. Zawarte w recenzji uwagi krytyczne mają charakter dyskusyjny i nie wpływają na pozytywną ocenę rozprawy. Wnoszę zatem o dopuszczenie wspomnianej rozprawy do publicznej obrony.

Sebastian Deorowicz