dr hab. Norbert Dojer
Wydział Matematyki, Informatyki i Mechaniki
Uniwersytetu Warszawskiego
tel. 662462016, e-mail dojer@mimuw.edu.pl

22.03.2019.

Recenzja rozprawy doktorskiej

Wojciecha Frohmberga

zatytułowanej:

GRASShopPER - wydajna metoda asemblacji de novo wykorzystująca strategię Overlap-Layout-Consensus

1. Problem badawczy i jego znaczenie

Rozprawa poświęcona jest problemowi asemblacji genomów z odczytów z sekwencjonowania DNA nowej generacji. Asemblacja jest niezbędna do zrekonstruowania sekwencji genomów organizmów, ponieważ odczyty z sekwencjonowania mają długość istotnie mniejszą od długości chromosomów - w najpowszechniej stosowanych techonologiach sekwencjonowania drugiej generacji jest to kilkaset nukleotydów wobec milionów, a nawet setek milionów nukleotydów w chromosomach. Co więcej, chociaż znane są już genomy dla tysięcy gatunków, praktyczne znaczenie asemblacji wydaje się nie maleć ze względu na rosnące zainteresowanie genomiką nowotworów, analizami różnorodności genomów w obrębie poszczególnych gatunków itp.
Zarazem asemblacja stanowi wysoce nietrywialny problem badawczy. Wyzwaniem jest z jednej strony osiągnięcie satysfakcjonujących parametrów jakościowych (które jest szczególnie kłopotliwe ze względu na wielość kryteriów oceny), z drugiej zaś rozmiar przetwarzanych danych (który wymaga starannego zaprojektowania algorytmów i struktur danych).
Przedmiotem rozprawy jest nowa metoda asemblacji oparta na klasycznym podejściu Overlap-Layout-Consensus. Podejście to było wiodące w czasach technologii sekwencjonowania metodą Sangera, po upowszechnianiu się drugiej generacji sekwencjowania utraciło swoje znaczenie na rzecz grafów de Bruijn'a, obecnie zaś przeżywa swój renesans w związku z powstaniem technologii sekwencjowania trzeciej generacji. Przedstawiona w rozprawie metoda przystosowana jest do asemblacji odczytów z sekwencjonowania drugiej generacji, ale wydaje się posiadać również duży potencjał do zastosowania z odczytami z sekwencjonowania generacji trzeciej.

2. Wkład autora

Rozprawa zawiera opis zastosowanej strategii, implementację metody w programie GRASShopPER oraz jej ewaluację na danych symulowanych i rzeczywistych. Strategia opiera się na klasycznym podejściu Overlap-Layout-Consensus, ale w jego realizacji zastosowano szereg interesujących rozwiązań, spośród których dwa poniższe uważam za szczególnie innowacyjne:

- zastosowanie charakterystyk *k*-merowych do selekcji obiecujących par odczytów w fazie Overlap,
- algorytm znajdowania znaczących rozwidleń w grafie nałożeń w fazie Layout.

Program GRASShopPER jest dziełem wieloautorskim, zadeklarowany wkład autora rozprawy zawiera:

- udział w opracowaniu ogólnej koncepcji metody,
- opracowanie szczegółowej koncepcji i implementację faz Layout oraz Consensus, w tym wspomnianego wyżej algorytmu znajdowania znaczących rozwidleń w grafie,
- ewaluację metody.

Uważam, że jest to wkład znaczący, w szczególności wystarcza on stwierdzenia, że doktorant wykazał się umiejętnością rozwiązywania złożonych problemów bioinformatycznych, w tym projektowania, implementacji i ewaluacji przeznaczonych do tego narzędzi.

3. Poprawność

Część pracy zawierająca formalną definicję problemu asemblacji oraz podstawy teoretyczne do jej sformułowania obarczona jest dwoma mankamentami. Po pierwsze, autor przesadnie często stara się formułować definicje i stwierdzenia za pomocą wyrażeń symbolicznych. Sens tych wyrażeń jest mało czytelny, a ponadto zawierają one liczne luki i nieścisłości, np.:

- w warunku (3.3.2c) rozważane zbiory powinny nie tyle mieć puste przecięcie, co raczej być parami rozłączne,
- w warunku (3.3.2d) rozważana suma zbiorów powinna być równa $I \cup RC(I)$,
- w definicji 3.4.1:
  - kolejność indeksów przy $v_{i,j}$ została zamieniona względem stosowanej wcześniej,
  - w nierówności (3.4.1b) sumy powinny zostać zastąpione przez maksimum, a $v_{|si|,2}$ przez $v_{i,2}$,
- wyrażenie (3.3.3b) zawiera niezdefiniowany operator trimEnds,
- brakuje wyjaśnienia, w jaki sposób funkcja kosztu zdefiniowana wyrażeniem (3.3.3) realizuje postulowaną własność, czyli karanie błędnych wydłużeń kontigów zawierających obszary repetytywne,
- algorytm opisany w pseudokodzie 2.4.1
  - pomija parametr *p* w liniach 7 i 9,
  - maksymalizuje funkcję kosztu zamiast ją minimalizować.

Drugim mankamentem teoretycznej części rozprawy jest jej słaby związek z częścią praktyczną. Kryterium oceny wyniku algorytmu asemblacji, wprowadzone w części teoretycznej jako *formalna definicja problemu asemblacji w postaci problemu optymalizacyjnego*, nie jest wykorzystane w dalszej części rozprawy. Dobrym uzasadnieniem wprowadzenia takiego kryterium byłoby zastosowanie go w zaproponowanej metodzie bądź w ewaluacji jakości wyników uzyskanych różnymi metodami. Jednak ewaluacja jakości opiera się na innych kryteriach, zaś z opisu strategii przyjętej w GRASShopPERze wynika, że kierowano się raczej zapewnieniem postulowanych własności rozwiązania problemu niż optymalizacją proponowanego kryterium. Nawet wyniki dotyczące złożoności problemu asemblacji, przytoczone w ostatnim akapicie rozdziału wprowadzającego kryterium, odnoszą się do odmiennego sformułowania tego problemu.

Powyższe uwagi nie podważają wartości głównego wyniku pracy, czyli programu GRASShopPER. Praktyczna część rozprawy zawiera staranny opis zastosowanego podejścia i jego implementacji, jak również rezultaty dość wszechstronnej ewaluacji na danych symulowanych i rzeczywistych. Pewien niedosyt budzi jedynie brak dyskusji doboru parametrów programu GRASShopPER do specyfiki danych wejściowych. W rozprawie przetestowano wpływ na wyniki zarówno parametrów programu, jak i właściwości danych wejściowych, jednak analizy te są od siebie oderwane. Tymczasem dla potencjalnego użytkownika kluczowa jest informacja, jak optymalnie dobrać parametry programu do analizowanych danych. O tym, że kwestia ta jest dalece nieoczywista, wydaje się świadczyć dobór

parametrów w eksperymentach porównujących GRASShopPER z innymi programami - parametry zastosowane do asemblacji poszczególnych zbiorów danych znacząco się różnią.

4. Wiedza kandydata

Rozdziały 1-3 rozprawy opisują istniejący stan wiedzy na temat problemu asemblacji. Opis ten zawiera m.in. szeroką prezentację interdyscyplinarnego tła problemu, które obejmuje zagadnienia z biologii, technologii sekwencjonowania, złożoności obliczeniowej, teorii grafów i bioinformatyki. Zawartość tych rozdziałów potwierdza rozległą wiedzę doktoranta w zakresie wymienionych dziedzin. Pewne zastrzeżenia mam do prezentacji zagadnień związanych z dopasowaniem sekwencji biologicznych (sekcja 2.4) oraz z definicją problemu asemblacji (sekcja 3.3). W obu sekcjach mieszają się klasyczne ujęcia omawianych zagadnień z ujęciami alternatywnymi, najwyraźniej zaproponowanymi przez autora (przynajmniej ja nie odnalazłem ich w referowanych pozycjach z bibliografii). Te ostatnie z jednej strony wzbudzają zastrzeżenia wymienione powyżej w sekcji dotyczącej poprawności, z drugiej zaś prowadzą do konfuzji pojęć – np. defnicja 2.4.3 wprowadza pojęcie dopasowania, które ma odpowiadać angielskiemu terminowi *multiple sequence alignment*, tymczasem definiowany obiekt jest raczej sekwencją consensusową dopasowania w sensie stosowanym powszechnie w literaturze.
Bibliografia pokrywa poruszane w pracy zagadnienia i można ją uznać za wyczerpującą. Pewien niedosyt budzi brak pozycji prezentujących w sposób syntetyczny i ujednolicony dwa zagadnienia:

- dopasowanie sekwencji biologicznych - wiedza na ten temat jest na tyle ugruntowana, że zawiera ją niemal dowolny podręcznik do podstaw bioinformatyki,
- asemblację - jest dostępnych kilka prac przeglądowych, warte odnotowania są także publikacje podsumowujące konkursy Assemblathon.

Mam wrażenie, że oparcie sekcji dotyczących powyższych zagadnień na tego rodzaju źródłach mogłoby uchronić rozprawę przed zastrzeżeniami wymienionymi powyżej.

5. Inne uwagi

- Definicja 2.4.1 jest zupełnie zbędna – lepiej ją usunąć, a wystąpienia definiowanego pojęcia *złożenie zbioru sekwencji* zastąpić zwrotem *pewna sekwencja*.
- Zmienna *C'* w pseudokodzie 5.1.1 nie jest zainicjalizowana, a jej znaczenie nie jest wyjaśnione w tekście, w szczególności na liście na górze strony 91 brakuje odpowiedniej etykiety.
- W opisie następującym po formule (5.1.6) brakuje wyjaśnienia, jak wartości $w_{k,s}$ zależą od czasu dodania odczytu *s*.
- Wartości w tabelach 7.5.1 i 7.5.2 zdają się sugerować, że scaffold'y są identyczne z kontigami – warto to sprawdzić i skomentować.
- W testach w sekcji 7.7 nie opisano ani testowego zbioru danych, ani wartości parametrów nietestowanych w poszczególnych eksperymentach.
- Na stronie 133 jest napisane, że precyzja charakterystyk *k*-merowych wynosi 98% dla okna rozmiaru 1000, tymczasem na stronie 78 napisano, że do takiej precyzji wystarcza okno rozmiaru 60.

6. Podsumowanie

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez artykuł 13 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym (z późniejszymi zmianami) moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problem naukowego?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka lub Automatyka i Robotyka?

[ ]	[X]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat umiejętność samodzielnego prowadzenia pracy naukowej?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

N. Dojer

Podpis