Streszczenie rozprawy doktorskiej Wojciecha Frohmberga pt. “GRASShopPER - wydajna metoda asemblacji *de novo* wykorzystująca strategię Overlap-Layout-Consensus”

1. Wstęp

Asemblacja DNA stanowi jeden z kluczowych elementów złożonego procesu sekwencjonowania materiału genetycznego. Istotą procesu asemblacji jest rekonstrukcja możliwie długiej sekwencji nukleotydów materiału genetycznego przy wykorzystaniu jego krótkich lecz silnie, przy tym nierównomiernie, nakładających się fragmentów. Fragmenty podlegające takiej rekonstrukcji, określane mianem *odczytów* (ang. *reads*), z powodu ograniczeń technologicznych nie mogą być wydłużone na etapie mechanicznego odczytu, a ich długość jest nieporównywalnie mniejsza od długości rekonstruowanej sekwencji. Proces asemblacji, przy obecnym stanie techniki stanowi zatem jedyny możliwy automatyczny sposób uzyskania sekwencji odczytywanych genomów w postaci ciągłej.

Z algorytmicznego punktu widzenia asemblacja stanowi uogólnienie problemu znajdowania minimalnego superciągu nad zadanym zbiorem sekwencji. Problem ten, nawet przy ograniczeniu do alfabetu binarnego, jest trudny obliczeniowo. Co więcej, jak dowiedziono w pracy [1], przy założeniu możliwych błędów w odczytach problem asemblacji jest silnie NP-trudny, a co za tym idzie opiera się możliwości utworzenia w pełni wielomianowego schematu aproksymacji (ang. *fully polynomial-time approximation scheme*, *FPTAS*). Stąd jedyne dostępne efektywne metody rozwiązujące problem bazują na założeniach heurystycznych bez oszacowania pesymistycznego błędu wartości rozwiązania. Sytuację dodatkowo pogarszają rozmiary realnych instancji problemu liczone w setkach milionów lub nawet miliardach odczytów. Żeby rozwiązać tak duże instancje problemu większość dostępnych na rynku metod asemblacji korzysta z heurystycznej strategii grafów dekompozycji (DG, często nadużywając nazwy grafów de Bruijna). Strategia ta

efektywność działania zawdzięcza wykorzystywanej strukturze danych kompresującej niejako odczyty, a właściwie ich krótkie fragmenty zwane k-merami, stanowiące podciągi sekwencji odczytów o zadanej długości. Proces tworzenia k-merów z sekwencji odczytów, zwany dekompozycją, jest stratny i nieodwracalny. Dzięki niemu jednak otrzymujemy efektywną strukturę grafową, która pomimo utraty części informacji jest w stanie, przy pewnych założeniach posłużyć do efektywnego odtworzenia dłuższych podciągów sekwencji wyjściowej. Istotne z punktu widzenia jakości otrzymanego rozwiązania jest to, iż uzyskaną efektywność uzyskuje się kosztem utraty części informacji na temat ciągłości odczytów. W przypadku krótkich odczytów będących wynikiem sekwencjonowania maszynami pierwszej generacji nie stanowiło to znaczącego ubytku informacji przekładającego się na stratę na jakości wyników, jednak wraz z rozwojem technologiii i wzrostem długości odczytów tracona informacja nt. ciągłości jest coraz większa i mniej akceptowalna.

Nowy potencjał rozwiązania problemu niesie strategia będąca poprzednikiem DG zwana Overlap-Layout-Consensus (OLC). Strategia ta sprowadza problem asemblacji do problemu komiwojażera nadając wierzchołkom etykiety sekwencji odczytów, łukom zaś przypisując tak zwane nałożenia (ang. overlap) między tymi sekwencjami. Wykorzystuje ona przy tym dość naturalne spostrzeżenie, iż koszty wynikające z nałożeń optymalnego uszeregowania (ang. layout) odczytów rozwiązania problemu asemblacji powinny być najmniejsze spośród wszystkich możliwych uszeregowań i wynikają jedynie z błędów sekwencjonowania. Jako że nie interesuje nas wynik w postaci sekwencji odczytów, ale raczej w postaci ciągu nukleotydów znalezione uszeregowanie odczytów musi być jeszcze spłaszczone poprzez ustalenie kompromisowych nukleotydów (ang. consensus).

2. Cel i zakres pracy

Głównym celem pracy jest stworzenie heurystycznego algorytmu asemblacji *de novo* o jak najlepszych właściwościach jakościowych dostarczającego aproksymację rozwiązania problemu asemblacji w akceptowalnym czasie. Naturalnym zatem wydaje się skorzystanie ze strategii Overlap-Layout-Consensus, która nie dopuszcza błędów wynikających z zaniedbania ciągłości sekwencji odczytów w trakcie ich

dekompozycji do struktury grafowej. Celem pochodnym jest również implementacja metody zwanej dalej GRASShopPER (ang. *GPU overlap GRaph ASSembler using Paired End Reads information for de novo assembly*) na podstawie zaproponowanego algorytmu oraz przeprowadzenie testów na zróżnicowanych zestawach danych włącznie z porównaniem wyników jakościowych oraz efektywnościowych z innymi metodami asemblacji będącymi aktualnym standardem.

Głównym wkładem doktoranta wykazywanym w dysertacji jest trzon nowej metody GRASShopPER. W tym również dwa algorytmy rozwiązujące podetapy procesu asemblacji. We wspomnianym kontekście pracę można podzielić na cztery główne części:

1. Formalna definicja problemu wraz z przedstawieniem tła badań
2. Omówienie aktualnego stanu wiedzy w tym użytych zewnętrznych metod niezbędnych w realizacji projektu
3. Określenie wprowadzonych przez doktoranta algorytmów z omówieniem ich złożoności i oczekiwanego wpływu na wynik asemblacji
4. Przeprowadzenie eksperymentów obliczeniowych oraz zebranie i omówienie wyników

W przedstawianej pracy autor skupił się na swoich autorskich pomysłach z przedstawieniem szerszego kontekstu ich aplikacji.

3. Podstawy teoretyczne oraz omówienie problemu

Na początku pracy doktorant wprowadza najistotniejsze, użyte w późniejszej części pracy, pojęcia, zarówno z zakresu terminologii biologicznej, jak i z pogranicza matematyki i informatyki (np. teorii grafów, ciągi czy rząd funkcji). Dokonuje również wprowadzenia w tematykę złożoności obliczeniowej niezbędnej do określenia klasy trudności rozwiązywanego przez niego problemu. W kolejnych podrozdziałach znajdziemy również niezbędne do opisu problemu asemblacji uściślenie pojęć dopasowania i złożenia sekwencji. Po wspomnianym wstępie następuje przybliżenie problemu asemblacji. Dokonane zestawienie pojęcia asemblacji z

sekwencjonowaniem płynnie przechodzi w zarys kontekstu problemu asemblacji. Można tutaj odnaleźć uściślenie terminów związanych z sekwencjonowaniem, ale także parametrami charakteryzującymi dane wejściowe problemu. Zwieńczeniem opisu problemu jest jego formalna definicja stanowiąca próbę jego obrazowego przedstawienia. Formalizacji dokonano poprzez sprowadzenie asemblacji do postaci optymalizacyjnego problemu kombinatorycznego. Obok definicji znajdziemy tutaj dyskusję na temat możliwych strategii rozwiązania problemu jak również określenie i charakterystykę konkurencyjnych metod asemblacji. Po analizie wspomnianych metod doktorant dochodzi do wniosku, iż w miejscu efektywnych metod wykorzystujących strategię OLC istnieje nisza w rynku asemblerów, co prowadzi go dalej do konkluzji, iż praca nad tym tematem może stanowić wartościowy przedmiot badawczy.

4. Metoda GRASShopPER

Wspomniane wnioskowanie stanowi podwaliny projektu GRASShopPER. W swojej dysertacji doktorant poświęcił obszerną, składającą się z trzech rozdziałów, część opisowi tej nowej metody asemblacji. Ze względu na rozległość przedsięwzięcia, projekt stworzony był etapami przez wielu autorów. Niezbędne zatem było precyzyjne ustalenie wkładu intelektualnego doktoranta w samą metodę ale także w poszczególne fazy metody. Pierwszy ze wspomnianych rozdziałów skupia się właśnie na precyzyjnym ustaleniu autorstwa poszczególnych fragmentów metody, stąd też poświęcona jest tu również spora uwaga na przedstawieniu ogólnego schematu działania metody ze szkicem działania podprocedur oraz przepływem danych. W dalszej części rozdziału następuje wnikliwy opis poszczególnych części metody, jednak z wyłączeniem tych części, które stanowią wkład autora.

Kolejne dwa rozdziały przedstawiają dwa algorytmy zaproponowane przez doktoranta:

1. Algorytm znajdowania znaczących rozwidleń grafu nałożeń
2. Algorytm progresywnego dopasowania sekwencji do znajdowania kompromisowej sekwencji wynikowej

Żeby przybliżyć pierwszy z wymienionych algorytmów warto najpierw umotywować potrzebę jego stosowania. Specyfika grafu nałożeń, w tym w szczególności sam koncept nałożeń, które łączą wszystkie możliwe wierzchołki/odczyty o podobnej podsekwencji, wywołuje redundancję możliwych ścieżek formujących identyczne sekwencje wynikowe. Z drugiej strony, dobrej jakości wyniki asemblacji powinny cechować się minimalną niepewnością użytych fragmentów. Wraz ze wzrostem ilości repetytywnych fragmentów genomu źródłowego przy braku możliwości automatycznej metody potwierdzania poprawności tworzonej sekwencji wzrasta ryzyko złożenia niepoprawnego wyniku. Stąd też jedynym sposobem radzenia sobie z niepewnością jest przerywanie dołączania odczytów w momencie pojawienia się alternatywnych ścieżek. Jednakowoż szukanie początków alternatywnych ścieżek, nazwane dalej szukaniem rozwidleń znaczącymi grafu nałożeń, przy możliwych błędach w odczytach jest nietrywialnym problemem. Trudność jest tym większa że realne instancje problemu posiadają gigantyczny graf nałożeń. Z punktu widzenia kwestii efektywnościowych najlepszym rozwiązaniem mogłaby być zatem heurystyka dokonująca optymalizacji jeszcze na etapie przechodzenia grafu, dla przykładu algorytm zachłanny, wyłaniający do wyniku sekwencje odczytów o najmniejszym aktualnym nałożeniu. Algorytm ten byłby jednak obarczony możliwością nieprawidłowego złożenia sekwencji, co wynika ze wspomnianej własności grafów nałożeń. Stąd zaproponowany algorytm wykorzystuje zachłanne podejście przechodzenia grafu obudowując je dodatkowo mechanizmem stanu zabezpieczającym przed zagłębianiem się w ścieżkę odpowiadającej sekwencji nukleotydowej niosącej ryzyko występowania w alternatywnej lokalizacji. Mechanizm ten wzbogacony o dobór następnika w ważonym głosowaniu, filtruje ślepe ścieżki wynikające z błędów sekwencjonowania.

Warto zwrócić uwagę, iż przechodząc po grafie nałożeń w poszukiwaniu znaczących rozwidleń znajdujemy sekwencję odczytów a nie nukleotydów sekwencjonowanego genomu, czego oczekuje się od algorytmu asemblacji. Stąd też w przypadku wszystkich metod z rodziny OLC niezbędne jest przeprowadzenie dodatkowego kroku spłaszczającego sekwencję odczytów do konsensusowego ciągu nukleotydów. Innymi słowy do uzyskania finalnego wyniku niezbędny będzie

algorytm rozwiązujący problem dopasowania sekwencji. Dopasowanie to musi jednak uwzględnić dość specyficzne warunki - znana jest zgrubna pozycja sekwencji odczytów, z których składamy nadsekwencję. Ta właściwość nie trywializuje jednak problemu, gdyż wciąż sekwencje odczytów mogą posiadać błędy, które efektywnie można zniwelować (oczywiście pod warunkiem odpowiednio dużej duplikacji informacji w odczytach) jedynie za pomocą algorytmu dynamicznego programowania. Warto zwrócić uwagę, że standardowe algorytmy dopasowania nie uwzględniają porządku wplatania kolejnych sekwencji do wyniku. Stąd też niezbędnym było zaproponowanie nowego algorytmu dopasowania zwanego w pracy dopasowaniem progresywnym.

W skrócie działanie algorytmu opiera się o zasadę ─ skoro znamy zgrubną pozycję sekwencji w rozwiązaniu wymuśmy umieszczenie sekwencji w pobliżu tej pozycji. Można tego dokonać poprzez ograniczenie przestrzeni przeszukiwanych komórek macierzy dynamicznego programowania określających punkt startowy wynikowej sekwencji. Powyższa technika wspomagana profilami nukleotydów zapewnia heurystykę budowy konsensusu rozwiązania, która jest odporna na zaburzenia w jakości odczytów instancji problemu asemblacji.

5. Eksperymenty obliczeniowe

Ostatni przed podsumowaniem, lecz nie najmniej istotny, rozdział doktoratu poświęcony jest eksperymentom obliczeniowym przeprowadzonym w celu ewaluacji jakości wyników metody GRASShopPER. Rozdział można podzielić na dwie zasadnicze części:

1. Zestawienie wyników jakościowych metody GRASShopPER z innymi wiodącymi metodami asemblacji.
2. Określenie wpływu wartości parametrów na jakość wyników działania metody GRASShopPER.

Nieodzownym elementem rzetelnej analizy wyników w przypadku każdych eksperymentów obliczeniowych jest precyzyjne określenie zbiorów danych testowych lub ewentualnie sposobu ich wygenerowania. W przypadku problemu

asemblacji *de novo* pojedyncza instancja zbioru testowego składa się z informacji na temat zestawu odczytów sparowanych oraz genomu referencyjnego służącego do porównania wyników działania algorytmu asemblacji. Notę na temat tego aspektu eksperymentów jak również informację na temat użytych miar oceny, środowiska uruchomieniowego i zużywanych przez program zasobów można znaleźć na początku wspomnianej części pracy.

5.1. Porównanie z wiodącymi metodami asemblacji

Eksperymenty porównawcze GRASShopPERa rozpoczyna uzasadnienie doboru konkurencyjnych metod asemblacji. Wybór ten będzie miał oczywisty wpływ na sprawiedliwość osądu jakości metody. Doktorant postanowił się tutaj oprzeć na obiektywnej wartości wskaźnika cytowalności artykułów, w których opublikowano poszczególne metody asemblacji. Korzystając z powyższego klucza wyłoniono sześć metod asemblacji:

- Velvet
- SPAdes
- SOAPdenovo2
- Celera
- SGA
- Platanus

Każdy z programów ze wspomnianego powyżej zestawu posłużył do uruchomienia eksperymentów obliczeniowe na zadeklarowanych trzech rzeczywistych zbiorach testowych. Jednocześnie starano się dobrać parametry programów (jeśli to było możliwe) tak by uzyskać jak najwyższe jakości wyników. Otrzymane wyniki zebrano w postaci tabelarycznej i opatrzono komentarzem.

Głównym wnioskiem z zaprezentowanych wyników jest to, iż spośród wymienionych metod podlegających porównaniu można wyróżnić trzy, które dominują nad pozostałymi. Są to GRASShopPER, SGA oraz SOAPdenovo2. Ciekawymi właściwościami (w stosunku do pozostałych metod) charakteryzuje się również program SPAdes tworząc bardzo długie kontigi pokrywające dużą część genomu

referencyjnego. Oprócz tego wyniki cechują się jednak dużym współczynnikiem błędów złożenia stąd nie zostały uwzględnione w fazie scaffolding'u.

5.2. Wpływ parametrów odczytów na jakość wyników

W tym kontekście doktorant postanowił zweryfikować wpływ zmienności parametrów odczytów takich jak pokrycie genomu referencyjnego oraz odsetek błędów sekwencjonowania na wyniki asemblacji metodą GRASShopPER. Niestety, co oczywiste, nie było dostępnych rzeczywistych danych obejmujących wspomniany zakres zmienności charakterystyki odczytów. Jednakowoż istnieją programy symulujące działanie sekwenatorów cechujące się bardzo zbliżonymi właściwościami produkowanych odczytów. Doktorant postanowił wykorzystać jeden z dostępnych programów do celu generowania odczytów rzeczywistego genomu referencyjnego organizmu modelując pożądaną cechę instancji za pomocą parametrów tegoż generatora. Zmienność parametrów jakościowych wyników programu GRASShopPER przedstawiono w postaci wykresów.

Główne wnioski ze wskazanych eksperymentów wskazują na istnienie dodatniego wpływu zwiększania pokrycia na długości kontigów generowanych przez metodę asemblacji. Począwszy jednak od pewnej wartości tegoż parametru (pokrycie powyżej 30x) pogarszają się niektóre wartości wskaźników jakości takie jak poziom duplikacji czy sumaryczna długość błędów złożeń. Jeśli chodzi z kolei o wpływ odsetka błędów w odczytach na wyniki jakościowe, wyniki wskazują dość naturalną negatywną zależność przyrostu błędów w danych na długości kontigów ale i pozostałe wskaźniki jakościowe wyników. Optymalna z punktu widzenia jakości danych wyjściowych miara jakości odczytów to przynajmniej phred24.

5.3. Dobór parametrów wejściowych programu GRASShopPER

Wyznaczenie najlepszych wartości dla niektórych parametrów metody GRASShopPER w sposób automatyczny może być bardzo trudne lub wręcz niemożliwe. W zależności od charakterystyki zbioru danych może okazać się więc konieczne dobranie innych niż domyślne wartości parametrów. Ostatni podrozdział eksperymentów obliczeniowych ma na celu opis najważniejszych parametrów

metody oraz ukazanie wpływu zmiany ich wartości na wyniki jakościowe kontigów, tak by użytkownik metody mógł wyrobić sobie pogląd jak dostosować wartości parametrów do swoich potrzeb. Parametry ilościowe, których wpływ na jakości wyników postanowiono przetestować to:

- wielkość okna sąsiedztwa charakterystyk k-merowych,
- długość podsekwencji minimalnego indeksu leksykograficznego,
- liczba dopuszczonych błędów w dopasowaniu,
- poziom tolerancji na zmniejszenie listy kandydatów.

Wyniki, podobnie jak w poprzednim zestawie eksperymentów, postanowiono zebrać w postaci wykresów zależności poszczególnych parametrów na wskaźniki jakościowe.

Jednym z ciekawszych wniosków, które można było wysnuć ze wspomnianych eksperymentów jest odporność metody na zmniejszanie okna sąsiedztwa charakterystyk k-merowych oraz zwiększanie długości podsekwencji minimalnego indeksu leksykograficznego w kontekście jakości kontigów. Warto zwrócić uwagę, że parametry te mają znaczący wpływ na efektywność działania metody. Innym ciekawym spostrzeżeniem może być to że metoda zwraca najlepsze wyniki w wypadku dopuszczenia dokładnie jednego błędu w nałożeniu dwóch odczytów. Jeśli chodzi o wpływ poziomu tolerancji zmniejszenia listy kandydatów na wartości jakościowe to parametr ten pozwala na wyważenie kompromisu pomiędzy długością kontigów a możliwymi błędami złożenia kontigów.

6. Podsumowanie

W czasach, gdy niemała część gatunków ma już całkiem dobrze poznane genomy mogłoby się nasunąć pytanie - “Czy na pewno wciąż potrzebujemy metody asemblacji *de novo*, skoro istnieje alternatywne podejście asemblacji przez mapowanie?”. Otóż warto przypomnieć, iż każdy organizm ma unikalny genom - różniący się w sposób mniej lub bardziej znaczący nawet od innych przedstawicieli tego samego gatunku. Różnice, które mogą tu występować to zarówno polimorfizmy pojedynczego nukleotydu (SNP), ale przybierają również postać różnych całych

podsekwencji genomu (ang. *polymorphic variants*). W tym drugim przypadku zdarza się, iż odczyty danego organizmu nie mają swoich odpowiedników lub są odnajdywane w niewłaściwych pozycjach genomu wynikiem polimorfizmu translokacji. Jako że znany genom referencyjny jest silnym bodźcem sugestywnym, wynik metody asemblacji może być tu tendencyjny (ang. *biased*). Powyższe rozumowanie przybliża nas do stwierdzenia, iż ze względu na swą brak wykorzystania genomu referencyjnego metody asemblacji *de novo* mogą lepiej nadawać się do odnajdywania polimorficznych wariacji w genomie organizmu.

Tematem streszczanej pracy jest GRASShopPER, tj. nowatorska metoda asemblacji *de novo* wykorzystująca w sposób hybrydowy technologię kart graficznych. Atutem metody jest przywrócenie do życia strategii Overlap-Layout-Consensus w swojej (na ile było to możliwe) pierwotnej formule. Strategia ta ma bez wątpienia wielki potencjał wraz z nowo-powstającymi technologiami sekwenatorów trzeciej generacji produkującymi coraz dłuższe odczyty. Konkurencyjna strategia grafów dekompozycji w tym wypadku traciłaby informację o ciągłości sekwencji zawartej w takich odczytach a w wypadku długich sekwencji byłaby to dość dokuczliwa strata. W tym miejscu warto zaznaczyć, iż opisana metoda nie odnosi się w żaden sposób do wspomnianych powyżej technologii a jedynie stanowi próbę dowodu, iż jesteśmy w stanie wykorzystać efektywnie strategię OLC, podwyższając przy tym jakość asemblacji.

Warto zaznaczyć, iż mimo zakończenia pewnego etapu, zwieńczonego publikacją [2], proces prac nad metodą GRASShopPER jeszcze się nie zakończył. Wraz z kolejnymi projektami planowane jest wykorzystanie GRASShopPERa do asemblacji *de novo* całego człowieka, co bez wątpienia na obecnym etapie stanowi duże wyzwanie. Do osiągnięcia tego celu niezbędne będą bez wątpienia dalsze optymalizacje metody.

Literatura

[1] Jacek Blazewicz and Marta Kasprzak. "Reduced-by-matching graphs: toward simplifying Hamiltonian circuit problem." *Fundamenta Informaticae* 118.3 (2012): 225-244.

[2] Aleksandra Swiercz, Wojciech Frohmberg, Michal Kierzynka, Pawel Wojciechowski, Piotr Zurkowski, Jan Badura, Artur Laskowski, Marta Kasprzak, and Jacek Blazewicz. "GRASShopPER – An algorithm for de novo assembly based on GPU alignments." *PLOS ONE*, 13(8):e0202355, 2018.