Michał Kierzynka
Politechnika Poznańska, Instytut Informatyki
Stypendysta projektu pt. „Wsparcie stypendialne dla doktorantów na kierunkach uznanych za strategiczne z punktu widzenia rozwoju Wielkopolski", Poddziałanie 8.2.2 Programu Operacyjnego Kapitał Ludzki

Dopasowanie sekwencji w problemie asemblacji DNA

Mimo, iż pierwsze próby sekwencjonowania DNA pojawiły się na świecie już pod koniec lat 70. XX wieku, do dziś nie opracowano uniwersalnej metody na odczytywanie genomu organizmów. Problem leży nie tylko w części biochemicznej eksperymentu (odczytywaniu pojedynczych nukleotydów, czyli sekwencjonowaniu) ale również w składaniu odczytanych fragmentów genomu w całość, a zatem w opracowaniu dobrego algorytmu rozwiązującego problem asemblacji *de novo*. Algorytmy do asemblacji DNA typu *de novo* są niezbędne, ponieważ istniejące technologie sekwencjonowania pozwalają odczytywać jedynie bardzo krótkie odcinki materiału genetycznego, które następnie trzeba złożyć w całość, podobnie jak puzzle.

Minione lata przyniosły bardzo szybki rozwój obu wspomnianych gałęzi: zarówno sekwencjonowania, jak i asemblacji. Obecnie na rynku istnieje kilku producentów maszyn służących do sekwencjonowania – tzw. sekwenserów. Niektóre z tych urządzeń są w stanie dostarczyć dodatkowych informacji na temat odczytywanych sekwencji DNA, np. o tzw. odczytach sparowanych, co stanowi cenną informację w procesie asemblacji. Z drugiej strony obserwowany wzrost mocy obliczeniowej jak i ilości pamięci operacyjnej w komputerach umożliwia stosowanie coraz bardziej zaawansowanych algorytmów asemblacji jak i mapowania. Niestety współczesne algorytmy tylko szczątkowo wykorzystują informacje o wspomnianym sparowaniu sekwencji pozostawiając tym samym duże pole do popisu badaczom. Ta niezwykle skrócona historia prowadzi nas do mojej pracy doktorskiej, której celem jest opracowanie efektywnego algorytmu do analizy sekwencji DNA w celu ich asemblacji z uwzględnieniem informacji pochodzącej z odczytów sparowanych.

Praca doktorska współfinansowana ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

Hipotezą stawianą w mojej pracy jest przydatność tzw. charakterystyk k-merowych sekwencji przy konstrukcji grafu DNA służącego do asemblacji. Wspomniane k-mery już od wielu lat stosowane są z powodzeniem do analizy sekwencji, nie tylko biologicznych. Innowacją w mojej pracy jest sposób, w jaki wykorzystuje je do obliczania podobieństwa między sekwencjami, dzięki czemu możliwe staje się przetwarzanie większych zbiorów danych w krótszym czasie. Innymi słowy, możliwa staje się asemblacja organizmów z dłuższymi genomami. Ponadto w swoim algorytmie wykorzystuję również wiedzę wynikającą z odczytów sparowanych, zwiększając tym samym dokładność działania zaproponowanej metody.

Kluczowym z punktu widzenia metodologii wspomnianych badań jest dostęp do danych pochodzących z realnych eksperymentów. Europejskie Centrum Bioinformatyki i Genomiki (ECBiG), mające swoją siedzibę w Poznaniu, posiada w swoim wyposażeniu nowoczesną aparaturę do sekwencjonowania, dzięki czemu możliwy staje się dostęp do takich danych. Sekwencjonowane organizmy cechują się występowaniem tzw. regionów repetytywnych w swoim genomie, czyli powtarzających się odcinków DNA, które wprowadzają szczególne utrudnienie podczas procesu asemblacji. Druga istotna cecha takich danych to błędy w odczytach pojedynczych nukleotydów, które wymuszają stosowanie algorytmów odpornych na takie błędy. Dzięki korzystaniu danych pochodzących z realnego sekwenatora już od najwcześniejszej fazy projektowania jak i pisania algorytmu możliwe stało się obranie poprawnego podejścia oraz zastosowanie adekwatnych technik.

Asemblacja łańcuchów DNA jest niewątpliwie problemem o dużej złożoności obliczeniowej. Oznacza to w praktyce, że potrzeba dużo czasu na złożenie odczytanych fragmentów DNA w genom. Proces ten można znacznie przyspieszyć korzystając z nowoczesnych narzędzi, które umożliwiają programowanie masywnie równoległe. W tym celu w swojej pracy wykorzystuję karty graficzne (ang. Graphics Processing Unit, GPU). Jest to być może trochę nieintuicyjne, ponieważ głównym zadaniem GPU jest generowanie obrazu, który następnie jest wyświetlany na monitorze komputera. Natomiast urządzenia te można również wykorzystywać do obliczeń równoległych niezwiązanych z grafiką komputerową. W dziedzinie dopasowywania sekwencji (DNA oraz aminokwasowych) na kartach graficznych odniosłem kilka sukcesów, które zaowocowały publikacjami w międzynarodowych czasopismach naukowych. Obecnie prowadzę również badania nad potencjalnym zastosowaniem innych równoległych architektur obliczeniowych, np. FPGA, do dopasowywania sekwencji oraz asemblacji.

Praca doktorska współfinansowana ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

Istnieje wiele praktycznych zastosowań opisywanej przeze mnie tematyki. Jednym z nich jest tzw. innowacyjne rolnictwo. Badania genetyczne prowadzone obecnie w tym zakresie skupiają się na takich problemach jak odporność uprawianych roślin na opryski czy szkodniki, bądź zwiększenie plonów. Efekty tego typu badań przekładają się na wymierne korzyści gospodarcze płynące ze zwiększenia konkurencyjności gospodarstw, które korzystają z tego typu innowacji. Co więcej, postęp jaki dokonuje się w genetyce, ujawnia coraz to większe jej znaczenie w wielu chorobach, np. w przypadku nowotworów takich jak rak piersi czy jajników, rak jelita grubego, prostaty, nerki, tarczycy czy też płuc. Obecnie szacuje się, że około 30% nowotworów złośliwych powstaje na skutek predyspozycji genetycznych, co stanowi jeden z najbardziej istotnych czynników ryzyka. Podczas moich badań nad asemblacją DNA typu *de novo* okazało się, że metoda, którą zaproponowałem, może być z powodzeniem użyta również do mapowania odczytów DNA na tzw. genom referencyjny. Narzędzia tego typu są podstawą resekwencjonowania, czyli składania genomów organizmów, dla których znany jest już tzw. genom referencyjny, włączając w to genom ludzki. Mam nadzieję, że zaproponowana metoda przyczyni się do poprawy jakości generowanych wyników, szczególnie w kontekście kompleksowych badaniach genetycznych człowieka. Obecnie badania genetyczne opierają się w znacznym stopniu na tzw. mikromacierzach, które pozwalają na zbadanie wystąpienia konkretnej, znanej obecnie mutacji wybranego genu. Inaczej rzecz ma się z sekwencjonowaniem całego genomu. Gdy człowiek jednokrotnie pozna swój genom, wówczas można wykonać na nim badania pod kątem wszystkich znanych chorób genetycznych. Co więcej, badania te można powtórzyć za kilka lat, gdy znane będą inne kombinacje mutacji genów powodujące choroby, przy czym kosztownego eksperymentu sekwencjonowania nie trzeba powtarzać. Ponadto warto nadmienić, że nawet w przypadku organizmów ze znanym genomem w niektórych przypadkach stosuje się asemblację typu *de novo*, np. gdy zachodzi potrzeba odtworzenia mocno „zniekształconego” przez nowotwór fragmentu genomu.

Współcześnie dzięki sekwencjonowaniu i asemblacji poznanie genomu bakterii bądź wirusa grożącego pandemią nie zajmuje już lat, lecz zaledwie tygodnie. Ma to oczywisty związek z bardzo dynamicznym rozwojem metod obliczeniowych, którymi się zajmuję. Warto podkreślić, że szerokie spektrum zastosowań pracy, którą opisuję, wynika z faktu, iż moje badania nie skupiają się na pojedynczej chorobie czy analizie konkretnych mutacji, lecz na opracowaniu jak najbardziej uniwersalnego podejścia do asemblacji całych genomów DNA.

Praca doktorska współfinansowana ze środków Unii Europejskiej w ramach
Europejskiego Funduszu Społecznego