Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
Informatyka Techniczna i Telekomunikacja

Rozprawa doktorska

Korekcja danych z sekwencjonowania genomów

Maciej Długosz

Promotor: prof. dr hab. inż. Sebastian Deorowicz

Gliwice, 2023

Korekcja danych z sekwencjonowania genomów

Streszczenie

Maciej Długosz
Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
Katedra Algorytmiki i Oprogramowania

Promotor: prof. dr hab. inż. Sebastian Deorowicz

W pracy poruszono temat korekcji odczytów uzyskanych w wyniku sekwencjonowania genomów urządzeniami marki Illumina. Do tej pory opracowano liczne algorytmy, których zadaniem jest detekcja oraz eliminacja błędów obecnych w danych sekwencjonowania. Skuteczność działania oraz zapotrzebowanie na zasoby obliczeniowe takich algorytmów różni się znacząco, rodząc trudności w doborze odpowiedniego do wybranego zagadnienia. Jednocześnie istniejące prace przeglądowe nie dają pełnego obrazu, wspomagającego taki wybór.

Opierając się na analizie cech istniejących algorytmów, zaproponowano nowy algorytm korekcji RECKONER. W ramach prac przedstawiono wymagania, jakie powinien spełniać, oraz opracowano zasadę jego działania. Algorytm został wyposażony w rozwiązania, pozwalające na poprawę skuteczności korekcji: korekcję błędów typu indel, strategię doboru wynikowego rozwiązania z zestawu wielu możliwości, weryfikację wyników w oparciu o oligomery dwóch długości. Wprowadzono nową metodę zliczania k-merów oraz wykorzystano strukturę danych generowaną przez narzędzie KMC. Ponadto wprowadzono automatyczną, opartą na danych empirycznych metodę doboru głównego parametru oraz wyznaczono złożoność obliczeniową algorytmu. Postawiono ograniczenia dotyczące liczby rozpatrywanych przez algorytm przypadków, skracając czas obliczeń i minimalizując zapotrzebowanie na pamięć operacyjną. Algorytm zapewnia wykorzystanie przetwarzania równoległego.

Algorytm RECKONER został poddany eksperymentom, mającym na celu ocenę skuteczności oraz obserwację zapotrzebowania na zasoby obliczeniowe. Dokonano analizy porównawczej z grupą konkurencyjnych algorytmów. Oparto się przy tym na zestawach odczytów symulowanych komputerowo przy pomocy dwóch metod, weryfikując liczbę wyeliminowanych błędów, oraz oceniono potencjał takich odczytów w ocenie algorytmów korekcji. Ponadto w eksperymentach wykorzystano liczne zestawy odczytów z rzeczywistych procesów sekwencjonowania DNA, obserwując wpływ korekcji na jakość przeprowadzonych w dalszej kolejności zadań asemblacji *de novo* i mapowania odczytów. Podobnej analizy dokonano opierając się na autorskiej metodzie oceny przy pomocy obserwacji wpływu korekcji na detekcję wariantów genomów. W eksperymentach dokonano pomiaru czasu, zapotrzebowania na pamięć oraz skalowalności algorytmów.

W wyniku wykonanych eksperymentów stwierdzono ograniczoną użyteczność oceny korekcji w oparciu o symulowane odczyty i konieczność dalszych prac nad wykorzystaniem w tym celu potoków detekcji wariantów genomów. Po-

nadto zaproponowano grupę najlepszych algorytmów korekcji, które są zalecane do wykorzystania w ramach prowadzonych prac związanych z przetwarzaniem odczytów sekwencjonowania genomów.