Algorytmy składania genomów umożliwiające diagnostykę zespołów genetycznych

Ustalenie kolejności nukleotydów, które stanowią podstawę przepływu informacji genetycznej, jest konieczne do zrozumienia aktywności biologicznej każdej żywej komórki. Proces określania tej kolejności nazywamy sekwencjonowaniem. Obecnie możemy sekwencjonować jedynie fragmenty DNA. Fragmenty takie nazywamy odczytami, a ich właściwości zależą od użytej techniki sekwencjonowania. Asemblacja to proces rekonstrukcji genomu na podstawie odczytów. Na algorytmiczną złożoność problemu asemblacji genomu mają wpływ czynniki związane charakterystyką odczytów, takie jak długość i dokładność oraz charakterystyka genomu, głownie dotycząca występowania regionów powtarzających się.

Obecne na rynku sekwenatory można podzielić na dwie grupy: odczytujące krótkie i dokładne fragmenty DNA długości od 50 do 500 nukleotydów oraz odczytujące dłuższe fragmenty od tysiąca do kilkuset tysięcy nukleotydów kosztem ich dokładności. Sekwenatory odczytujące krótkie sekwencje DNA pojawiły się wcześniej i mają wiele zastosowań w biologii i medycynie. Jednak charakterystyki tej technologii, dotyczące między innymi ograniczonej możliwości mapowania elementów repetytywnych i punktów złamań rearanżacji chromosomowych, pozostawiają istotną cześć ludzkiego genomu niedostępną dla analiz biologicznych czy medycznych. Sekwenatory odczytujące długie fragmenty DNA zostały wprowadzone by sprostać problemom asemblacji złożonych, bądź repetytwnych regionów.

Niemniej jednak, nawet najnowocześniejsze technologie nie pozwalają na odpowiedź na żaden biologiczny bądź medyczny problem bez odpowiedniej bioinformatycznej obróbki dostarczanych przez nie danych oraz narzędzi umożliwiających interpretację otrzymanych wyników. Obecnie dane z długich odczytów są wykorzystywane głównie w badaniach naukowych, lecz posiadają wielki potencjał do zastosowań klinicznych, zwłaszcza gdy zostaną wzbogacone o wiedzę ekspercką o regionach genomu istotnych z punktu widzenia medycznego.

W projekcie planujemy skupić się na dwóch problemach z ważnymi zastosowaniami klinicznymi, które mogą być rozwiązane z użyciem sekwencjonowania długimi odczytami:

- określenie architektury złożonych wariantów strukturalnych (translokacji, insercji, delecji, duplikacji lub inwersji) mogących obejmować więcej niż dwa chromosomy;
- ustaleniu sekwencji złożonych, repetytywnych regionów zawierających tak zwane segmentalne duplikacje.

Planujemy również implementację algorytmów i stworzenie narzędzi pozwalających na kliniczną interpretację otrzymanych wyników.

W projekcie chcemy opracować algorytmy, które w oparciu o istniejące narzędzia pozwalające na określenie miejsc złamań za pomocą danych z sekwencjonowania długich odczytów, stworzą graf reprezentujący strukturę chromosomów po rearanżacjach. Algorytmy te będą korzystały z danych opisujących regiony repetytywne oraz baz danych zawierających wariacje strukturalne obecne w populacji a także powodujące choroby genetyczne. Stworzymy również narzędzia pozwalające na określenie patogenetyczności znalezionych wariantów strukturalnych.

Drugim celem tego projektu jest opracowanie algorytmów pozwalających na określenie sekwencji nukleotydów zawierających segmentalne duplikacje (rejony zawierające samopodobne fragment DNA, które występują w ludzkim genomie w więcej niż jednej kopii). W ramach projektu zajmiemy się problemem znajdowania najbardziej prawdopodobnego scenariusza rearanżacji genomowej mediowanej przez segmentalne duplikacje z kilku możliwych scenariuszy opisanych w literaturze. Zastosowane algorytmy będą korzystały z metod teorii grafów oraz wnioskowania bayesowskiego.

Oczekujemy, że metody i narzędzia zaproponowane w projekcie, dzięki wykorzystaniu analitycznych i obliczeniowych procedur, będą stanowić kolejny krok ku lepszemu diagnozowaniu pacjentów z złożonymi rearanżacjami chorobowymi. Projekt pozwoli na metodyczne badanie wariacji genetycznych w rejonach pomijanych w czasie standardowych procedur wykrywających choroby genetyczne, lecz często je powodujących. W dłuższej perspektywie projekt z pewnością pozwoli na lepsze zrozumienie struktury, różnorodności i wpływu wariacji strukturalnych oraz wariacji genetycznych obejmujących segmentalne duplikacje na fenotyp, ewolucję oraz występowanie chorób genetycznych. Wydaje się, że największą wartością proponowanych badań będzie stworzenie nowych metod bioinformatycznych użytych do badania wariacji genetycznych oraz ich potencjał diagnostyczny.