Postęp, który notuje się w ostatnich latach w bioinformatyce jest ogromny. W szczególności coraz dostępniejsze (m.in. z uwagi na cenę) jest wykonywanie sekwencjonowania genomów. Technika ta powoli zaczyna być dostępna nawet do celów diagnostycznych, co skutkuje tym, że opowieści o medycynie spersonalizowanej zaczynają nabierać realnych kształtów.

Z sekwencjonowaniem wiążą się jednak pewne bardzo istotne problemy, z których jednym z kluczowych jest rozmiar pozyskiwanych danych. Przykładowo w typowym przypadku sekwencjonowania genomu jednego człowieka powstaje około 300 GB surowych danych. Jeśli weźmiemy pod uwagę także wyniki kolejnych etapów analizy, to z łatwością rozmiar ten wzrasta do ponad 1 TB. O ile w przypadku jednej osoby można się tym za bardzo nie przejmować, to należy mieć na uwadze, że w medycynie spersonalizowanej taka czynność będzie wykonywana rutynowo dla wielu pacjentów. To zaś rodzi olbrzymie wyzwania przed laboratoriami, które będą zmuszone do przechowywania gigantycznych ilości danych. W celu wyobrażenia sobie skali problemu warto wspomnieć o jednym z ostatnich oszacowań, według którego w 2025 roku dane z sekwencjonowania będą pozyskiwane w tempie około 1 zettabajta (10^{21}) na rok, z czego około 2–40 eksabajtów (10^{18}) będzie musiało być przechowywane długoterminowo. Są to wartości olbrzymie, a warto wspomnieć, że obecnie realny koszt rocznego przechowywania oraz 15-krotnego przesyłu danych rozmiaru 1 TB to w przybliżeniu 1 tysiąc dolarów amerykańskich.

Tradycyjnym środkiem pozwalającym na zmniejszenie skali tego problemu jest stosowanie kompresji danych. Dzięki redukcji ilości zajmowanej pamięci dane skompresowane mogą być szybciej (i taniej) przesyłane oraz przechowywane, co ma duże znaczenie praktyczne. Osiągane współczynniki kompresji zależą jednak bardzo mocno od specyfiki danych a także od tego czy dopuszczalna jest tzw. kompresja stratna (zezwalająca na kontrolowaną utratę części informacji).

W naszym projekcie skupimy się na danych będących wynikiem jednych z ostatnich etapów analiz, tzn. na gotowych sekwencjach genomowych osobników bądź też na opisie różnic (wariantów) pomiędzy osobnikami a tzw. genomem referencyjnym (będącym pewnym przybliżeniem genomów wszystkich osobników danego gatunku). Obecnie właśnie na podstawie takich danych podejmuje się próby oceny skłonności m.in. do chorób o podłożu genetycznym.

Główne cele naukowe projektu są następujące:

- opracowanie algorytmu kompresji kolekcji danych genomowych osobników tego samego gatunku,
- opracowanie skompresowanej struktury danych umożliwiającej bardzo oszczędne przechowywanie informacji o wariantach występujących w osobnikach w dużej kolekcji z zapewnieniem szybkich mechanizmów dostępu,
- opracowanie uniwersalnego kompresora plików VCF (Variant Call Format) służących do przechowywania wyników analiz genomowych, w szczególności dla dużych kolekcji osobników,
- opracowanie kompresora plików gVCF zawierających szczegółowe wyniki analizy dla pojedynczego osobnika.

Elementem wspólnym wszystkich tych celów jest opracowanie nowych algorytmów kompresji bądź skompresowanych struktur danych dla danych genomowych różnego rodzaju. W wyniku naszych prac powinniśmy opracować narzędzia znacznie redukujące koszty przechowywania, przesyłu i analizy danych genomowych. Dążymy również do tego, aby dzięki naszym pracom analizy genomowe były dostępne dla szerszego grona badaczy dzięki temu, że przynajmniej część z nich nie będzie wymagała dostępu do kosztownej infrastruktury informatycznej. Wiele z badań będzie można wykonywać za pomocą niewielkich stacji roboczych, bądź nawet laptopów.