W ostatnich latach notuje się gwałtowny postęp w bioinformatyce. W szczególności opracowywane są coraz tańsze i łatwiej dostępne technologie sekwencjonowania genomów. Skutkuje to m.in. lawinowo narastającą ilością danych, których rozmiary stanowią problem nie tylko od strony analizy. Wyzwaniem dla istniejących infrastruktur sprzętowych jest już nawet samo przechowywanie i transfer tych danych.

Żeby docenić skalę problemu wystarczy zauważyć, że od roku 2003 do 2015 realny koszt zsekwencjonowania pojedynczego genomu ludzkiego spadł z ok. 30 mln dolarów do ok. 4 tys. dolarów. Co więcej najnowsze urządzenia takie jak Illumina TenX, choć jeszcze bardzo drogie, pozwalają na obniżenie tego kosztu do ok. 1 tys. dolarów. Urządzenia sekwencjonujące najnowszej generacji (IonTorrent, Oxford Nanopore) potrafią "wyprodukować" odczyty charakteryzujące się, co prawda, dużą stopą błędu, ale także bardzo długie, co w niedalekiej przyszłości pozwoli na znacznie dokładniejszą analizę sekwencji genomowych niż to ma miejsce obecnie. Spersonalizowana medycyna wydaje się być już nie odległą wizją, ale kwestią kilku najbliższych lat.

Warto także wspomnieć, że największa placówka sekwencjonująca genomy, Beijing Genomics Institute (BGI) w Chinach, używa obecnie 230 urządzeń sekwencjonujących, z czego większość to nowoczesne maszyny Illumina HiSeq 2000/2500. Ich teoretyczna przepustowość w roku to ok. 1.5 Pbp danych surowych, co przekłada się docelowo na ponad 10 PB. Trudno dokładnie określić jakie generuje to koszty dla takich instytucji jak BGI, ale dobrym oszacowaniem mogą być ceny przechowywania i transferu danych oferowanych przez jednego z największych dostawców usług przechowywania danych i obliczeń w chmurze jakim jest Amazon EC2. I tak, koszt rocznego przechowywania 1 TB danych oraz 15-krotnego ich pobrania wynosi ponad 1 tys. USD, czyli tylko ok. 4 razy mniej niż koszt zsekwencjonowania genomu jednego człowieka. W związku z problemami z rozmiarami koniecznych pamięci masowych, sporo ośrodków decyduje się na usuwanie danych archiwalnych i przechowywanie tylko najnowszych, a przez archiwalne dane mogą być rozumiane nawet takie, które zostały uzyskane zaledwie dwa lata wcześniej.

Tradycyjnym środkiem ułatwiającym przetwarzenie ogromnych danych są techniki kompresji. Metody te pozwalają nie tylko na redukcję miejsca i tym samym przyspieszenie obiegu danych (np. wymiany ich między placówkami naukowymi), ale w wielu przypadkach również przyspieszają analizę danych bezpośrednio w reprezentacji skompresowanej czy też pozwalają na obsługę większych ilości danych na konkretnym komputerze, posiadającym ograniczoną pamięć wewnętrzną (RAM). Jednym z pokrewnych podejść jest także wykorzystanie pamięci masowej (m.in. dysków twardych) nie tylko do przechowywania (bądź przenoszenia) danych, ale także jako znacznie tańsze, choć również istotnie wolniejsze, rozszerzenie pamięci operacyjnej (RAM) komputera.

Główne naukowe cele projektu są następujące:

- opracowanie skompresowanej i szybkiej struktury danych reprezentującej informacje o genomach wielu osobników tego samego gatunku,
- opracowanie nowych algorytmów analizy wyników z sekwencjonowania w celu uzyskania genomów sekwencjonowanych osobników; algorytmy te mają wykorzystywać wiedzę o genomach wielu do tej pory zsekwencjonowanych osobników tego samego gatunku,
- opracowanie nowych algorytmów dopasowywania sekwencji genomowych i kompresji wyników tego dopasowywania,
- rozwinięcie opracowanych wcześniej algorytmów zliczania k-merów w danych genomowych i opracowanie na ich podstawie nowych algorytmów analiz genomowych wykorzystujących m.in. statystki występowania k-merów.

Wspólnym mianownikiem tych celów jest kompresja danych, dzięki której możliwe będzie zredukowanie zapotrzebowania na pamięć dyskową konieczną do przechowywania danych. Przede wszystkim jednak zwarta reprezentacja danych w pamięci operacyjnej (bądź roboczej pamięci dyskowej), pozwoli na uruchamianie opracowanych algorytmów, w przeciwieństwie do algorytmów istniejących, na stosunkowo niedrogich stacjach roboczych. W dalszej perspektywie pozwoli to znacząco zredukować koszty przetwarzania takich informacji w ośrodkach obliczeniowych, a co więcej pozwoli na prowadzenie zaawansowanych badań na ogromnych danych za pomocą łatwo dostępnych komputerów klasy PC w miejsce drogich rozwiązań serwerowych.