Streszczenie popularnonaukowe

Sekwencja każdego ludzkiego genomu koduje niesamowitą historię rozwoju pojedynczej komórki do niezwykle złożonego organizmu. Ludzkie ciało składa się z dziesiątek bilionów komórek, a każda z nich, mimo niewielkiego rozmiaru, zawiera ogrom podstruktur. Genom składa się z czterech podstawywch jednostek budulcowych - nukleotydów - znajdujących się w 23 parach chromosomów. Chromosomy składają się łącznie z kilkudziesięciu tysięcy genów, które zawierają instrukcje wyznaczające kierunek rozwoju organizmu. Wymienione liczby pokazują skalę informacji, jakie są przechowywane w każdej komórce. Nasuwa się naturalne pytanie - czego można się dowiedzieć, gdybyśmy zrozumieli język programujący rozwój organizmów? Odpowiedzią na to pytanie zajmuje się genomika, dziedzina nauki wywodząca się z genetyki. Współczesna genomika to nauka bardzo interdyscyplinarna starająca się zrozumieć m.in. funkcje i ewolucje struktur genomicznych. Jednym z głównych narzędzi używanych w genomice jest sekwencjonowanie, którego wynikiem jest skład badanego odcinka genomu. Na przykład, możemy otrzymać macierz ekspresji genów, która zlicza ile jakich genów się tam znajduje. Dane z sekwencjonowania potrafią być ogromne - analiza gen po genie jest praktycznie niemożliwa. To pozostawia ogromne pole do popisu dla matematyków i informatyków, by wytworzyli narzędzia i oprogramowanie, które będą potrafiły przetworzyć badane dane i wyciągnąć z nich kluczowe informacje, które potem jest łatwo zinterpretować biologom w istotnie krótszym czasie.

W trakcie ostatniej dekady nastąpił dynamiczny rozwój technologii do sekwencjonowania pojedynczych komórek (ang. *single-cell sequencing*). Wcześniej, wszelkie analizy były przeprowadzane na grupach komórek (ang. *bulk sequencing*), co było podejściem łatwiejszym technologicznie, tańszym, ale o wiele mniej dokładnym. Rozdzielczość pojedynczej komórki w sekwencjonowaniu pozwala na poznanie każdej komórki oddzielnie. To prawdziwy przełom w genomice, który został nagrodzony tytułem *Metody Roku 2013* przez *Nature Methods*. Dzięki następnym rozwojom technologii, pojedyncze komórki można sekwencjonować na różnych poziomach molekularnych, np. chromatyny, RNA, czy też białek. Jednym z najnowocześniejszych podejść jest sekwencjonowanie multimodalne - czyli sekwencjonowanie genomu na różnych poziomach molekularnych jednocześnie. Dzięki temu, mamy o wiele bogatsze informacje o każdej z komórek, gdyż żaden z poziomów molekularnych nie jest w stanie samodzielnie zakodować pełnej informacji o genomie (centralny dogmat biologii molekularnej).

Celem proponowanego projektu jest rozwój nowego algorytmu z rodziny *modelowania tematów* (ang. *topic modeling*), który będzie potrafił skompresować dane z multimodalnego sekwencjonowania pojedynczych komórek do postaci, która ma drastycznie mniejszy rozmiar i jest łatwo interpretowalna. Modelowanie tematów wywodzi się z analizy dużych zbiorów tekstowych i polega na znajdowaniu grup słów, które często występują w podobnych kontekstach. Współcześnie, jednym z najbardziej udanych podejść do modelowania tematów jest *generatywne modelowanie Bayesowskie*, które używa probabilistycznego języka, by opisać relacje pomiędzy słowami i dokumentami tekstowymi. Okazuje się, że modelowanie tematów (rozumiejąc dokumenty tekstowe jako komórki, a słowa jako np. geny czy białka) potrafi świetnie przetwarzać dane genomiczne. Jednak, według naszej najlepszej wiedzy, nikt dotąd nie próbował modelować tematów multimodalnych dla danych z pojedynczych komórek. Efektem tego są mniej dokładne przetworzone reprezentacje wyników sekwencjonowania. Użycie pełnej informacji multimodalnej do tworzenia tematów odblokuje możliwość o wiele dokładniejszych analiz, które przy użyciu obecnych metod są niemożliwe. Do realizacji poszczególnych zadań badawczych wykorzystamy nowoczesne techniki z zakresu Bayesowskiego rachunku prawdopodobieństwa. Zaproponujemy różne architektury modeli bazujące na istniejących, dobrze działających modelach, by dopasować je do danych genomicznych. Modele będziemy ulepszać iteratywnie - testując na danych z sekwencjonowania i usprawniając to, co uznamy za wymagające poprawy. Otrzymane modele zostaną zaprogramowane i udostępnione na publicznym repozytorium jako pakiet open-source, by każdy mógł ich użyć za darmo.