# Streszczenie

Współczesne metody sekwencjonowania RNA służą w przeważającej liczbie przypadków do oceny poziomu ekspresji genów i ilościowego porównywania ekspresji w różnych próbach biologicznych. Dane transkryptomiczne można jednak rozpatrywać również w kontekście jakościowym, znajdując różnice, podobieństwa i wyciągając na ich podstawie wnioski.

Pierwszym celem niniejszej pracy było opracowanie metody jakościowej analizy sekwencyjnych danych transkryptomicznych w doświadczeniach wieloczynnikowych. Drugim - dołączenie metody interpretacji efektów biologicznych do tej analizy statystycznej. Wreszcie jako trzeci cel obrane zostało sprawdzenie, czy porównanie jakościowe może również poprawić lub wprowadzić nowy kontekst do analizy ilościowej.

Aby uzyskać porównania zbiorów danych w kontekście wspólnie i rozłącznie występujących sekwencji nukleotydów, wykorzystana została odległość Mash. Jest ona podstawą prowadzonej następnie analizy zmienności jakościowej transkryptomów i identyfikacji efektów statystycznie istotnych ze względu na ich zróżnicowanie. Ponadto, zaproponowano użycie ontologii GOSlim GOA w celu uzyskania bardziej szczegółowej interpretacji biologicznej wyników. Przedstawiono przykład zastosowania metody na rzeczywistych danych eksperymentalnych oraz omówiono uzyskane tym sposobem wyniki.

Rzeczywiste dane stanowiły dane transkryptomiczne z eksperymentu trójczynnikowego (czynniki: genotyp, dzień, temperatura), przeprowadzonego na jęczmieniu (Mikołajczak i in., 2022), uzyskane przez sekwencjonowanie za pomocą platformy Illumina. Dane te zostały poddane kontroli jakości i składaniu w skafoldy. Dla każdej pary zbiorów została obliczona odległość Mash, w wyniku czego otrzymano macierz odległości, która została poddana analizie PCoA, a następnie jedno- i wielowymiarowej analizie wariancji. Obliczenia przeprowadzone zostały dla surowych danych, dla skafoldów oraz wspólnie dla zbioru surowych i złożonych danych, w celu porównania wyników. Wyniki analiz wykazały największy wpływ czynnika temperatura oraz czynnika dzień na różnice jakościowe w zbiorach transkryptomów. Etap składania zauważalnie różnicował zbiory między sobą, natomiast wyniki osobno dla skafoldów i dla surowych danych były podobne.

Zbiory danych złożonych zostały podzielone na podzbiory ze względu na terminy GOSlim przypisane do poszczególnych sekwencji. Dla każdego terminu, dla którego podzbiory danych były niepuste, przeprowadzona została analiza jak ta opisana powyżej. Podobnie jak dla całych zbiorów, dla podzbiorów na ogół wykazano istotności efektów czynników: temperatury i dnia.

Dane zostały poddane różnicowej analizie ilościowej za pomocą dwóch modeli: modelu liniowego analizy wariancji oraz z wykorzystaniem modelu analizy kowariancji, gdzie zmiennymi objaśniającymi były współrzędne główne. Liczby genów wykazujących zróżnicowany poziom ekspresji uzyskane przy wykorzystaniu informacji zawartych w odległościach Mash, tzn. w modelu analizy kowariancji, były niższe niż te uzyskane przy użyciu modelu liniowego.

Najważniejsze wnioski ogólne wynikające z przedstawionej pracy to: odległość Mash jest odpowiednim narzędziem do wprowadzenia kontekstu jakościowego do analizy danych transkryptomicznych, opracowana metoda jest stosunkowo mało wymagająca w kontekście czasu i zasobów, a zastosowanie ontologii GOSlim pozwoliło na wzbogacenie analizy danych transkryptomicznych o element funkcjonalny. Wykorzystanie wyników z PCoA wykonanej na macierzy odległości Mash umożliwiło poprawienie wyników różnicowej analizy ilościowej.

Słowa kluczowe: odległość, Mash, transkryptom, GO, GOSlim, doświadczenia wieloczynnikowe, interakcja genotypu i środowiska