

Streszczenie

Metabolomika, dziedzina zajmująca się badaniem metabolitów, analizuje dynamiczne zmiany w stężeniach molekuł o niskiej masie cząsteczkowej (ok. 1-1.5 kDa), aby dostarczyć wgląd w biologiczne warunki panujące pod wpływem różnych stresorów. Identyfikacja metabolitów jest zazwyczaj realizowana za pomocą technik analitycznych, takich jak magnetyczny rezonans jądrowy (NMR) i spektrometria mas (MS), często w połączeniu z technikami separacji. W ostatnich latach liczne badania skupiały się na korelacjach pomiędzy stężeniami metabolitów a stanami chorobowymi, zaburzeniami genetycznymi oraz progresją nowotworów. Pomimo swojego znaczenia w zrozumieniu złożonych procesów biochemicznych, wiele istniejącego oprogramowania do metabolomiki pozostaje niedostępne, niewspierane lub płatne, a większość narzędzi koncentruje się na pojedynczych etapach analizy, zamiast oferować kompleksowe, zautomatyzowane rozwiązania.

Niniejsza praca odpowiada na rosnącą potrzebę kompetentnych i powtarzalnych potoków przetwarzania danych (ang. *pipelines*) w metabolomice. Aby sprostać tym wyzwaniom, stworzono NASQQ, otwarty i ogólnodostępny potok przetwarzania danych, zaprojektowany do analizy jednowymiarowych protonowych widm NMR. NASQQ integruje istniejące rozwiązania z zaawansowanymi modelami uczenia maszynowego i analizą wzbogacenia ścieżek, aby zapewnić stabilne i powtarzalne rozwiązania dla badań metabolomicznych. Implementacja NASQQ obejmuje modułową metabolomiczną sekwencję zadań (ang. *workflow*) napisaną w języku Nextflow, obejmującą przetwarzanie widm i moduły analizy danych, w tym zarówno podejścia jednowymiarowe, jak i wielowymiarowe oraz analizę ścieżek biologicznych.

W ramach pracy potok przetwarzania danych został zwalidowany za pomocą otwartego zestawu danych obejmującego Dysautonomię rodzinną (ang. *Familial Dysautonomia*), obejmującego analizę surowych widm surowicy zarówno pacjentów z tą chorobą, jak i ich zdrowych krewnych. Ewaluacja wykazała skuteczność i przydatność NASQQ w badaniach nad chorobami. Co więcej, zastosowanie NASQQ do próbek Dysautonomii rodzinnej podkreśliło w testach wielowymiarowych oraz w analizie wzbogacenia ścieżek biologicznych istotne odkrycia w przetwarzanych widmach. Wykorzystując otwarte narzędzia bioinformatyczne, niestandardowe funkcje i uczenie maszynowe, NASQQ oferuje ogólnodostępny i skuteczny przepływ pracy, który standaryzuje metody przypisywania sygnałów, redukuje błędy operacyjne i zwiększa powtarzalność poprzez automatyzację i równoległe przetwarzanie w stabilnym środowisku opartym na kontenerach (ang. *containerized environment*). Potok przetwarzania danych skutecznie łączy surowe dane widmowe z interpretacją biologiczną, a także otwiera nowe perspektywy na przyszłe usprawnienia i rozszerzone zastosowania w badaniach metabolomicznych.

słowa kluczowe: Bioinformatyka, Metabolomika, 1D ^1H magnetyczny rezonans jądrowy, Uczenie maszynowe, Potok przetwarzania danych, Nextflow.