Abstract

Metabolomics, a comprehensive study about metabolites, investigates dynamic changes in the concentrations of low-molecular-weight molecules (approx. 1-1.5 kDa) to provide insights into biological conditions influenced by various stressors. The identification of metabolites is commonly achieved using analytical techniques such as nuclear magnetic resonance (NMR) and mass spectrometry (MS), often coupled with separation techniques. Recently, numerous studies have explored the correlation between metabolite concentrations and conditions like diseases, genetic disorders, and cancer progression. Despite its significance in understanding complex biochemical processes, much of the existing metabolomics software remains inaccessible, unsupported, or behind paywalls, with most tools focusing on isolated steps of the analysis process rather than offering comprehensive, automated solutions.

This thesis addresses the growing need for robust and reproducible pipelines in metabolomics. To tackle these challenges, it introduces NASQQ, an open-source Nextflow pipeline specifically designed for the analysis of 1D proton NMR spectra. NASQQ integrates existing solutions with advanced machine learning models and pathway enrichment analysis to provide a robust and reproducible solution for metabolomics research. The implementation of the NASQQ pipeline includes a modular metabolomic workflow written in Nextflow framework, covering spectral processing and data analysis modules, including both univariate and multivariate approaches and pathway analysis.

In this work, the pipeline was evaluated using an open dataset on *Familial Dysautonomia*, involving the analysis of raw serum spectra from both patients and healthy relatives. This evaluation demonstrated the pipeline's effectiveness and applicability in disease studies. The application of NASQQ on *Familial Dysautonomia* samples highlighted significant findings from the spectral processing, univariate tests, machine learning assessments, and pathway enrichment analysis. By leveraging open-source bioinformatics tools, custom functions and machine learning, NASQQ offers an accessible end-to-end workflow that standardizes signal assignment methodologies, reduces operational confusion, and enhances reproducibility through automation and parallelization in a stable containerized environment. The pipeline effectively links raw spectral data with biological interpretation, while also paving the way for future improvements and expanded applications in metabolomics research.

keywords: Bioinformatics, Metabolomics, 1D ^{1}H nuclear magnetic resonance, Machine learning, Pipeline, Nextflow.