Biostatistics and machine learning in MALDI MS research

Lukáš Pečinka[1,2], Lukáš Moráň[3,4], Monika Vlachová[5], Petra Kovačovicová[2,3], Sabina Ševčíková[5], Aleš Hampl[2,3], Petr Vaňhara[2,3], and Josef Havel[1,2]

[1]Faculty of Science, Masaryk University, Brno, Czech Republic; [2]International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic; [3]Faculty of Medicine, Masaryk University, Brno, Czech Republic; [4]Research Centre for Applied Molecular Oncology (RECAMO), Masaryk Memorial Cancer Institute, Brno, Czech Republic; [5]Babak Myeloma Group, Department of Pathophysiology, Faculty of Medicine, Masaryk University, Brno, Czech Republic
lukas.pecinka@med.muni.cz

With increasing demands on precise analyses of biological samples in complex biological matrices, there is also a need to develop and optimize mass spectrometric (MS) methods. MS analysis of whole cells, plasma samples, and other biological materials is of great importance for monitoring and elucidating biological processes in the organism and provides important information regarding organism pheno/genotype. In two topics presented herein, different techniques for whole cell samples and peripheral blood plasma will be presented. The whole cell MALDI TOF MS is already used in clinical microbiology and diagnostics. In recent years it has been introduced also to cell biology, immunology, and cancer biology.

The first project focuses on classifying ovarian cancer cells with different percentages of cell populations with a knockout of a single gene (TUSC3). Different cell types (4 in total) from different organisms (human and mouse) were introduced to MS analysis. MS method was combined with multivariate statistical and machine learning algorithms (PLS-DA, ANN, and RF for example) using an R programming language. Data obtained from MS were analysed via an in-house developed R-script. In total 5 optimized classifiers based on different algorithms were established and compared for 175 mass spectra divided into 5 groups. PLS-DA was determined as a model with the best performance with 100% accuracy (95% confidence interval, CI = 94.7-100%) for the test data. The method described above was further used for other studies; to follow the differentiation process of hESCs to ELEPs for example. We visualized the full differentiation trajectory based on spectral data only and revealed also some phenotypic abnormalities linked to passage number, and by proxy aneuploidy status of hESCs.

The second project is dealing with the development method for the analysis of human plasma samples using MALDI TOF MS. This project aims to discriminate multiple myeloma (MM) patients and patients with similar diseases like plasma cell leukemia (PCL) and extramedullary multiple myeloma (EMD). The two steps protein extraction protocol was developed for the classification of MM, PCL, and EMD patients. Intensity across the whole *m/z* range increased approx. 50 times when extraction protocol was used (compare to dilute direct plasma samples). The accuracy of classification models using ML algorithms (RF, PLS-DA, and ANN) was 80-90% for the training dataset and 80-85% for the test dataset. These findings may help accelerate the integration of MALDI MS into a clinical application as the diagnosis of MM, PCL, and EMD is rather inaccurate nowadays.

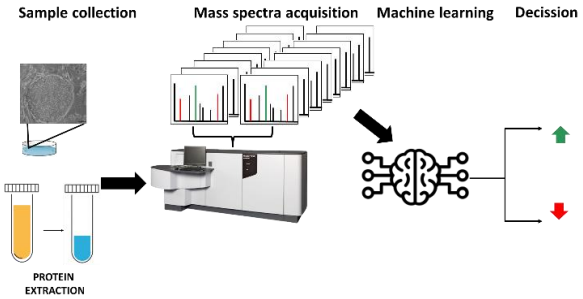

Figure 1: Generalized workflow scheme

Supported by a grant of the Czech Ministry of Health NU21-03-00076 and the project National Institute for Cancer Research (Programme EXCELES, ID Project No. LX22NPO5102)—Funded by the European Union—Next Generation EU. Also supported by MHCZ-DRO FNBr, 65269705, by Masaryk University MUNI/A/1370/2022, MUNI/A/1298/2022, and MUNI/11/ACC/3/2022