Hierarchical bi-clustering of mass spectrometry imaging data using recursive rank-2 non-negative matrix factorization

Melanie Nijs[1], Etienne Waelkens[2], Bart De Moor[1]
1 KU Leuven, ESAT-STADIUS Leuven, Belgium.
2 KU Leuven, Dept. Cellular and Molecular Medicine, Leuven, Belgium

Melanie.Nijs@esat.kuleuven.be

Introduction: Mass spectrometry Imaging (MSI) is an advanced analytical technique that measures the abundance and spatial mapping of molecules present within a sample. One of the main challenges in MSI data analysis remains the complexity of the m/z-spectra caused by noise, matrix effects and other artifacts (Alexandrov et al., 2020). We observe that molecules with a high abundance suppresses those with lower m/z-ratios and can misdirect classical tools for data analysis, such as principal component analysis (PCA). Therefore, the observed significance of a molecule may not always be directly related to its abundance. In this work, we present a recursive rank-2 non-negative matrix factorization (rr2-NMF) algorithm that automatically returns spectral and spatial visualization of colocalized molecules, both highly and lowly abundant.

Method: The proposed method performs a hierarchical inspection of the MSI of a tissue as it iteratively divides the data in two subclusters according to the similarity of the spectra, using the additive characteristic of NMF. We further examine two stopping criteria for the recursion: (i) a predefined depth of recursion, and (ii) convergence of the cosine similarity between the spectra of the pixels in each cluster. When the stopping criterion is met for a subcluster, the corresponding spectral and spatial distributions are plotted, together with a table containing the values of the peaks. To increase computational efficiency, we also provide a parallel implementation. We demonstrate performance and ease-of-use on MALDI-TOF MSI data from pancreatic tissues from healthy mice.

Results: The results are easy-to-interpret spectral and spatial visualizations and demonstrate an increased level of detail compared to standard NMF (Nijs et al., 2021). The rr2-NMF algorithm starts from the full dataset and runs automatically until the stopping criterion is met. In contrast to previous analyses (Smets et al., 2019, Derwae et al., 2023) of our samples, rr2-NMF reveals detailed spatial and spectral distributions without parameter tuning or manual interaction. For the mouse pancreatic tissue, we detect known proteins present within the high abundances, including insulin (± 5800 Da) and the protein present in the acinar cells (± 6650 Da). Moreover, the method detects several molecules of interest with lower abundances, such as: ± 5821 Da, ± 5840 Da, 3923 Da, ± 3124 Da and ± 2901 Da, each of which exhibiting a particular spatial correlation. The spectral and spatial signatures of the clusters are confirmed by visualization of the ion images of the peak values. Our tool, implemented using Python's NMF package, performs unsupervised bi-clustering of the 1.6 GB sample in a few minutes, and allows for a 15% speed increase when executed in parallel. Finally, we compare the individual sample analyses with a multi-sample analysis across three pancreatic slices and demonstrate the scalability of the algorithm.

Novelty: The proposed method already reveals detailed spatial and spectral distributions by tackling the dominance of the molecules in high abundance.

Preliminary data: Using rr2-NMF, we automatically detect relevant m/z-values in healthy mice pancreas and demonstrate scalability to multiple samples and bigger datasets.