Title: Development of advanced clustering methodologies in mass spectrometry imaging for a comprehensive exploration of cancerous tissues

Intra-tissue heterogeneity is an essential factor in studying diseased tissue microenvironment, evolution of disease progression and rare cell populations. Current methodologies include a number of tissue-based tests and targeted assays (immunohistochemistry (IHC), haematoxylin and eosin staining (H&E), fluorescence microscopy, etc.) which need to be performed sequentially to reflect the global cellular state, phenotypical changes and biochemical pathways. However, targeted assays are time consuming, require prior knowledge and limit un-targeted analysis as well as discovery-based research of biomarkers. Thus, there is a need for a new label-free tool which will distinguish these subpopulations by enabling a complex and rapid detection of discovery-based and spatially resolved molecular signatures in a single experiment. It is therefore quite natural that molecular spectroscopic imaging, particularly based on mass spectrometry, has developed in the field of cell and tissue biology. Mass spectrometry imaging (MSI), particularly Matrix-Assisted Laser Desorption Ionization (MALDI), has demonstrated its vast potential for biomarker discovery and has played a central role in clinical research, linking classical histology and molecular analysis. This high spatial resolution (reaching 5µm) technique allows us to directly map molecular species (proteins, peptides, lipids, and metabolites). MALDI imaging allows us today to systematically acquire tens of thousands of mass spectra and even more on a region of interest of a tissue section. In order to get closer to what is done in conventional histopathology, we want to highlight from such spectrometric data sets similar areas at the molecular level and this is why experts in the field use clustering methods (such as the well-known *K-means* algorithm) from machine learning. Even if such a multivariate approach goes in the right direction by proposing a non-targeted and therefore unbiased approach, one quickly realizes that its current use remains problematic in several respects. First, the number of clusters to be considered for a given dataset is often selected by a trial-and-error procedure based on the observation of the generated clustering maps. The optimal number of clusters is thus selected on the basis of the biological structures that are expected to be found in this clustering map, which is not satisfactory in the context of an exploration without *a priori*. We must not forget that humans always give credence to what they think they see and that mathematical methods can generate very surprising images. A second, more insidious problem lies also in the fact that the clustering algorithms used today in spectroscopic imaging take very poorly into account so-called unbalanced clusters, i.e. clusters containing very different numbers of spectra. In the context of the characterization of cancerous tissue, we are often in a position to easily detect healthy and cancer areas in a tissue section, but the subpopulations of cancerous cells are very often not highlighted because they correspond to a much more limited number of spectra expressed by a small variance. A last aspect often underestimated in the use of clustering in spectral imaging is the choice of the default Euclidean distance. The distance between two spectra is indeed a way to measure their potential similarity but if its choice is not optimal the final clustering map is not representative of the biological reality of the sample. The objective of this thesis is to solve these different problems and to propose a global methodology to rationalize the use of clustering in mass spectrometry imaging (MALDI, SpiderMass MSI,…) for an exhaustive characterization of heterogeneities in cancer tissues. This thesis will be carried out in the framework of a collaboration between two laboratories of the University of Lille recognized at the international level for their respective competences, namely the LASIRE laboratory (UMR CNRS 8516) specializing in spectroscopy and machine learning (i.e. chemometrics) and the PRISM laboratory (INSERM U 1192) specializing in mass spectrometry imaging applied to biological tissue characterization.

PhD advisor: Pr. L. Duponchel, LASIRE lab (ludovic.duponchel@univ-lille.fr)

Co-supervisor: Dr. Nina Ogrinc (nina.ogrinc@univ-lille.fr)

Application deadline: May 6th, 2022

How to apply: Please send a CV, a letter of motivation and your transcripts (first and second year of Master) by email to Ludovic.duponchel@univ-lille.fr