IMAGING MASS SPECTROMETRY IN DETECTING TUMOR HETEROGENEITY

El. Ibrahim, Y. Mostafa, T. Renad, A. Mariem and Abd El. Donia

I. INTRODUCTION

Mass Spectrometry Imaging (*MSI*) is a technology that simultaneously provides the **spatial distribution** of hundreds of biomolecules directly from tissue lead to **minimal loss** of **histological information**. The most common technique used for it is "Matrix-Assisted Laser Desorption/Ionization (*MALDI*)". Accordingly, the same tissue section can be histologically assessed and registered to the MSI dataset and be accessed using any programming language (e.g. *MATLAB*, *Python*, etc.). This high cellular specificity is behind the increasing popularity of MSI in cancer research and its proven ability to identify diagnostic and prognostic **biomarkers**.

Significance: MSI provides **untargeted spatial-molecular information** necessary to uncover molecular Intratumor heterogeneity. The challenge has been to identify those **tumor subpopulations** that drive patient outcomes within the highly complex datasets (*hyper-dimensional data, Intratumor heterogeneity, and patient variation*). Here we report an automatic, unbiased pipeline to nonlinearly map the hyper-dimensional data into a **3D space**, and identify molecularly distinct, clinically relevant tumor subpopulations. We demonstrate this pipeline's ability to uncover subpopulations statistically associated with patient survival in primary tumors of **gastric cancer** and with metastasis in primary tumors of **breast cancer**.

Keywords: MSI, MALDI, Intratumor heterogeneity and hyper-dimensional data.

II. MATERIALS AND METHODS

1. MSI

The core of MSI is to detect **mass spectrum** of atoms, many techniques could be used for that as mentioned above but now let's focus on the main technique.

MSI machine receives the **sample** taken from patient, **heats** it till vapor state and the resultant charge of atoms could be then **accelerated** using ions accelerator, then it is time for **magnet** to make them focused and control the movement of atoms with high and low mass-to-charge (*m/z*) ratio by concept of **deflection** which states that atoms with *higher* m/z ratio will deflect *less* and **vise-versa** and as a final step we have the **detector** with receives this values (*peaks*) and graph them.

2. MALDI

It is an **ionization technique** that uses a **laser** energy absorbing **matrix** to create ions from large molecules with **minimal fragmentation**. Here comes the benefit of MSI technique which allows us to work with **tissue section**, the process of MALDI could be defined as in (*Figure1*).

We can then extract different images according to single **m/z** values integrated over all pixels.

3. Dimension Reduction

In machine learning we are having too many factors on which the final classification is done. These factors are basically, known as variables. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play.

It is the **transformation** of data from a **high dimensional space** into a **low-dimensional space** so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension. Working in high-dimensional spaces can be undesirable for many reasons; raw data are often sparse as a consequence of the **curse of dimensionality**, and analyzing the data is usually computationally intractable. Dimensionality reduction is common in fields that deal with large numbers of observations and/or large numbers of variables, such as *signal processing, speech recognition, neuroinformatics, and bioinformatics*.

Dimension Reduction dives into two essential components which are: **Feature Selection** and **Feature Extraction**

Feature Selection Methods

- Filter
- Wrapper
- Embedded

Dimension Reduction methods can be divided into two categories. The first one is **linear** technique and the second one is **non-linear** technique. Each

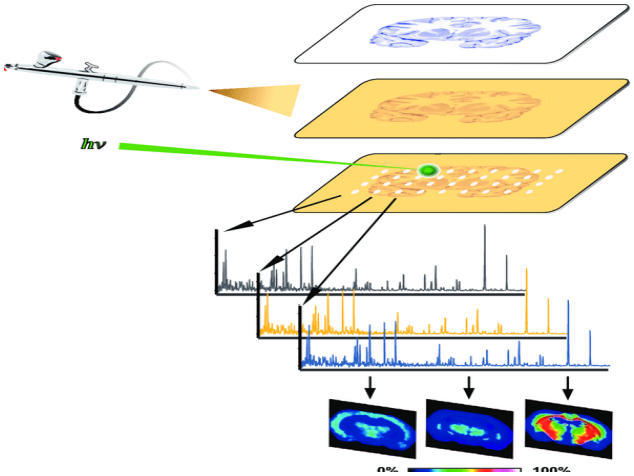

Figure1: MALDI technique where sample is taken from patient and appeared in tissue section, then applying matrix on it with some analyte and finally a laser beam is focused on them to obtain mass spectrum of each pixel

one has its own advantages and disadvantages, but we will go deeper into one example of each one of them.

Principle Component Analysis (PCA)

It is the process of computing **the principal components** and using them to perform a change of basis on the data and also it is a **linear** dimension reduction technique.

PCA is used in exploratory data analysis and for making **predictive models**. It is commonly used for **dimensionality reduction** by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. The first principal component can equivalently be defined as a direction that **maximizes** the variance of the projected data (*Figure2*).

t-distributed Neighborhood Stochastic Embedding (t-SNE)

It is a **machine learning algorithm** for **visualization** based on **Stochastic Neighbor Embedding** and also it is a **non-linear** dimension reduction technique. The t-SNE algorithm comprises **two** main stages. First, t-SNE constructs a **probability distribution** over pairs of high-dimensional objects in such a way that *similar objects are assigned a higher probability while dissimilar points are assigned a lower probability*. Second, t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it **minimizes** the Kullback–Leibler divergence (*KL divergence*) between the two distributions with respect to the locations of the points in the map. While the original algorithm uses the **Euclidean distance** between objects as the

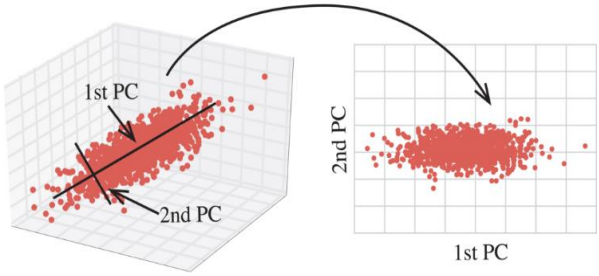

Figure2: PCA dimension reduction technique which shows the variance between data from high dimension space to its principal components

base of its similarity metric, this can be changed as appropriate.

t-SNE has been used for visualization in a wide range of applications, including *computer security research, music analysis, cancer research, bioinformatics and biomedical signal processing*. It is often used to visualize high-level representations learned by an *artificial neural network*.

Figure3: t-SNE dimension reduction technique applied on MNIST dataset

III. RESULTS