Machine Learning and Mass Spectrometry Imaging: Application in Tumor Heterogeneity

Hendwan Abozide, Youssef Saleh, Rokaia Mohammed

Cairo University, Faculty of Engineering, Department of Healthcare Engineering and Management
Supervised by of Dr. Walid Abdelmoula, Research Scientist at Brigham and Women's Hospital / Harvard Medical School

Abstract

Identification of tumor subpopulations that detriment the disease prognosis and the patient outcomes is essential for personalized therapy. Studies showed that Mass spectrometry imaging (MSI) could visualize intra-tumor heterogeneity and helped identify diagnostic and prognostic biomarkers. We studied the methods introduced by Abdelmoula et al. and used it as reference for our work. Our goal was to study the intra-tumor heterogeneity of gastric and breast cancer data. We implemented our project using Python instead of using MATLAB that was used by our main reference. We used machine learning techniques and statistical methods to identify the significant tumor subpopulations that affect the patient status. Dimensionality reduction was performed using t-distributed stochastic neighbor embedding (t-SNE) and K-means clustering was applied to discretize the data. Statistical analysis was done on the discretized data to deduce the significant clusters, hazardous cluster in gastric cancer data and metastatic cluster in breast cancer data. Significance Analysis of Microarrays (SAM) was used to reveal the statistically significant protein ions with FDR < 0.001 associated with these significant clusters. We built a predictive model using the Support Vector Machine (SVM) to predict the survival status of gastric cancer patients and the metastatic status of breast cancer patients.

Introduction

Mass spectrometry-based Imaging (MSI) is a technique that analyzes the complex-protein molecules in the tissue sample. The most common technique used in the clinical applications of MSI is **Matrix-Assisted Laser Desorption/Ionization (MALDI)** that can reveal the early onset of cancer biomarkers.

Tumor heterogeneity is defined as the presence of different tumor subpopulations that contribute to distinct phenotype and influences cancer evolution. It helps in tracking tumor progression and ultimately determining the patient status outcome. MALDI-MSI technique is used to visualize the different tumor subpopulations existing in the same sample, as shown in Fig. 1, which is identified as **intra-tumor heterogeneity.**

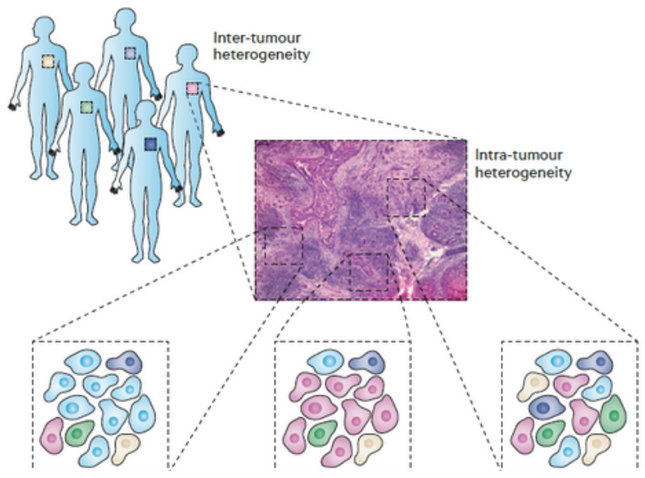

Fig. 1. Showing the inter and intra-tumor heterogenity

Methadology & Results

t-SNE and K-means

Intra-tumor heterogeneity was clearly visualized using a non-linear dimensional reduction technique: **t-distributed stochastic neighbor embedding (t-SNE).** It can accurately represent high-dimensional data in a low-dimensional space (2D and 3D). T-SNE showed the molecular differences between tumor subpopulations that contribute to cancer evolution. Fig. 2 shows the t-SNE maps of the two MALDI – MSI datasets we used (Gastric & Breast).

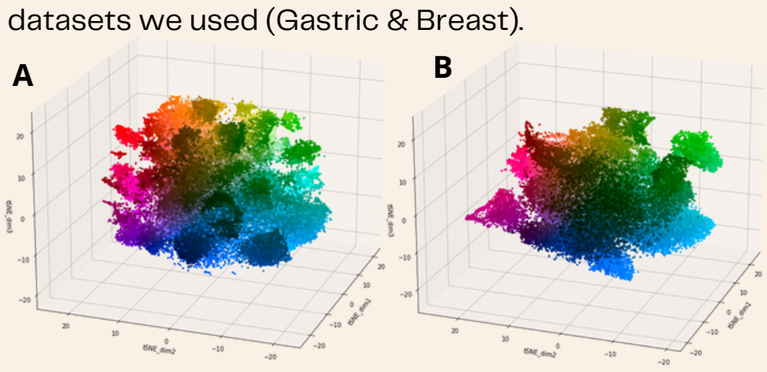

Fig. 2. (A) t-SNE scatter map of the gastric cancer represented in lower dimensional space. (B) t-SNE scatter map of the breast cancer represented in lower dimensional space.

Better visualization of tumor subpopulations was obtained by discretization of the t-SNE maps using K-means clustering method. The optimum K clusters were **K=4** in gastric cancer and **K=8** in breast cancer.

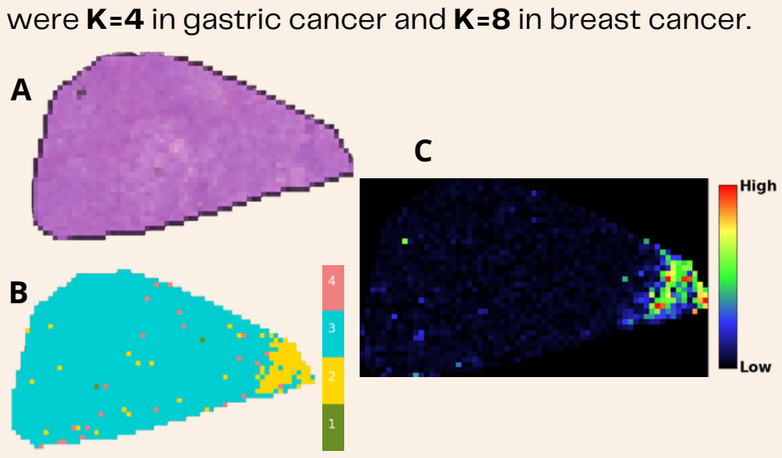

Fig. 3. (A) Showing histology image of a patient sample (B) K-clustered image of patient sample at K=4 (C) MSI of protein ion at mz = 3374

Fig. 3B & Fig. 3C demonstrate that K-means clustering and the MSI data show more spatial distribution that the histology image and that although the subpopulations are molecularly and phenotypically distinct, they are histologically identical.

Survival Analysis of Gastric Cancer

Kaplan-Meier is used to calculate the survival probability of the patient to a certain point in time. Cox regression is then used to calculate the Cox Hazard ratio for each cluster. Fig. 4 shows Kaplan – Meier curves and the Cox hazard ratios highlighting that clusters 2 & 3 are significantly different.

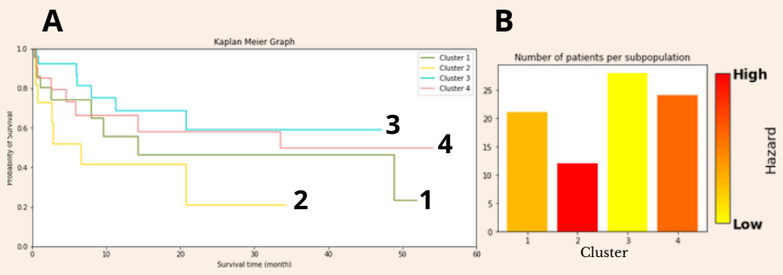

Fig. 4. (A) Kaplan-Meier curves showing the survival distribution for each cluster (B) Bar plot colored using Cox hazard ratio showing the number of patients in each cluster.

Metastasis Analysis of Breast Cancer

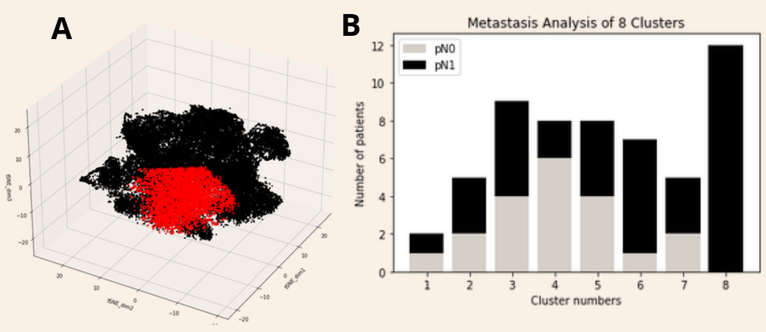

Fig. 5. (A) The fully metastatic cluster, highlighted in red, is shown in the t-SNE map. (B) Distribution of the metastatic (black) and nonmetastatic (grey) patients.

A fully metastatic cluster was observed in Fig. 5B when illustrating the distribution of metastatic and non-metastatic patients for each cluster using a histogram. This fully metastatic cluster (cluster 8) is highlighted in the t-SNE clustered image in Fig. 5A and can be compared against other clusters

SVM Classifier Results

By performing a validation method called **Leave One Patient Out (LOPO)**, we drop one patient and train the classifier and test it on the left out patient, and this process is repeated on all patients. Fig. 6 shows the patient statuses prediction results.

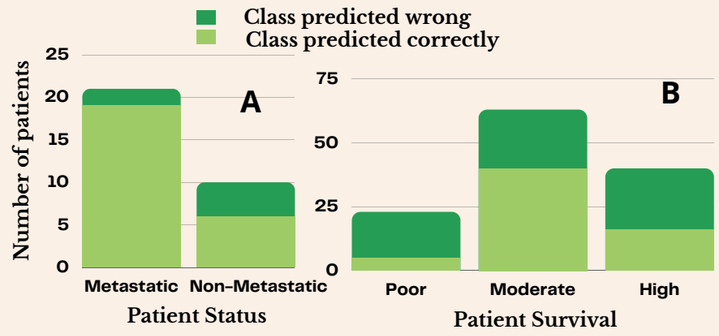

Fig. 6. (A) Breast cancer Metastatic and Non-Metastatic predictions (B) Gastric cancer Poor, Moderate and High predictions.

The most stable protein ions identified (most recurring) after completing LOPO runs are shown in Fig. 7 for gastric and breast cancer.

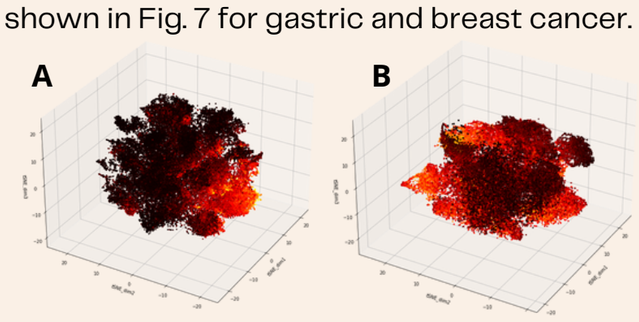

Fig. 7. (A) Gastric cancer t-SNE map highlighting m/z protein ion intensity: **3374** (B) Breast cancer t-SNE map highlighting m/z protein ion intensity: **4999**

References

Main reference: W. M. Abdelmoula et al., "Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of Mass spectrometry imaging data," Proc Natl Acad Sci U S A, vol. 113, no. 43, pp. 12244–12249, May 2016, doi: 10.1073/PNAS.1510227113.

[Fig.1]: A. Marusyk, V. Almendro, and K. Polyak, "Intra-tu-mour heterogeneity: a looking glass for cancer?," Na-ture Reviews Cancer 2012 12:5, vol. 12, no. 5, pp. 323–334, Apr. 2012, doi: 10.1038/nrc3261.

Conclusion

The applied machine learning techniques in this paper helped in analyzing the intratumor heterogeneity, which is a crucial factor in identifying the tumor evolution and subsequently the patient status. Tumor subpopulations affect the cancer evolution, and we were interested in identifying unique tumor subpopulations that influence the patient's outcome. Statistical analysis of the data helped in determining the tumor subpopulations with the most detrimental impact on the patient's status. Using the identified significant subpopulations, we built a classifier capable of predicting the metastasis status in breast cancer patients and survivability of gastric cancer patients. This classifier proves that we could identify early onset of cancer biomarkers, unlike histological images, and allows focused treatment for cancerous patients.