

Statistical Classification and Visualization of MALDI-Imaging Data

Marc Gerhard, Sören-Oliver Deininger
Bruker Daltonik GmbH, 28359 Bremen, Germany
{sod,mag}@bdal.de

Frank-Michael Schleif
University Leipzig, Dept. of Medicine, 04107 Leipzig, Germany
schleif@informatik.uni-leipzig.de

Abstract

Proteomic profiling based on mass spectrometry (ms) is an important tool for studies at the protein and peptide level. Thereby, the identification of relevant masses for a specific kind of disease, discriminating classification models as well as a reliable visualization is complicated. For the analysis of tissue sections the new technique of MALDI-Imaging has been introduced, which results in new data analysis challenges. Here we present a method for the analysis and visualization of MALDI Imaging spectra applied on a clinical cancer data set.

keywords: *data analysis, maldi imaging, class imaging, clinical proteomics*

1 Introduction

Mass spectrometry is a quantitative tool for the analysis of larger sample cohorts, such as cancer patients and requires advanced statistical analysis. Here we focus on the analysis of MALDI imaging measurements. Imaging mass spectrometry can be used to measure the spatial arrangement and relative concentration of compounds in biological samples. One recently developed imaging technology utilizes matrix-assisted laser desorption ionization (MALDI) and time-of-flight mass spectrometry to generate profiles and two-dimensional ion density maps of peptide/protein signals from thin tissue sections [3]. This allows to determine specific information of the local proteomic composition, relative abundance and spatial distribution of these components. Such imaging experiments yield a wealth of information, allowing the measurement and comparison of the molecular components of the section and to gain a better understanding of biological processes. For the analysis of MALDI imaging data we applied different statistical tools to access the information and integrated the method in a unique tool package. Unsupervised principal component analysis (PCA) [1] is discussed to access tissue specific variance in the data. A supervised classification approach based on support vector machines (SVM) [4] is used to classify specific tissue types¹.

2 Methods

A clinical breast cancer study with multiple tissue section of labeled breast cancer, connective and normal tissue has been analyzed. The sections were given by a standard pathologic staining (HE) see Figure 1(a) and without staining used for MALDI analysis. One observes tumor regions (left) and connective tissue lower right region. The experiment focused on the generation of classification models upon the measured ms spectra. From the given slice ms spectra were generate by application of matrix. This is complicated

¹Hardware and software by Bruker Daltonik GmbH, Bremen, Germany

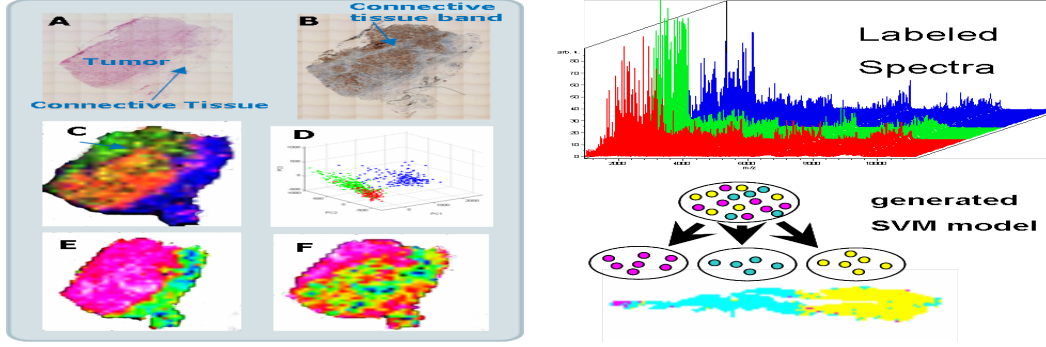


Figure 1. Left: Breast cancer sections: a) HE stain, b) staining by Her2 immunostaining, c) MALDI image colored in accordance to selected masses, d) 3d PCA plot of the data, e) MALDI image colored in acc. to 1st pc and f) in acc. to 2nd pc. Right: On top different labeled spectra without obvious class differences are shown. A SVM model is trained on spectra peaks. The model is used to generate a visualization of the tissue slices with respect to the classification decision. Thereby the cancer regions are the small piece on the left and the larger region on the right side. (color version available on request)

for tissue sections, but could be optimized using the ImagePrepTM station. All spectra were recorded on a linear autoflex III + smartbeam laser and automatically processed by the data analysis package ClinProTools 2.1. The MALDI-Imaging analysis is integrated into the tool package flexImaging, whereby each spectrum has been baseline corrected, the spectra set has been recalibrated and a peak picking procedure was applied to obtain a feature matrix (1012×94) for further statistical analysis (details [2]). Subsequently a PCA + SVM approach has been used. Thereby a standard PCA has been applied (see [1]) whereas for the SVM model generation an improved variant of a linear SVM classifier has been implemented.

SVMs are powerful learning machines with the idea to map input vectors \mathbf{x} into a high-dimensional feature space Z through some potential nonlinear mapping. In this space, an optimal separating hyperplane is constructed [4]. The determination of the hyperplane can be defined as a quadratic optimization problem [4]. The construction of a linear classifier for separating two classes of data in a D_V dimensional space may be equivalently viewed as to find a linear projection $\mathbb{R}^{D_V} \rightarrow \mathbb{R}$, represented by a normalized $D_V \times 1$ unit projection vector \mathbf{v} , such that a predefined separability measure between the two classes of data is maximized. Considering the data V as N_S column vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_S}$ and their corresponding class membership as $c_1, c_2, \dots, c_{N_S} \in \{-1, 1\}$, the expression for finding an optimal hyperplane for linearly separable data with non-negative slack variables ξ_1, \dots, ξ_{N_S} representing errors in the constraints is:

$$\text{Minimize} : \frac{1}{2} \mathbf{v}' \cdot \mathbf{v} + \sum_{i=1}^m p_i \xi_i$$

subject to $c_i(\mathbf{v}' \cdot \mathbf{x}_i + b) \geq 1 - \xi_i$ where coefficients p_i are positive constraints representing the relative importance of individual data points (here $p_i = 1 \forall i$) and b is a bias. The dual problem with Lagrangian multipliers α_i is given as:

$$Q(\alpha) \stackrel{Max!}{=} -\frac{1}{2} \sum_{i,j=1}^{N_S} \alpha_i \alpha_j c_i c_j \mathbf{x}_i \mathbf{x}_j + \sum_{i=1}^{N_S} \alpha_i$$

with constraints $\sum_{i=1}^{N_S} \alpha_i c_i = 0$ and $0 \leq \alpha_i \leq p_i$. The above equation can be easily extended to the nonlinear case substituting the inner product $\mathbf{x}_i \mathbf{x}_j$ by a non-linear kernel

function $K(\mathbf{x}_i\mathbf{x}_j)$ [4]. For the linear case the kernel may be just the ordinary inner product $K(\mathbf{x}_i\mathbf{x}_j) = \mathbf{x}_i^\top \cdot \mathbf{x}_j$. The unique solution of this optimization problem can be taken as a classifier. For linear SVM v can be interpreted as a feature ranking, which was used to generate a k-NN classifier [1] upon the best c features. The whole modeling processes including feature selection has been cross validated. The obtained classification model has been applied on a separate test data set and visualized in a $2d$ color image of the original MALDI image overlaid with the obtained classification coloring.

3 ClassImaging

The data of a multiple spectra from a single patient tissue slice has been first analyzed within a PCA analysis as shown in Figure 1(a). Here a clear separation of connective tissue from the cancer regions can be observed. The first principal component (pc) of the feature matrix explains 80% of variance and the 2nd pc 10% of the variance. On a single tissue section, the PCA correlates well with the histology and a clear separation of the different tissue parts can be observed (c.f. 1(a)). One major limitation for the use of PCA arises when it comes to sample cohorts from different patients. The main variance analyzed by PCA is rather between patients than interesting tissue subtypes. Further common large variations of the remaining amount of blood in the sample material are measured by the MALDI technology and lead to PCA models mainly focusing on the blood intensity variance in the sample. This however is in general not correlated with the tissue material leading to unimportant PCA models out of focus of the study. Hence a supervised analysis instead of PCA is needed which has been done by a SVM as explained above, thereby labeled data have been used in a model generation within a 10-fold cross validation procedure. A recognition accuracy of nearly 95% and a crossvalidation accuracy of 89% was obtained. The relevant features taken from the SVM ranking highlighted mass positions found to be interesting as potential biomarker candidates and are currently under evaluation. Subsequently the spectra are classified according to the model, the classification results have been used to generate the ClassImaging image (c.f. Fig. 1(b)).

4 Conclusions

Multivariate statistical methods as PCA and SVM are useful to analyze MALDI imaging data. PCA provides an automatic feature extraction leading to a compact representation of high dimensional data and supervised classification can be used to obtain reliable classification models. Both approaches are able to determine feature rankings. Thereby we observed that a simple PCA model is in general inappropriate and a supervised analysis is needed. In the extended version of an SVM classifier a good model and feature ranking could be observed indicating potential biomarkers which are currently under evaluation by clinical experts. ClassImaging allows a direct comparison of classification results with the histology improving the validation significantly².

References

- [1] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.
- [2] R. Ketterlinus, S-Y. Hsieh, S-H. Teng, H. Lee, and W. Pusch. Fishing for biomarkers: analyzing mass spectrometry data with the new clinprotocols software. *Bio techniques*, 38(6):37–40, 2005.
- [3] M. Stoeckli, P. Chaurand, D. Hallahan, and R. Caprioli. Imaging mass spectrometry: A new technology for the analysis of protein expression in mammalian tissues. *Nature Medicine*, 7(4):493–496, 2001.
- [4] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer, NY, USA, 1995.

²ACKNOWLEDGMENT: We are grateful to the MALDI-Imaging team (Bruker Daltonik GmbH Bremen, Germany) and A. Walch (GSF-Inst. for Pathology, Neuherberg, Germany)