Editorial

For reprint orders, please contact: reprints@future-science.com

Bioanalysis

Machine learning techniques for mass spectrometry imaging data analysis and applications

Ye Zhang[1] & Xin Liu*,[2]

[1]Department of Biological Sciences, University of Notre Dame, 142 Galvin Life Sciences Center, Notre Dame, IN 46556, USA

[2]Department of Chemistry & Biochemistry & the Harper Cancer Research Institute, University of Notre Dame, 140 McCourtney Hall, Notre Dame, IN 46556, USA

* Author for correspondence: Tel.: +1 979 888 8868; xliu13@nd.edu

First draft submitted: 19 December 2017; Accepted for publication: 25 January 2018; Published online: 21 March 2018

Keywords: clinical application • machine learning • MSI

Mass spectrometry imaging

As a powerful label-free analysis tool, mass spectrometry imaging (MSI) enables simultaneous detection and visualization of molecular species, such as lipids, proteins, peptides, glycans, metabolites and therapeutics, in a variety of biological samples [1]. Among the MSI techniques currently available, two important ones are MALDI imaging and desorption electrospray ionization imaging. In an MSI experiment, followed by simple but careful sample preparation, the mass spectrometer ionizes the analytes at each x, y coordinate on the section surface, resulting in an ordered array of mass spectra. Computational software and statistical analysis can then be used to reconstruct the ion density maps of various m/z values. After the candidate m/z values are determined, accurate mass matching to databases of known molecules or tandem MS (MS/MS) fragmentation can then be performed to determine the identities further.

In recent years, with advances in instrumentation and techniques, MSI has become more widely applied in different fields including pharmacology and clinical practice [2,3]. However, since the sensitivity, accuracy, acquisition speed, spatial resolution, mass resolution and throughput of MSI analysis has improved tremendously, the amount, dimensionality, as well as the complexity of datasets generated by MSI has also significantly increased. To interpret this huge amount of data, and extract essential chemical and spatial information more comprehensively and efficiently, there is an increasing interest in applying informatic approaches based on specialized machine learning (ML) algorithms to match the current demand [4,5]. The aim of this article is to provide the reader a brief overview onto how ML approaches could help process and elaborate complex imaging data, to provide a valuable molecular insight into different biological specimens, and make the MSI techniques more versatile and translatable in solving clinical problems.

ML approaches in MSI

During MSI statistical analysis, ML algorithms are implemented to detect patterns and structures within the data. Specifically, supervised and unsupervised ML algorithms are widely used for data classification and data clustering, respectively [5].

Supervised ML & applications

Supervised ML algorithms are effective in predicting patterns and features of a dataset with labels, which generally involves three steps: labeled training data selection, model optimization/validation and prediction of a new unlabeled dataset. It is frequently used to address classification problems to discriminate between groups of samples under different conditions. Clinically, the two most commonly used supervised algorithms are support vector machines (SVM) and random forest (RF) [5]. SVM can go beyond the computational limitation of linear classification by introducing kernel function. While RF is efficient in analyzing large datasets and robust in handling overfitting issues.

newlands press

SVM-based classification and RF algorithms have been successfully applied to derive information from MSI results on a variety of biological specimens, such as cancer patient tissue samples. It was reported that these approaches were utilized to accurately and reliably discriminate different cancer types, including thyroid cancer, breast cancer, colon cancer and liver cancer [6,7], by analyzing MALDI imaging data acquired from bioptic samples, thus assisting in determining the origin of the tumorigenesis irrespective of the metastatic sites. ML-assisted MSI has also been applied to assess patient tissue samples in identifying survival and recurrence-associated protein signatures in melanoma metastases [8] and sarcomas [9]. Another example is that HER2 expression level significantly correlates with breast cancer and gastric cancer patient outcomes. With implementing supervised ML algorithms on MALDI data from patient tissue samples, HER2 level could be deduced with high accuracy, sensitivity and specificity [10,11].

Supervised ML algorithms could also be used to depict tumor margins and microenvironment in the clinic. An analysis using ML methods to identify the margins of clear cell renal cell carcinoma suggested that adapting such approaches can better define tumor margins, resulting in more thorough tumor extirpation and reducing local recurrence [12]. In addition, a study using desorption electrospray ionization lipid imaging-derived classifier rapidly and precisely classified gliomas into different subtypes and grades [13], which provides critical surgery-related information to surgeons and pathologists, and is potentially applicable to real-time analyses.

In short, supervised ML classification is an invaluable tool to elucidate subtle molecular characteristics of different biological samples, thus enhancing the quality and accuracy of information obtained from a MSI analysis.

Unsupervised ML & applications

Compared with supervised ML, unsupervised ML methods do not require sample labeling, or previous knowledge, to highlight categories of the datasets. Thus, unsupervised learning could be used for exploratory data analysis either due to the lack of comprehensive information of the samples or the expectation to find hidden patterns that have not previously been discovered. In MSI, as a classical unsupervised task, clustering/segmentation is used to group spatially resolved spectra based on similar characteristics they share, and new samples can further be assigned to the identified clusters according to their spectral similarity. In clinical studies, this approach can also be applied as partially supervised, since pathologists could impose the correct number of clusters (dendrogram) based on visual inspection.

The common unsupervised algorithms include *k*-means, hierarchical clustering, partitioning around medoids (PAM) and density-based spatial clustering of applications with noise [14]. *k*-means is a well-known method, but it is sensitive to anomalous data points and outliers, while PAM is more robust to process data containing outliers. Hierarchical clustering is another widely used approach, which has an added advantage of interactive analysis of the clustering dendrogram over *k*-means but needs a large amount of memory to keep the full distance matrix. Density-based spatial clustering of applications with noise has also become more popular due to its superior capability to handle noisy data and pick up outliers.

MSI combined with clustering analysis has achieved many successes in recognizing the heterogeneity of morphologically similar, and neighboring cells. A good example is to utilize this approach clustering subpopulations of tumor cells or regions in tumor sections. It has been used to depict gastric cancer tissue subareas from patient sample sections more in detail than histological staining, helping to define and evaluate the sample more efficiently and accurately [15]. Unsupervised ML was also applied on MSI data to reveal subpopulations in sarcomas [16] and 3D colorectal adenocarcinoma biopsies [17] to find heterogeneous types of cells, which are difficult to be identified with staining methods. The trained model can quickly and precisely process 3D MSI datasets, thus eliminating time-consuming procedures needed for individual images. In a more recent study, MALDI imaging was implemented with PAM clustering analysis on colorectal cancer large-scale tissue microarrays, revealing a handful of clinically relevant implications, such as status, grade and prognosis simultaneously [18]. In summary, these studies provided insights of tumor heterogeneity and prospective biomarkers by looking at samples at the molecular level based on the robust unsupervised ML analysis of MS imaging datasets.

Other considerations

In most of the instances, analytical and technical variability or artifacts exists in MSI studies. For raw data collected, preprocessing steps including smoothing, baseline correction, normalization, peak alignment and picking are required to improve the data quality, and make spectra acquired comparable within the same experiment and in distinct analyses. Additionally, data dimensionality still presents a big limitation. Due to the intrinsic high-dimensionality of MSI data containing both molecular and spatial information of each pixel, dimensionality

reduction is also needed for eliminating noise and irrelevant features, while preserving important information and making the datasets more suitable for a robust and efficient ML analysis. Several techniques based on either linear or nonlinear mapping of the data to the low-dimensional space are available, such as principal component analysis, probabilistic latent semantic analysi, self-organized maps and *t*-SNE [17]. However, more better-performing methods are still highly needed, and these various data reduction methods should always be tested in new circumstances to identify the most reliable one.

Conclusion & future perspective

With the technological developments in specificity, sensitivity and resolution, MSI has become a promising tool in solving clinical problems to help with diagnosis, prognosis and individualized therapy by providing invaluable molecular insights in specimens. 3D and single-cell resolution imaging using MS are also emerging as new frontiers. To help with analyzing the large imaging datasets produced with modern instruments, statistical workflows and the state-of-the-art ML approaches, as well as the improvements in software and hardware will greatly promote the MSI technologies to be employed in the daily clinical routine practice. It is also conceivable that in the near future, automation will be more accessible in generating MSI data, which will lead to a continuous exponential growth in the data volume. In the era of big data, more powerful algorithms are in great need to make automatic, efficient and easy-to-use software to meet the demand for the automated data analysis.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

1. Liu X, Hummon AB. Mass spectrometry imaging of therapeutics from animal models to three-dimensional cell cultures. *Anal. Chem.* 87(19), 9508–9519 (2015).
2. Huang KT, Ludy S, Calligaris D *et al.* Rapid mass spectrometry imaging to assess the biochemical profile of pituitary tissue for potential intraoperative usage. *Adv. Cancer Res.* 134, 257–282 (2017).
3. Buchberger AR, DeLaney K, Johnson J, Li L. Mass spectrometry imaging: a review of emerging advancements and future insights. *Anal. Chem.* 90(1), 240–265 (2018).
4. Galli M, Zoppis I, Smith A, Magni F, Mauri G. Machine learning approaches in MALDI-MSI: clinical applications. *Expert Rev. Proteomics* 13(7), 685–696 (2016).
5. Smith A, Piga I, Galli M *et al.* Matrix-assisted laser desorption/ionisation mass spectrometry imaging in the study of gastric cancer: a mini review. *Int. J. Mol. Sci.* 18(12), 2588 (2017).
6. Galli M, Zoppis I, De Sio G *et al.* A support vector machine classification of thyroid bioptic specimens using MALDI-MSI data. *Adv. Bioinformatics* 2016, 3791214 (2016).
7. Meding S, Nitsche U, Balluff B *et al.* Tumor classification of six common cancer types based on proteomic profiling by MALDI imaging. *J. Proteome Res.* 11(3), 1996–2003 (2012).
8. Hardesty WM, Kelley MC, Mi D, Low RL, Caprioli RM. Protein signatures for survival and recurrence in metastatic melanoma. *J. Proteomics* 74(7), 1002–1014 (2011).
9. Lou S, Balluff B, Cleven AHG, Bovée JVMG, McDonnell LA. An experimental guideline for the analysis of histologically heterogeneous tumors by MALDI-TOF mass spectrometry imaging. *Biochim. Biophys. Acta* 1865(7), 957–966 (2017).
10. Rauser S, Marquardt C, Balluff B *et al.* Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. *J. Proteome Res.* 9(4), 1854–1863 (2010).
11. Balluff B, Elsner M, Kowarsch A *et al.* Classification of HER2/neu status in gastric cancer using a breast-cancer derived proteome classifier. *J. Proteome Res.* 9(12), 6317–6322 (2010).
12. Oppenheimer SR, Mi D, Sanders M, Caprioli RM. A molecular analysis of tumor margins by MALDI mass spectrometry in renal carcinoma. *J. Proteome Res.* 9(5), 2182–2190 (2010).
13. Eberlin L, Norton I, Dill A, Golby A. Classifying human brain tumors by lipid imaging with mass spectrometry. *Cancer Res.* 72(3), 645–654 (2012).
14. Nagpal A, Jatain A, Gaur D. Review based on data clustering algorithms. *Inf. Commun. Technol. 2013 IEEE Conf.* 13, 298–303 (2013).
15. Deininger S-O, Ebert MP, Fütterer A, Gerhard M, Röcken C. MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *J. Proteome Res.* 7(12), 5230–5236 (2008).

16. Willems SM, Van Remoortere A, Van Zeijl R, Deelder AM, McDonnell LA, Hogendoorn PC. Imaging mass spectrometry of myxoid sarcomas identifies proteins and lipids specific to tumour type and grade, and reveals biochemical intratumour heterogeneity. *J. Pathol.* 222(4), 400–409 (2010).

17. Inglese P, McKenzie JS, Mroz A *et al.* Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer. *Chem. Sci.* 8(5), 3500–3511 (2017).

18. Hinsch A, Buchholz M, Odinga S *et al.* MALDI imaging mass spectrometry reveals multiple clinically relevant masses in colorectal cancer using large-scale tissue microarrays. *J. Mass Spectrom.* 52(3), 165–173 (2017).