Intraoperative prediction of tumor cell concentration from Mass Spectrometry Imaging

Vandana Mohan, Ivan Kolesov, Ferenc A. Jolesz, Nathalie Y.R. Agar and Allen R. Tannenbaum

***Abstract*— This work is motivated by the problem of accurately locating tumor boundaries during brain tumor surgery. Currently, such boundary is typically localized using preoperative images and neuronavigation tools. While improved prognosis is associated with minimal residual tumor, an added challenge arises in surgical-decision making to completely excise the tumor and preserve eloquent cortex. We propose that an objective assessment of patterns of tumor cell concentration will help in performing this boundary location by identification of local minima of the tumor cell concentration as tumor boundaries. In this work, we aim to relate the mass spectrometry data - acquired from tissue sections by the Desorption Electrospray Ionization (DESI) approach - to histopathological scores of tumor cell concentration (as evaluated by the neuropathology expert), towards demonstrating that a system can be trained apriori on available tissue samples with known scores, and can be used intraoperatively as an integrated DESI probe to predict the score of the tissue under analysis. We apply the Relevance Vector Machine technique towards learning a "model" that allows us to estimate the tumor cell concentration given the mass spectra. We quantify the performance of this model by testing the framework on real mass spectrometry data acquired from brain tumors (gliomas) of different grades and subtypes with promising results in prediction, and further motivate its intraoperative application.**

I. Introduction

Surgery is the most important, and often the first treatment modality for the majority of brain tumors. In neurosurgical interventions for brain cancers, the principal challenge is to maximize the resection of tumor, while minimizing the potential for neurological deficit by preserving critical tissue. We aim to develop a real-time analysis of the tissue using mass spectrometry in conjunction with radiology for image-guided neurosurgery. A surgical probe is being integrated with an atmospheric pressure Desorption Electrospray Ionization (DESI) mass spectrometry source [1], [2] and a neuronavigation system to allow Real-Time Stereotactic Mass Spectrometry (RTSMS) analysis of surgical tissue for the molecular detection of tumor margins. Mass spectrometry is introduced as a molecular imaging approach, complementing or bypassing needs for systemic injections of molecular probes required for standard methods. In this paper, we propose a framework where we use a training set of multiple subjects and build a predictive model to determine tumor cell concentration from Mass Spectrometry Imaging (MSI), towards detecting tumor margins intraoperatively. In the absence of a physical model linking the data (profile spectra) to the output (the histopathological score), we frame the problem as one of regression using the Relevance Vector Machine (RVM) techniques of Tipping et al [3]. In Section III, we present the theory behind the proposed framework, specifically the details on the preprocessing of the raw Mass Spectrometry data towards denoising and feature(/peak) selection, the details of the RVM-based model for the tumor cell concentration as a *function* of the spectra and the details of the implementation of the proposed framework, complete with the libraries and parameters used to obtain the results reported in this paper. In Section IV, we discuss the experiments conducted to validate the proposed framework, and present the results obtained. Finally, in Section V, we make concluding remarks and present ideas for extending this work in the future.

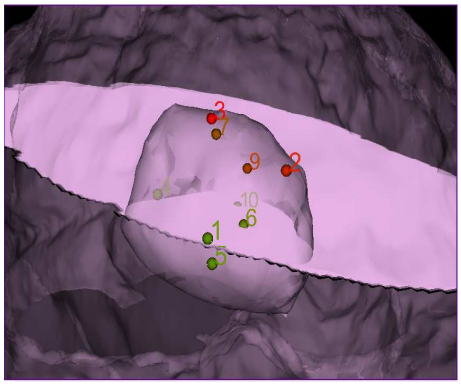


Fig. 1. Visualization of stereotactic sampling sites on tumor surface for 1 subject

In Figure I, we show the stereotactic sampling sites visualized throughout the tumor for 1 subject with site markers colored by the absolute intensity values of the *m/z* 653.5834 (selected to illustrate the variance of molecular distribution in the sampling sites) and sized proportional to the true size of the samples acquired for analysis. The numbers 1 to 10 indicate the order of acquisition. Figure 2 shows a sample profile spectrum and corresponding stained tissue section.

II. Prior Work

Mass spectrometry is a well-established analytical technique used for the identification and characterization of

V. Mohan, I. Kolesov and A. Tannenbaum are with the Schools of Electrical & Computer and Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA; Tannenbaum is also with the Department of Electrical Engineering, Technion-IIT, Haifa, Israel gth115a@mail.gatech.edu, tannenba@ece.gatech.edu

N.Y.R. Agar and Ferenc A. Jolesz are with the Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA nagar@bwh.harvard.edu,

ISBN 978-963-311-370-7

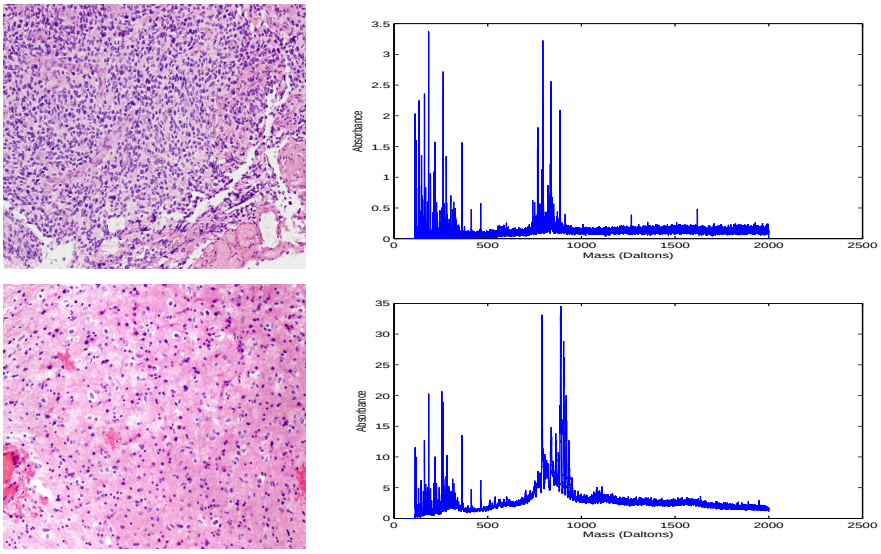


Fig. 2. Sample tissue section stained with hematoxylin and eosin (H&E)s (left) and corresponding Profile Spectra (right): tumor cell concentration of 90% (top row) and 30% (bottom row)

molecules based upon their mass. With the recent adaptation of mass spectrometers to accommodate direct tissue analysis, an increasing number of studies support the usage of mass spectrometry as an efficient tool to detect and delineate pathology directly from tissue specimens [4], [5]. In 2004, Desorption ElectroSpray Ionization (DESI) was introduced as an atmospheric pressure ionization source and subsequently proven to be applicable to tissue imaging [1], [2], [6]. In addition to being a soft ionization method that allows the analysis of intact biomolecules, DESI provides direct sampling of the surface of interest, overcoming time limitations that normally result from specialized sample preparation. This technique paves the way for the integration of Mass Spectrometry into the clinical environment for *in vivo* molecular imaging of tumors, without the systemic injection of contrast agents. Primary brain tumors span a large range of cell types and grades (World Health Organization (WHO) classification system [7]); moreover within individual tumors there may be considerable heterogeneity leading to sampling error. Since the appropriate choice of treatment, as well as estimates of prognosis, rely upon an accurate diagnosis, current histopathological approaches need to be augmented by detailed molecular profiles from an individual's tumor to maximize treatment efficiency.

Quantitative analysis of Mass Spectrometry imaging involves two broad classes of tasks: preprocessing (which includes denoising and peak selection) and statistical analysis (classification, modeling, etc.). An excellent comparison between various feature selection methods such as the student t-test and the P-test as well as an evaluation of dimensionality reduction techniques (such as PCA) as applied to MS data, can be found in [8]. In [9], the authors use the undecimated discrete wavelet transform to decompose mass spectra into noise and signal components, and combine this denoising step with a baseline correction towards improved peak detection, and apply this to the analysis of breast cancer data. In [10], the authors put forth a work flow view of MS image processing and illustrate the value of preprocessing particularly with regard to the benefits associated with denoising. The work in [11] puts forth a novel approach for dimensionality reduction in mass spectrometry data, by combining binning with KS-test based feature selection and further reducing the space dimension by restricting the coefficient of variation. Most existing literature with respect to statistical analysis addresses the classification of mass spectrometry data (e.g. into healthy/diseased) and the identification of biomarkers, and the modeling of mass spectrometry data. We note that the majority of this literature when it uses histological information for analysis, primarily uses it in classification (for training in a supervised classification paradigm) or for validation of statistical frameworks. [12], [13], [14] which address classification and [15], [16] which address the identification of biomarkers towards the characterization/understanding of a disease are notable examples under category. However, the current work falls under the mathematical purview of modeling and we note that prior work in the field of modeling with regard to mass spectrometry data is focused greatly in modeling mass spectra themselves as a function of time and/or *m/z*. For example, [17] puts forth a framework for efficiently pooling mass spectra by modeling the statistical properties of single-shot spectra in the framework of linear regression, which yields a maximum-likelihood estimator that is then employed in pooling at each time interval, and [18] puts forth a regression-based approach for protein expression profile modeling. Note that while these are very effective modeling approaches, their fundamental application lies in the preprocessing of mass spectrometry data. Our interest however lies in modeling histopathological information (specifically the tumor cell concentration) as a function of mass spectra,

and to the best of our knowledge, this is the first work that addresses this specific modeling problem.

III. Proposed Framework for Estimation of Tumor Cell Concentration

A. Modeling Tumor Cell Concentration

We formulate the problem of modeling the tumor cell concentration as a function of profile spectra, in the linear regression framework. We have the following requirements for the technique to be used - the ability to handle limited number of training inputs, the ability to generalize well and low computational cost. Hence, we propose to use Relevance Vector Machines (RVM) [3], [19] based regression for this problem, using a Gaussian kernel.

Let $\mathbf{x}_{(i,j)}$ denote the profile spectrum of the i-th sampling site from the j-th subject in our training population, and let $h(\mathbf{x})$ denote the tumor cell concentration of the profile spectrum given by $\mathbf{x}$.

In the regression framework, $\mathbf{x}$ are the input vectors and the scalar target t can be written as:

$$t = h(\mathbf{x}) + \epsilon \tag{1}$$

where the error ϵ is modeled as independent zero-mean gaussian with variance σ^2.

We model $h(\mathbf{x})$ as:

$$h(\mathbf{x}) = \sum_{m=1}^{M} w_m \phi_m(\mathbf{x}), h = \phi(\mathbf{x})\mathbf{w} \tag{2}$$

By the error model assumed, we get:

$$p(t_n|\mathbf{x}_n) = N(t_n|h(\mathbf{x}_n, \mathbf{w}), \sigma^2) \tag{3}$$

Since t_n are assumed to be independent, the likelihood of the complete data set can be written as:

$$p(t|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{\frac{-N}{2}} \exp\{\frac{-1}{2\sigma^2}\|t - \phi\mathbf{w}\|\} \tag{4}$$

By the RVM approach, here we constrain the parameters $\mathbf{w}$ by prior distributions using hyperparameters α as follows:

$$p(\mathbf{w}|\alpha) = \prod_{i=1}^{M} N(w_i|0, \alpha_i^{-1}) \tag{5}$$

Using Bayes' rule, we can write the posterior parameter distribution conditioned on the data as follows:

$$p(\mathbf{w}|t, \alpha, \sigma^2) = \frac{p(t|\mathbf{w}, \sigma^2)p(\mathbf{w}|\alpha)}{p(t|\alpha, \sigma^2)} = N(\mu, \Sigma) \tag{6}$$

where

$$\Sigma = (A + \sigma^2\Phi^T\Phi)^{-1} \tag{7}$$

$$\mu = \sigma^2\Sigma\Phi^T t \tag{8}$$

$$A = diag(\alpha_1, \alpha_2 ... \alpha_M) \tag{9}$$

Formulating the Sparse Bayesian Learning as a Type-II Maximum Likelihood procedure, we write the Marginal Likelihood as:

$$L(\alpha) = \log p(t|\alpha, \sigma^2) \tag{10}$$

$$= \frac{-1}{2}\left[N\log(2\Pi) + \log|C| + t^T C^{-1} t\right] \tag{11}$$

with

$$C = \sigma^{-2} I + \Phi A^{-1} \Phi^T \tag{12}$$

By maximizing $L(\alpha)$ with respect to α and σ, we get maximum probable estimates of the two quantities as α_{MP} and σ_{MP} respectively.

We finally obtain the predictive distribution for any new input $\mathbf{x}_{new}$ as:

$$p(t_{new}|t, \alpha_{MP}, \sigma^2_{MP}) \tag{13}$$

$$= \int p(t_{new}|\mathbf{w}, \sigma^2_{MP}) p(\mathbf{w}|t, \alpha_{MP}, \sigma^2_{MP}) d\mathbf{w}$$

$$= N(t_{new}|h_{new}, \sigma^2_{new}) \tag{14}$$

where

$$h_{new} = \mu_{MP}^T \Phi(\mathbf{x}) \tag{15}$$

$$\mu_{MP} = \sigma_{MP}^{-2} \Sigma \Phi^T t$$

$$\sigma^2_{new} = \sigma^2_{MP} + \Phi(\mathbf{x}_{new})^T \Sigma \Phi(\mathbf{x}_{new})$$

Thus, given a new profile spectrum x_{new}, we will use Equation 15 to predict the tumor cell concentration.

B. Preprocessing

The high dimensionality of mass spectrometry poses a challenge in the proposed framework. Firstly, since the number of samples typically available for training such a framework are limited, high dimensionality of input reduces the robustness of the *learned* model. Secondly, the dimensionality of the input directly affects the complexity of the framework. To address these issues, we employ peak selection techniques to select the variables to regress over. We experiment with two techniques: using t-statistics and KS-statistics. We compute t-statistics (KS-statistics) with respect to the binned tumor cell concentration of the input samples (as classified by the expert as part of the histopathological evaluation). We then order the t-statistics (KS-statistics) and select the masses with highest (lowest) values of t-statistics (KS-statistics) for use in the modeling step i.e. these are the peaks at which the kernels in the model of Equation 2 are centered to begin with.

C. Implementation

In the scope of this work, we implement the proposed framework in MATLAB in conjunction with C++ code utilizing the *Dlib* open source library [20]. The preprocessing and peak selection are performed in MATLAB and the RVM-based modeling is carried out using MEX code that employs the *Dlib* library. We use the sigmoid kernel in the RVM model.

IV. Results

A. Data

The DESI Imaging data used for the experiments in this work was acquired from tissue specimens prepared at the Brigham and Women's Hospital, and analyzed in the Aston Laboratories at Purdue University. A lab-built prototype (configured as described in [2]) was used for the DESI ion source. The spray solvent used for the acquisition was

methanol:water (50:50) with the application of a 5 kV spray voltage, the nitrogen gas pressure was 150 psi and the solvent flow rate was 1.5 L/min. 2D images were recorded consisting of arrays of pixels, with each pixel covering an area of 200 x 200 m2, and the samples were then analyzed in negative (*m/z* 150 1000) ion modes. The MS data was acquired using a LTQ linear ion trap mass spectrometer controlled by the XCalibur 2.0 software (from Thermo Fisher Scientific, San Jose, CA, USA). Spatially accurate images were then assembled using the software Biomap (freeware). Finally, for the modeling experiments presented in this paper, five profile spectra corresponding to five different regions of each specimen were used. In all, 28 tissue samples of gliomas were used leading to 140 mass spectra as input for the modeling process. The glioma samples span different subtypes (anaplastic astrocytoma, glioblastoma, oligoastrocytoma and oligodendroglioma) and three tumor grades (WHO grades 2-4). Also, histopathological evaluation was performed by a pathologist on all the samples, yielding assessment of tumor cell concentration per sample, which was also an input into the modeling process.

B. Experiments and Results

We focus in this study on the mass spectrometry data between (*m/z*) 650-1000 Daltons, since we are focused on understanding and using the relationship of the lipid profile to all characteristics of gliomas (including grades, subtypes and tumor cell concentration). We perform peak selection using two techniques: t-statistics and KS-statistics. We bin the tumor cell concentration into three categories ($0-35\%$, $35-70\%$ and $70-100\%$) and use these categories to divide the available data into subgroups over which we compute the statistics. We then sort the statistics and select the masses with the highest(lowest) t-statistics (KS-statistics) as peaks to regress over. We conduct experiments to compare the performance of the two peak selection approaches, as well as the number of peaks used. We also explore via our test cases the effect of preliminary normalization on the mass spectrometry data.

We present our results from the comparison of preprocessing schemes in Table IV-B. We compare four schemes against the original data - scaling to lie in the $(0,1)$ range, scaling by the mean and standard deviation, scaling by the minimum and standard deviation, and preliminary baseline correction using a moving average. Figure 3 shows a visual comparison of these preprocessing schemes with the original data, for three sample spectra chosen to represent three different tumor cell concentration levels ($5\%, 40\%, 90\%$). Table IV-B shows the corresponding quantitative comparison of these schemes, with the modeling and prediction errors that were obtained for the respective schemes while using a sigmoid kernel-based RVM for modeling with 4 peaks selected using t-statistics. From extensive experiments over both peak selection methods and over different numbers of peaks, we observe that the preprocessing scheme that yields the best performance for this data set is the one where we scale the data by standard deviation and force the minimum to zero.

Finally, we conducted experiments to ascertain the optimal number of peaks to employ in this framework and these results are presented in Table IV-B. We compare a wide range of values for the number of peaks with both peak selection approaches. In looking at the modeling and prediction errors, we find that for this data set, 4 and 5 are the optimal number of peaks for the t-statistics-based and KS-statistics- based approaches respectively. Using a leave-N-out approach, we train the model on 80% of the available data set and test it on the remaining 20%. The errors from the respective sets are designated here as the modeling and prediction errors respectively, with these numbers being estimated as the average over a 100 independent trials. It is worth noting that while it can be expected that increasing the number of peaks would increase accuracy in the model by allowing more degrees of freedom, in our study, we observed diminishing returns when increasing the number of peaks beyond the optimal described above. We attribute this to the limited size of the population that we use for training the model, and we expect that increasing the number of samples analyzed will lead to higher accuracy (and allow this for a greater number of peaks).

V. Conclusion and Future Work

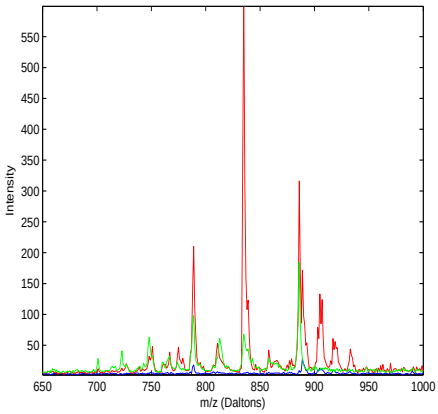
In this paper, we proposed a novel method for intraoperative prediction of tumor cell concentration from MSI, based on a Relevance Vector Machine approach to be used in conjunction with appropriate peak selection methods for dimensionality reduction. The prediction of tumor cell concentration, given a previously unseen spectrum acquired by the DESI-MS modality, was found to be achievable in real-time which makes an excellent case towards integrating the framework with the surgical workflow.

The framework yields promising results with a reasonably low error of 14% as tested by a leave-N-out approach to quantifying the prediction capability of the *learned* model. These results are especially promising given the limited size of the population we regress over in training the RVM. This also motivates our first goal for future work on this front, which is to increase the robustness of the modeling results and identify more suitable peaks by including a larger population of mass spectra into the analysis. We will also explore alternate peak selection and preprocessing techniques. We will especially pursue approaches to increase the model accuracy and its prediction capability by reducing inter-subject, inter-subtype and inter-grade variance. We aim to develop appropriate techniques for normalization of mass spectra towards this goal. Finally, we will address the issues of tumor density (as different from tumor cell concentration which is used in this work) and spatial interpolation, towards enabling the intraoperative use of the proposed framework for tumor boundary location.

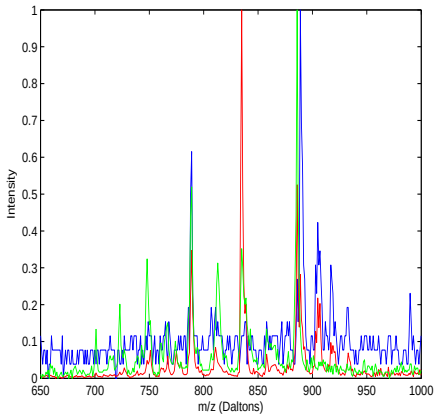
Acknowledgements

This work was supported in part by grants from NSF, AFOSR, ARO, as well as by a grant from NIH (NAC

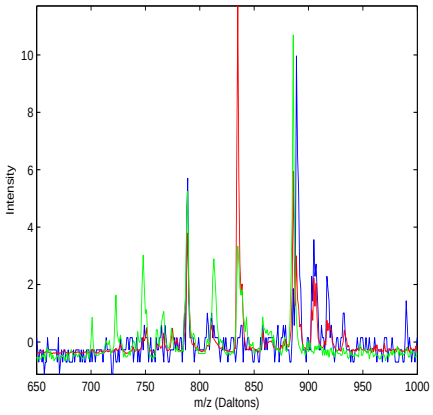
Fig. 3. Visual comparison of preprocessing schemes: sample spectra with tumor cell concentrations of 5%(red), 40%(green) and 95% (blue)



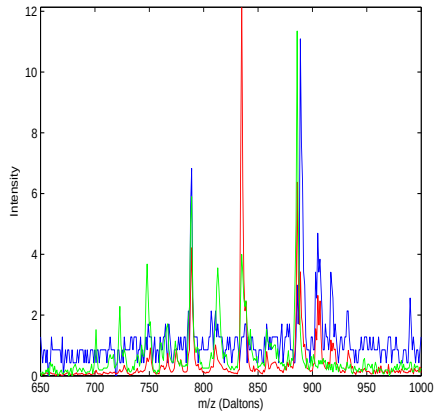
(a) Original spectra



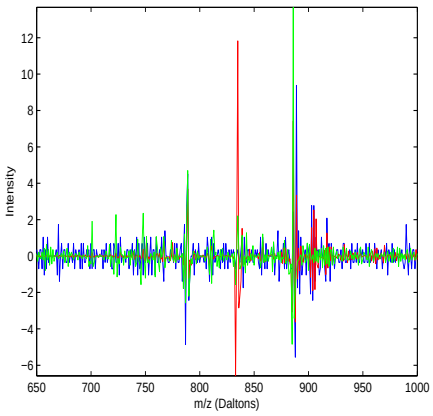
(b) Scaling to lie in $(0, 1)$



(c) Scaling by mean and standard deviation



(d) Scaling by standard deviation, with 0 minimum



(e) Basic baseline correction (using moving averages)

TABLE I

COMPARISON OF PREPROCESSING SCHEMES: MODELING AND PREDICTION ERRORS USING THE SIGMOID KERNEL FOR MODELING, WITH T-STATISTICS-BASED SELECTION OF 4 PEAKS

Preprocessing scheme	Modeling Error (%)	Prediction Error(%)
None *Original data*	25.80	25.72
Scaling (to lie in [0,1])	24.22	23.95
Scaling (by mean and standard deviation)	25.21	25.31
Scaling (by standard deviation) and minimum zero	13.21	14.59
Baseline correction (by moving average)	22.66	25.48

TABLE II

QUANTIFYING THE MODEL PERFORMANCE: COMPARING DIFFERENT PEAK SELECTION METHODS AND THE NUMBER OF PEAKS USED

Peak selection method	Number of peaks	Modeling error (%)	Prediction error (%)
T-statistics	50	25.86	25.66
T-statistics	20	25.78	25.82
T-statistics	10	20.40	20.28
T-statistics	6	18.75	18.84
T-statistics	5	17.34	17.71
T-statistics	**4**	**13.21**	**14.59**
T-statistics	3	19.31	21.45
KS-statistics	50	25.74	25.94
KS-statistics	20	14.05	14.72
KS-statistics	10	13.37	14.22
KS-statistics	6	13.38	14.13
KS-statistics	**5**	**13.31**	**14.02**
KS-statistics	4	13.36	14.31
KS-statistics	3	14.07	15.01

P41 RR-13218) through Brigham and Women's Hospital. This work is part of the National Alliance for Medical Image Computing (NAMIC), funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54 EB005149. Information on the National Centers for Biomedical Computing can be obtained from `http://nihroadmap.nih.gov/bioinformatics`.

The authors would like to thank Livia S. Eberlin and Dr. R. Graham Cooks of the Aston Labs at Purdue University for their collaboration in the acquisition of the DESI-MS data used in this work, and their valuable inputs that aided in our understanding of this data.

REFERENCES

[1] Z. Takats, J. Wiseman, B. Gologan, and R. Cooks, "Mass spectrometry sampling under ambient conditions with desorption electrospray ionization," pp. 471–473, 2004.

[2] R. Cooks, Z. Ouyang, Z. Takats, and J. Wiseman, "Ambient mass spectrometry," pp. 1566–1570, 2006.

[3] M. Tipping, "Relevance vector machine," Oct. 14 2003, uS Patent 6,633,857.

[4] R. Caldwell and R. Caprioli, "Tissue profiling by mass spectrometry: a review of methodology and applications," *Molecular & Cellular Proteomics*, vol. 4, no. 4, p. 394, 2005.

[5] P. Chaurand, M. Sanders, R. Jensen, and R. Caprioli, "Proteomics in diagnostic pathology: profiling and imaging proteins directly in tissue sections," *American Journal of Pathology*, vol. 165, no. 4, p. 1057, 2004.

[6] J. Wiseman, S. Puolitaival, Z. Takats, R. Cooks, and R. Caprioli, "Mass spectrometric profiling of intact biological tissue by using desorption electrospray ionization," *Angew. Chem. Int. Ed*, vol. 44, pp. 7094–7097, 2005.

[7] J. Smirniotopoulos, "The new WHO classification of brain tumors." *Neuroimaging Clinics of North America*, vol. 9, no. 4, p. 595, 1999.

[8] I. Levner, "Feature selection and nearest centroid classification for protein mass spectrometry," *BMC bioinformatics*, vol. 6, no. 1, p. 68, 2005.

[9] K. Coombes, S. Tsavachidis, J. Morris, K. Baggerly, M. Hung, and H. Kuerer, "Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform," *Proteomics*, vol. 5, no. 16, pp. 4107–4117, 2005.

[10] J. Norris, D. Cornett, J. Mobley, M. Andersson, E. Seeley, P. Chaurand, and R. Caprioli, "Processing MALDI mass spectra to improve mass spectral direct tissue analysis," *International journal of mass spectrometry*, vol. 260, no. 2-3, pp. 212–221, 2007.

[11] J. Yu, S. Ongarello, R. Fiedler, X. Chen, G. Toffolo, C. Cobelli, and Z. Trajanoski, "Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data," *Bioinformatics*, vol. 21, no. 10, p. 2200, 2005.

[12] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q. Le, "Sample classification from protein mass spectrometry, by peak probability contrasts," *Bioinformatics*, vol. 20, no. 17, p. 3034, 2004.

[13] M. Wagner, D. Naik, and A. Pothen, "Protocols for disease classification from mass spectrometry data," *PROTEOMICS-Clinical Applications*, vol. 3, no. 9, pp. 1692–1698.

[14] N. Agar, J. Malcolm, V. Mohan, H. Yang, M. Johnson, A. Tannenbaum, J. Agar, and P. Black, "Imaging of Meningioma Progression by Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry," *Analytical Chemistry*, pp. 97–109.

[15] E. Diamandis, "Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations," *Molecular & Cellular Proteomics*, vol. 3, no. 4, p. 367, 2004.

[16] J. Li, Z. Zhang, J. Rosenzweig, Y. Wang, and D. Chan, "Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer," *Clinical Chemistry*, vol. 48, no. 8, p. 1296, 2002.

[17] M. Skold, T. Ryden, V. Samuelsson, C. Bratt, L. Ekblad, H. Olsson, and B. Baldetorp, "Regression analysis and modelling of data acquisition for SELDI-TOF mass spectrometry," *Bioinformatics*, vol. 23, no. 11, p. 1401, 2007.

[18] H. Liu, A. Bonner, and A. Emili, "Modeling protein tandem mass spectrometry data with an extended linear regression strategy," in *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, vol. 2, 2004.

[19] C. Bishop and M. Tipping, "Variational relevance vector machines," in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, 2000, pp. 46–53.

[20] D. King, "dlib c++ library," March 2010, `http://dlib.net/`.