Bioinformatics, 31, 2015, i142–i150
doi: 10.1093/bioinformatics/btv251
ISMB/ECCB 2015

OXFORD

Deconvolving molecular signatures of interactions between microbial colonies

Y.-C. Harn[1], M. J. Powers[2], E. A. Shank[2,3,4] and V. Jojic[1,*]

[1]Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599-3175, USA, [2]Department of Biology, University of North Carolina, Chapel Hill, NC 27599-3280, USA, [3]Department of Microbiology and Immunology, University of North Carolina, Chapel Hill, NC 27599-7290, USA and [4]Curriculum of Genetics and Molecular Biology, University of North Carolina, Chapel Hill, NC, USA

*To whom correspondence should be addressed.

Abstract

Motivation: The interactions between microbial colonies through chemical signaling are not well understood. A microbial colony can use different molecules to inhibit or accelerate the growth of other colonies. A better understanding of the molecules involved in these interactions could lead to advancements in health and medicine. Imaging mass spectrometry (IMS) applied to co-cultured microbial communities aims to capture the spatial characteristics of the colonies' molecular fingerprints. These data are high-dimensional and require computational analysis methods to interpret.
Results: Here, we present a dictionary learning method that deconvolves spectra of different molecules from IMS data. We call this method MOLecular Dictionary Learning (**MOLDL**). Unlike standard dictionary learning methods which assume Gaussian-distributed data, our method uses the Poisson distribution to capture the count nature of the mass spectrometry data. Also, our method incorporates universally applicable information on common ion types of molecules in MALDI mass spectrometry. This greatly reduces model parameterization and increases deconvolution accuracy by eliminating spurious solutions. Moreover, our method leverages the spatial nature of IMS data by assuming that nearby locations share similar abundances, thus avoiding overfitting to noise. Tests on simulated datasets show that this method has good performance in recovering molecule dictionaries. We also tested our method on real data measured on a microbial community composed of two species. We confirmed through follow-up validation experiments that our method recovered true and complete signatures of molecules. These results indicate that our method can discover molecules in IMS data reliably, and hence can help advance the study of interaction of microbial colonies.
Availability and implementation: The code used in this paper is available at: https://github.com/frizfealer/IMS_project.
Contact: vjojic@cs.unc.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

1.1 Background

Microbial metabolites have been of great importance as a source of clinically relevant bioactive compounds. One microbial colony may secrete different metabolites, or alter the quantity of certain metabolites when encountering other colonies. Understanding the effects of co-culturing may provide a means to manipulate metabolite production. However, few studies have investigated how changes in microbial community composition shape metabolite production by its members (Hoefler *et al.*, 2012; Watrous *et al.*, 2012). One technology useful for such investigations is matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) imaging mass spectrometry (IMS), or MALDI-IMS. MALDI-IMS has been successfully applied to a variety of biological systems over the last decade. It is a promising tool because it can be used on a wide array of sample types. In addition, it has ability to provide information about spatial distribution of molecular species abundances (Alexandrov, 2012). However, the molecular species and their abundances are not directly measured. Rather, they are reflected in ion counts across over $10^4 - 10^5$ mass-charge ratios. Each molecular species may contribute to multiple mass-charge ratio measurements and different

molecular species may contribute to the same measurement. An ability to uncover the molecular species and their abundances from these mass spectra would make MALID-IMS especially useful in unraveling the mechanisms of interactions in microbial communities. Due to the complexity of these data, direct inspection is not practicable. Thus, it is natural to resort to statistical methods to obtain succinct summaries of MALDI-IMS data.

Computational and statistical methods for MALDI-IMS analysis have been proposed (Alexandrov, 2012; Jones *et al.*, 2012; Kobarg *et al.*, 2013; Trede *et al.*, 2012). These methods can be divided into several groups depending on their focus. One group of methods applies supervised learning, focusing on finding important features (biomarkers) in the data for purposes of phenotype classification. Such approaches include use of symbolic discriminant analysis (Lemaire *et al.*, 2007), Support Vector Machine (Groseclose *et al.*, 2008), genetic algorithms (Cazares *et al.*, 2009), elastic net (Hong and Zhang, 2010) and aNN (Rauser *et al.*, 2010). Another group of methods focuses on segmentation of IMS data with unsupervised clustering of spectra (Alexandrov *et al.*, 2013a,b).

Our work, in contrast, belongs to a group of methods focusing on concise data representation. Previous methods in this group used principal component analysis (PCA) (Leendert *et al.*, 2007; Plas *et al.*, 2007), independent components analysis (ICA), non-negative matrix factorization (NNMF) (Siy *et al.*, 2008) and probabilistic latent semantic analysis (pLSA) (Hanselmann *et al.*, 2008). All of these methods showed good results in decomposing IMS signals into the components of biological interest. PCA and ICA use negative values in the decomposition; negative values do not have a natural interpretation as components of the mass spectrum or abundances of molecular species. On the other hand, NNMF and pLSA produce non-negative components. Nevertheless, NNMF uses alternating least squares aimed at continuous values to decompose a count data matrix. This approach lacks a generative model and thus its output, while non-negative, is still not easily interpreted. Finally, pLSA has a generative model for the data, allowing for a simple interpretation, but its application has been focused only on modeling an anatomical area in tissues or other biological meaningful unit that corresponds to a composition of a large array of different compounds. For example, the pLSA implemented in Bruker's software (ClinProTools 3.0) was used in analyzing a tumor dataset in MALDI-IMS (Hanselmann *et al.*, 2008; Sören-Oliver and and Klaus Meyer, 2015). The report showed that the method can recover signature spectra of cell types from the data. However, identification of cell types can be accomplished using few discriminative spectra. Hence, this approach is not suitable for fine-grained analysis of the whole repertoire of microbial metabolites.

In this article, we model IMS data in terms of molecular signatures and abundances of these molecules. By decomposing mass spectra into contributions from different molecular species, we can infer directly both novel molecules and their abundances. To accomplish this, we propose a Dictionary Learning method, drawing on work in machine learning (Balasubramanian *et al.*, 2012; Lee *et al.*, 2006; Maurer *et al.*, 2013; Mehta and Gray, 2013; Olshausen and Field, 1997). We deem this method Molecular Dictionary Learning (**MOLDL**).

The key contributions of our work are outlined next. We introduce the first generative model of IMS data. This model takes into account the count nature of the data, background information on the ionization types and the spatial organization of the data. Recovery of abundances and spectra of molecular species can be seen as probabilistic inference in our model. For this purpose, we derived and implemented an efficient bi-convex optimization algorithm. Crucially, our method does not require users to specify the number of the molecular species in the sample, as this number is uncovered automatically. We conducted computational experiments to demonstrate performance of this method on both synthetic and real datasets.

1.2 Notation and terminology

We will assume that each spectrum is of length s, and we assume $w \times h$ such spectra are arranged on a grid of width w and height h. We will use $\mathbf{y}^{i,j}$ to denote spectrum measured at location i, j on the grid, and $\mathbf{Y}$ to denote all of those spectra. Note that $\mathbf{Y}$ is a tensor of size $s \times w \times h$. To denote measurement of ion counts of i^{th} mass/charge (m/z) in spectrum $\mathbf{y}$, we will write y_i. We will call each location on the grid a **grid cell**.

We will refer to molecules with the same molecular weight and ionization preferences as a **molecular species** (note that different molecules can appear indistinguishable to a measurement technology such as mass spectrometry and contribute to the same molecular species). A measured spectrum can be seen as a linear combination of spectra of different molecular species. We will organize spectra of molecular species into a matrix $\mathbf{D}$, referred to as a **dictionary**, with each column of $\mathbf{D}$ being a single molecular species' spectrum.

We will assume that dictionary has s columns, hence s different dictionary spectra. Hence, we will have a *capacity* to model as many different molecular species as there are different mass-charge ratios. However, this is simply an upper bound on the number of molecular species. We will denote the k^{th} spectrum in the dictionary with $\mathbf{d}_k$. Hence, the i^{th} charge of the k^{th} dictionary element will be a scalar $d_{i,k}$, an entry in the i^{th} row and k^{th} column. To differentiate between spectra in a dictionary and spectra in actual experimental data, we will refer to spectra in the dictionary as **dictionary elements**.

We will refer to level of contributions from different dictionary elements as **abundances** and denote them using $\mathbf{w}$. A vector of abundances $\mathbf{w}$ will be of length s, since we can use at most s dictionary elements. In addition because each position i, j on the grid has its own set of abundances, we will use $\mathbf{W}$, a tensor of size $s \times w \times h$, to denote the set of all abundances on a grid, and $\mathbf{w}^{i,j}$ to denote an abundance vector belonging to a location i, j. For each location i, j, there is a offset w_0 added in the linear combination of predictors. Therefore $\mathbf{W}_0$ is a matrix of size $w \times h$ to denote the set of all offsets on a grid. We will use $\mathbf{1}$ to denote a vector of all 1s.

2 Methods

In order to explain our model, we will incrementally approach the full model by first introducing dictionary construction in Section 2.1, followed by the space independent model in Section 2.2, culminating in space dependent model in Section 2.3. We show that both space-dependent and space-independent models give rise to biconvex objectives. We then provide an algorithm that can fit both models in Section 2.5. We must emphasize here that the inputs of this algorithm are only the IMS-data $\mathbf{Y}$ and a dictionary pattern that is based on MALDI mechanism and is independent of data. Molecular species number that is an input of many dimension reduction algorithm is learned automatically by our algorithm.

2.1 Constructing a dictionary non-zero pattern based on MALDI-IMS's prior knowledge

Only ionized molecules have signals in mass spectrometry. Molecules, depending on their characteristics, can be ionized with positive or negative charge; therefore, mass spectrometry with

different charge modes are applied to detect different molecules. Different molecules have different ionization preferences. Because MALDI mass spectrometry is a soft ionization technique, molecules are rarely fragmented; however, numerous ion adducts are possible. Common ion types of MALDI-TOF in both positive and negative mode often observed in microbial MALDI-IMS data are listed in Table 1 (Gross, 2011). Given a putative peak M + H, corresponding to a molecule of weight M, we can compute other possible peaks for this molecule by adding the mass differences between different ionization types. We will refer to the set of differences as Δ. For example, a molecule yielding ions M + H with *m/z* 301.1 in the positive mode can produce other ions with *m/z* values of 301.01 +17.03 (M + NH_4), 301.01 + 21.98 (M + Na), ... , 301.01 + 76.18 (M + 2 K-H). Hence, given an *m/z* value for M + H, denoted by *r*, we compile a full list of putative peaks as $r + \Delta = \{r + \delta | \delta \in \Delta\}$. Since an *m/z* value may come from any of the *six* ion types listed in the table (in the positive mode), a dictionary element associated with a particular M + H ion has *six* candidate non-zero entries corresponding to these peaks (the same idea applies to the negative mode).

For a given molecular species, a dictionary element is allowed to have non-zero entries only corresponding to putative peaks arising due to different ionization types. More precisely, the only non-zero entries will be $D_{\lceil s+0.5ppm \rceil, \lceil s+q+0.5ppm \rceil}$, $q \in s + \Delta$ (0.5 ppm accounts for the measurement error of ±0.5 ppm Da *m/z*). If the *m/z* values we infer do not appear in our data, we discard these values (peaks). We call these non-zero entries in a dictionary a **dictionary pattern**. Importantly, this pattern only determines the sparsity of the dictionary; all putative non-zeros in the dictionary are still treated as parameters that have to be learned. Consequently, if the data show no support for a particular molecular species giving rise to an ion type, the corresponding entry in the dictionary will be zero. An illustration of such a dictionary and its relationship to theoretical data is shown in Figure 1a.

A **dictionary pattern** is constructed based on the prior knowledge of MALDI-IMS, which is shown in Table 1. Using this pattern greatly reduces the problem complexity. For a dataset containing *n* *m/z* values, without any dictionary patterns, a general assumption of it would be that every *m/z* value could come from one molecular species, and every molecular species could gathering any of the *m/z* values. This assumption leads to a dictionary of dimension $O(n^2)$. In contrast, using dictionary pattern, we know one molecular species with molecular weight M only generates the ion types listed in the Table 1, so the number of parameters in a dictionary reduces to $O(cn)$, where *c* is the number of ion types. Therefore, dictionary patterns reduce the chance for overfitting, the number of local minima, and the running time. We demonstrate the utility of using a dictionary pattern in a synthetic experiment.

Table 1. The common ion types and their mass (the row of this table is sorted by their *m/z*, and all *m/z* s listed here are rounded to the second decimal points) for positive and negative mode

(a) Positive mode	
Ionization types	*m/z*
M + H	M + 1.01
M + NH_4	M + 18.04
M + Na	M + 22.99
M + K	M + 39.10
M + 2Na-H	M + 44.97
M + 2K-H	M + 77.19
(b) Negative Mode	
Ionization types	*m/z*
M-H_2O-H	M-19.02
M-H	M-1.01
M-Na-2H	M + 20.98
M + Cl	M + 34.97
M-K-2H	M + 37.08

2.2 Space independent model with poisson noise

In our space independent model, we assume that spectra in neighboring locations on the grid are independent from each other. Furthermore, once we determine the abundances, **w**, of different dictionary elements, different entries of spectrum **y** will be independent from each other. This is a typical independence assumption used in dictionary models. Due to the type of the data collected (i.e. non-negative counts rather than continuous values), we model the data in each y_k as Poisson distributed. Hence, we have

$$y_k | \mathbf{D}, \mathbf{w} \sim \text{Poisson}(\mathbf{D}_{k,:}\mathbf{w} + w_0).$$

We note a difference here in comparison to the standard Poisson regression. Loosely, the Poisson regression can be seen as modeling log counts. However, we assume that the contributions of ions with particular *m/z* from two different compounds add up rather than multiply to yield the measured counts. Hence, in the generalized linear model view, we use an identity link function rather than a logarithm.

We note here that $\mathbf{D}_{k,:}$ is a row of matrix **D**, and hence not the same as $\mathbf{d}_k$ which is a column. This row reflects a proportion of each molecular species ions that have k^{th} *m/z* ratio.

Even though the dictionary could have a large number of potential molecular species, only a subset of ions have non-negligible counts across all grid cells. Moreover, even smaller set of them are measured in any one grid cell. Hence, given our assumption of dictionary structure described in Section 2.1, only a subset of dictionary

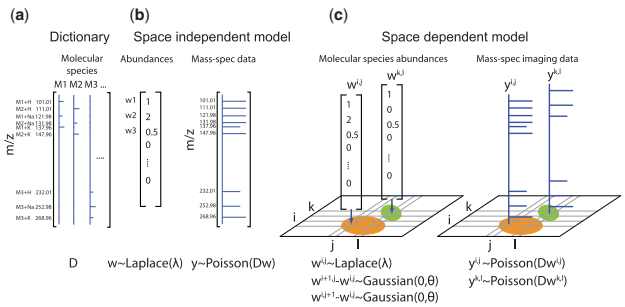

Fig. 1. The dictionary learning framework for deconvolving molecular signatures from MALDI-IMS data. (**a**) Our probabilistic model of IMS data utilizes a dictionary, **D**, to represent peaks associated with molecular species. Each column of the dictionary corresponds to a potential molecular species. A sparsity pattern in a column reflects prior knowledge of possible ion types. (**b**) For our space *independent* model, we assume abundances of most molecular species are zero as we do not expect to see every molecular species present in the sample. A linear combination of molecular species in the dictionary, according to their abundances **w**, gives rise to observed counts, **y**. (**c**) In mass-spec imaging data, spectra are measured on a grid covering the biological sample of interest. In the illustration, orange and green areas correspond to two microbial populations. The grid areas covering the same population are expected to have similar abundances of molecular species. Our space *dependent* model captures this homogeneity expectation by assuming that abundances in nearby locations on grid are frequently similar

elements will be used: those that can produce ions with non-negligible counts. In order to encode this kind of sparsity assumptions, we introduce an ℓ_1 penalty or, equivalently, a Laplace prior, on molecular species abundances. Similarly, we do not assume a priori that each molecular species will generate each of its putative ion types and hence place a similar penalty on dictionary entries as well. This additional penalty is only meaningful for the entries we deemed candidate non-zeros, the rest are by definition zero. Hence, we have

$$\mathbf{w}|\lambda \sim \text{Laplace}(\lambda), \mathbf{D}|\phi \sim \text{Laplace}(\phi)$$

This model is illustrated in Figure 1b.

2.3 Space dependent model with fusion penalty

It is natural to assume that MALDI-IMS measurements taken at nearby locations will have very similar abundances of molecular species. For an illustration see Figure 1c. One way to introduce this assumption in the model is to penalize differences in parameters at neighboring locations. In particular, one such penalty is

$$-\sum_{i,j,l} \theta \|w_l^{i,j} - w_l^{i+1,j}\|_2^2 + \theta \|w_l^{i,j} - w_l^{i,j+1}\|_2^2.$$

This sum of squares penalty promotes shrinkage of *the differences* of abundances between nearest neighbors. Penalties on the parameter differences are referred to fusion penalties. Here, we employ ℓ_2^2, or a sum-of-squares penalty, on the differences.

Hence, the penalized objective for the space dependent model is

$$\begin{aligned}\text{FLP}(\mathbf{W}, \mathbf{W}_0, \mathbf{D}, \lambda, \phi, \theta; \mathbf{y}) = &\sum_{i,j,k} y_k^{i,j} \log\{(D_{k,:}\mathbf{w}^{i,j} + w_0)\} \\ &- \sum_{i,j,k} (D_{k,:} w_l^{i,j} + w_0^{i,j}) - \sum_{i,j,k} \log\{y_k^{i,j}!\} \\ &- \lambda \sum_{i,j,l} |w_l^{i,j}| - \phi \sum_{k,m} |d_{k,m}| \\ &- \sum_{i,j,l} \theta \|w_l^{i,j} - w_l^{i+1,j}\|_2^2 + \theta \|w_l^{i,j} - w_l^{i,j+1}\|_2^2.\end{aligned}$$

This model is illustrated in Figure 1c. **MOLDL** optimizes the function FLP by optimizing $\mathbb{W}$ given D and optimizing D given $\mathbb{W}$ alternatingly. The details of these two algorithms and the details of choosing hyperparameters are shown in the Supplementary data.

2.4 Biconvexity of space dependent model's objective

THEOREM 1:The objective in Equation 2.3 is biconvex in abundances, $\mathbf{W}$ and $\mathbf{W}_0$, and dictionary, $\mathbf{D}$.

PROOF: Sketch: To show that the function FLP is bi-convex in $(\mathbf{W}, \mathbf{W}_0)$ and $\mathbf{D}$, we need to show that the function is convex in $(\mathbf{W}, \mathbf{W}_0)$ for a fixed $\mathbf{D}$ and vice versa. For a fixed $\mathbf{D}$ the objective is a sum of a convex function with an affinely transformed argument $\sum_{i,j,k} \log\{\mathbf{D}_{k,:}\mathbf{w}^{i,j}\}$, a linear functions of $\mathbf{w}$ and w_0, a convex function $|w_l^{i,j}|$ and another convex function with an affinely transformed arguments $\|w_l^{i,j} - w_l^{i,j+1}\|_2^2$. As a sum of convex functions, the function FLP is convex for a fixed $\mathbf{D}$.

For a fixed $(\mathbf{W}, \mathbf{W}_0)$, the function FLP has three terms that are influenced by D. The first term is a convex function with an affinely transformed argument linear function of $\mathbf{D}$, $\sum_{i,j,k} \log\{D_{k,:}\mathbf{w}^{i,j}\}$. The second term is a linear function of D, $-(\sum_{i,j,k} D_{k,:}\mathbf{w}^{i,j})$. The third term is a convex function $|d_{k,m}|$. As a sum of convex function, the function FLP is convex for a fixed $\mathbf{W}, \mathbf{W}_0$. Hence, the function FLP is biconvex in abundances and dictionary. □

The statement of biconvexity holds even for $\theta = 0$ so fitting the space independent model is also a biconvex optimization problem.

2.5 Algorithm

Algorithm 1 shows pseudo-code for **MOLDL**. We initialize $\mathbf{z}_0$ as $\log(\mathbf{y} + 1)$ and $\mathbf{D}^{init}$ can be any matrix which honors the non-zero pattern constructed in Section 2.1. Here, we use NNMF on non-zero entries for dictionary initialization in real cases. Also, we use 'updateW-ADMM' to refer to an implementation of $\mathbf{W}$ updates, outlined earlier. The algorithm iteratively updates $\mathbf{W}, \mathbf{W}_0$, and $\mathbf{D}$ until the function FLP converges or the change in $\mathbf{W}$ and $\mathbf{D}$ becomes smaller than a certain value.

Algorithm 1. Molecular Dictionary Learning

Input: $\mathbf{Y}, \mathbf{D}^{init}, \lambda, \theta, \phi$, loopNum, tol1, tol2
Output: $\mathbf{D}, \mathbf{W}, \mathbf{W}_0$
$\mathbf{W}^{(0)} = 0$; $\mathbf{W}_0^{(0)} = 0$; $\mathbf{D}^{(0)} = \mathbf{D}^{init}$;
$\mathbf{z}_0^{(0)} = \{\mathbf{y} + 1\}$;
prevLp $= -$inf; lp $=$ FLP$(\mathbf{W}^{(0)}, \mathbf{W}_0^{(0)}, \mathbf{D}^{(0)}, \lambda, \theta, \phi; \mathbf{y})$;
for $i = 1$; $i <$ loopNum; i++ **do**
  $\mathbf{W}^{(i)}, \mathbf{W}_0^{(i)} =$
  updateW $-$ ADMM $(\mathbf{D}^{(i-1)}, \mathbf{W}^{(i-1)}, \mathbf{W}_0^{(i-1)},$
  $\mathbf{z}_0^{(i-1)}, \mathbf{z}_1^{(i-1)}, \mathbf{z}_2^{(i-1)}, \lambda, \theta, \mathbf{y})$;
  $\mathbf{D}^{(i)} = \text{argmax}_{\mathbf{D}} \text{FLP}(\mathbf{W}^{(i)}, \mathbf{W}_0^{(i)}, \mathbf{D}^{(i-1)}, \phi; \mathbf{y})$;
  if $|\text{lp} - \text{prevLp}| <$ tol1 *or*
  $(max(\mathbf{W}^{(i)} - \mathbf{W}^{(i-1)}) <$ tol2 *and*
  $max(\mathbf{D}^{(i)} - \mathbf{D}^{(i-1)}) <$ tol2) **then**
    break;
  end
  prevlp $=$ lp; lp $=$ FLP$(\mathbf{W}^{(i)}, \mathbf{W}_0^{(i)}, \mathbf{D}^{(i)}, \lambda, \theta, \phi; \mathbf{y})$;
end

3 Results

We ran our method on three synthetic datasets to show its performance. **MOLDL** was then applied to two real datasets and evaluated based on its ability to recover dictionary elements corresponding to known molecular species in the sample.

3.1 Synthetic data results

We present three synthetic experiments with different purposes. Because the ground truth dictionaries of these experiments are known, we quantify performance of our and other's methods in terms of dictionary recovery. Here, we use an entry-by-entry comparison and cross-dictionary coherence as indices of the recovery. If a dictionary size is small, we can compare the ground truth dictionary and the dictionary learned by computational methods entry-by-entry. If the dictionary size is large, we use cosine similarity to compute agreement between pairs of dictionary elements. For a ground truth dictionary element $\mathbf{d}_m$, and a learned dictionary element $\tilde{\mathbf{d}}_m$, we will refer to a pair of dictionary elements with the same index as matched. For example, if $\mathbf{d}_m$ and $\tilde{\mathbf{d}}_m$ are matched, all other pairs are mismatched. The cosine similarity of the two $\tilde{\mathbf{d}}_m, \tilde{\mathbf{d}}_m$ vectors is computed as $\cos(\mathbf{d}_m, \tilde{\mathbf{d}}_m) = \frac{\mathbf{d}_m \cdot \tilde{\mathbf{d}}_m}{\|\mathbf{d}_m\| \|\tilde{\mathbf{d}}_m\|}$. Hence, given two dictionaries, $\mathbf{D}$, and $\tilde{\mathbf{D}}$ we can compute a cosine similarity matrix $c_{i,j} = \cos(\mathbf{d}_i, \tilde{\mathbf{d}}_j)$. We note that cosine similarity matrix between a non-negative dictionary and itself may not be a diagonal matrix.

A reconstructed dictionary might contain the same elements as the ground truth dictionary, but in a different order. To obtain

optimal matching between elements of dictionaries, we run the Hungarian algorithm (Kuhn, 1955) on the negative of the cosine similarity matrix. This ordering has the benefit of maximizing overall cosine similarity between matched dictionary elements.

3.2 Synthetic experiment 1: the advantage of using a dictionary pattern

A dictionary pattern can help recover the dictionary more accurately by reducing the chance of learning false positive entries in the dictionary. To show this statement is true, we compared the recovered dictionaries from our method with and without a dictionary pattern. In the first synthetic experiment, we made a ground truth dictionary that has the dictionary pattern $[1\,1\,0; 0\,1\,1; 1\,0\,0]$. One means entries have values and zero means entries do not have values. As defined, a dictionary pattern only decides the sparsity of the dictionary; therefore, the values of these non-zero entries are still undecided. To make the simulation simpler, we set the values of each entries to be the same. And because the constraint on a dictionary element is to have its L2-square value equal to one, the values of the ground truth dictionary entries were $[0.707\,0.707\,0; 0\,0.707\,0.707; 1\,0\,0]$ ($0.707^2 + 0.707^2 \approx 1$). This dictionary has three molecular species, with the first two elements have at most two possible values that are non-zeros, and the third element has at most one possible value. The values of $\mathbf{W}$ were generated by taking absolute values of the sample from a normal distribution ($\mu = 0, \sigma = 10$). The sample size (width times height) is 20×20; the $\mathbf{W}$ is a tensor of the size $3 \times 20 \times 20$.

The result in Figure 2 shows that the dictionary learning without the pattern learned one false positive value in entries 6 (the yellow bars) and two false negative values in entries 3 and 5. Also, using the pattern makes it easier for the algorithm to converge: **MOLDL** with the pattern took 11 iterations to converge, while without the pattern it took 21 iterations. Our argument is that even if the ground truth dictionary is simple and the sample size is relatively large considering the variables to be learned, it is still possible that dictionary learning recovers a dictionary with false entries. Hence using the dictionary patterns improves both speed and accuracy of the method.

3.3 Synthetic experiment 2: deconvolution of molecular species

In the second synthetic experiment, we addressed a situation where multiple molecules contribute to the same peak in the spectra and their abundances are diffused across the sample in 2D space, and we compared the results of different computational methods in this situation. In this experiment, we set a ground truth dictionary as a three-by-three matrix, with the first, second, third column being $[0.89, 0.45, 0], [0, 0.89, 0.45], [0.89, 0, 0.45]$, respectively. Thus each of the three molecular species shares peaks with the other two. A ground truth abundances were generated as follows: a location on the grid for a particular molecular species was chosen and diffusion of its abundance was performed. The diffusion was emulated using a mean filtering kernel of size 3×3 iterated 10 times. This is to simulate the real case of microbial secretion that would lead the abundances of nearby locations to be similar. The ground truth abundance is a $3 \times 20 \times 20$ tensor. The computational methods we compared our method to are NNMF and pLSA, and we set the variable numbers (molecular species number) for all methods to 3.

The result is shown in Figure 3. Our method (**MOLDL**, the light-blue bars) decomposed dictionary elements correctly. Both NNMF (the yellow bars) and pLSA (the red bars) were unable to recover dictionary elements correctly in this simulation. Note that in this experiment there is no sparsity penalty on either abundances or dictionary.

3.4 Synthetic experiment 3: dictionary recovery evaluation

In the third synthetic experiment, we simulated a larger dataset using a ground truth dictionary with 38 *m/z* values and 20 molecular species. To simulate the dictionary pattern used in the real data, we generated this dictionary pattern by extracting part of the pattern from the pattern generated in Section 2.1. We extracted the first 20 dictionary elements; there are 38 different values for their respective *m/z* values. To simulate the real case that not all locations (sample grids) contain all molecular species, we made $\mathbf{W}$ in some locations zero, so different molecular species existed in different locations of the synthetic sample. The ground truth abundance is a $20 \times 30 \times 30$ tensor. We generated ground truth $\mathbf{W}$ according to the method used in synthetic experiment 2. For ground truth $\mathbf{D}$, while we assume that non-zero pattern is known, the values of actual entries in the dictionaries need to be learned. In our experiments, we generated those by taking the absolute value of the sample from a normal

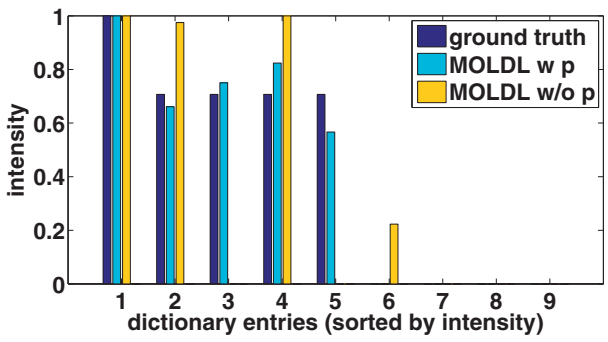

Fig. 2. The entry-by-entry comparison of the ground truth dictionary (the dark-blue bars), **MOLDL** with the pattern (**MOLDL** with p, the light-blue bars), and the one without the pattern (**MOLDL** w/o p, the yellow bars). The entries are sorted in descend order of the intensity in the ground truth dictionary, so the three entries with zero intensities in the ground truth dictionary are the last three entries in the figure

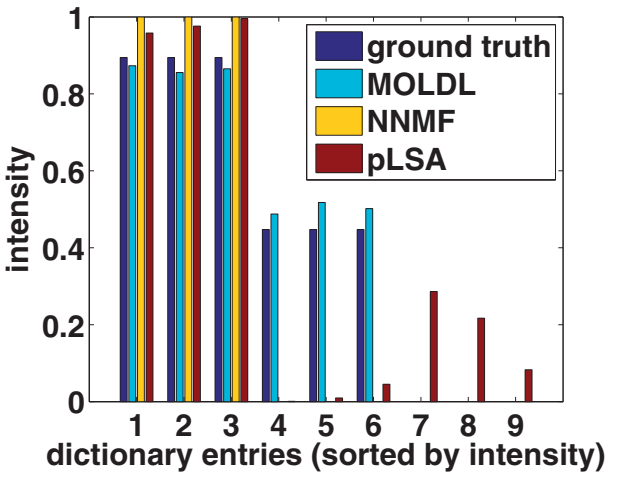

Fig. 3. The entry-by-entry comparison of the ground truth dictionary (the dark-blue bars) and the learned dictionaries from our method (**MOLDL**, the light-blue bars), NNMF (the yellow bars), and pLSA (the red bars). Ground truth dictionary consists of three elements and nine entries. The entries are sorted in descend order of the intensity in the ground truth dictionary as in Figure 2

distribution. The computational methods we compared are NNMF and pLSA. All hyperparameters in MOLDL were learned by held-out validation mentioned in the Supplementary data. **MOLDL** learned the molecular number automatically by the hyperparameter λ. We set both NNMF and pLSA's molecular number to 20. This comparison result is shown in the Supplementary data.

To make pLSA and NNMF more comparable to our algorithm, we added sparsity constraints on pLSA (sparse-pLSA) and NNMF (sparse-NNMF) according to the algorithms of Li and Ngom (2013)and Liu *et al.* (2010). We used ℓ_1 regularization on abundances in both algorithms.

In the Figure 4, we compared the learned dictionaries from each method to the ground truth dictionary. First, learned dictionary elements were matched by the Hungarian algorithm, then we computed the cosine similarity for all element pairs from the learned dictionary and the ground truth dictionary. We divided these similarities into a matched dictionary elements group and a mismatched dictionary element group and computed a histogram with bin size equal to 100 for both groups. This histogram was then normalized because the number of true dictionary elements and the number of false ones were different. We also show in Figure 4A the comparison of the ground truth dictionary to itself as a reference.

As we see in Figure 4A all true dictionary elements have a cosine similarity of 1, the maximum value. In Figure 4A, some mismatched dictionary elements have high cosine similarity because these elements are very similar. These give rise to the bars with small heights in Figure 4A because the abundances of them are few. In Figure 4B and C, the comparisons of the results of sparse-NNMF and sparse-pLSA are shown. For sparse-NNMF and sparse-pLSA ∼20% of the matched dictionary elements' cosine similarities lie in the bin of 1, and the cosine similarities, as a whole, lie in a broad range, from 0.2 to 1. However, there are also ∼10% of the matched dictionary elements' cosine similarities that lie in the bin of 0.4 in both algorithms and 10% of the matched dictionary elements' cosine similarities that lie in the bin of 0.2 in sparse-NNMF. In contrast, **MOLDL** performed better: there are ∼95% of the matched dictionary elements' cosine similarities larger then 0.5 and 80% of their cosine similarities larger then 0.8 (shown in Fig. 4D). One matched dictionary elements' cosine similarity is low (0.3) because this element is similar to other elements in this synthetic experiment. So the contributions of this element to the signals are learned as the contributions come from the other elements. In **MOLDL**, there are 5% of the matched dictionary elements that have their cosine similarities below 0.5; while using sparse-NNMF and sparse-pLSA methods there are 47%.

In terms of the mismatched dictionary elements, their distribution in **MOLDL** is similar to that of the ground truth dictionary (Fig. 4A). In contrast, there are some mismatched dictionary elements with large cosine similarities in sparse-NNMF and sparse-pLSA. This synthetic experiment shows the performance of sparse-NNMF and sparse-pLSA is more vulnerable to noise while **MOLDL** offers more robust performance.

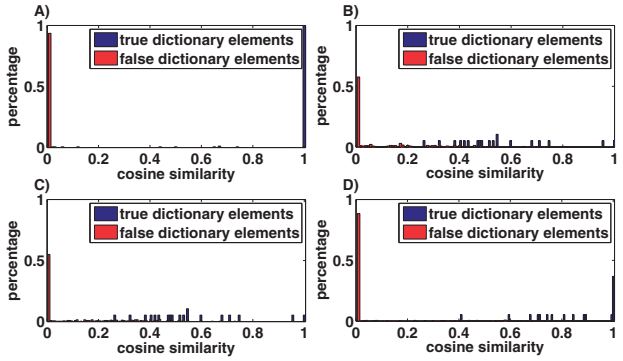

Fig. 4. The cosine distribution of the true dictionary elements and false dictionary elements for each pair of methods comparison. (**A**) ground truth dictionary compared to itself. (**B**) The dictionary came from sparse-pLSA compared with the ground truth dictionary. (**C**) The dictionary came from sparse-NNMF compared with ground truth dictionary. (**D**) The dictionary came from **MOLDL** compared with the ground truth dictionary

3.5 Real data results

Strains and Media Preparation. *Bacillus cereus* ATCC14579 and *Bacillus subtilis* NCIB 3610 were resuspended into Luria Broth from growth on agar plates, and resuspended to an OD_{600} of 0.5 One μl of these cell suspensions were then spotted onto 10 ml agar plates (0.1X Luria Broth, Lennox: 10 g tryptone, 5 g yeast extract, 5 g NaCl and 15 g Bacto-agar per L). Four bacterial spots (two of each bacterial species) were put onto the agar plates in a line, with the two spots of the same species next to each other at a 1 cm distance, and the spots of the different bacterial species 0.5 cm away from each other. Colonies were grown at 30°C for 12 or 40 hr before being harvested for MALDI-TOF imaging.

MALDI-IMS Sample Preparation. Agar-grown microbial samples were prepared for MALDI as described in Yang *et al.* (2012). Briefly, rectangular regions of agar containing the bacterial co-cultures and the distal control colonies were excised from the agar plate, placed onto a MALDI-TOF ground steel target plate (Bruker part no. 224990) and covered with Universal MALDI matrix (Sigma, Fluka 50149) using a 53 μm stainless steel sieve (Hogentogler & Co, part 1312). After matrix application, the sample was dried overnight at 37°C. Excess matrix was physically removed to clean the plate, and a peptide calibration standard was spotted onto it (Bruker part no. 206195, Pepmix4).

MALDI-IMS Experiment Protocol. After mass calibration using the Pepmix standard, samples were imaged using a MALDI-TOF mass spectrometer (Microflex LRF, Bruker) with a Microscout ion source (Nitrogen UV laser, $\lambda = 337$ nm) in both linear positive and linear negative mode. FlexControl and FlexImaging software (Bruker) was used for image acquisition with 80 shots averaged from each pixel of 400 to 800 μm across across an m/z of 0 to 5000 Da.

Preprocessing of the Raw Data. The raw data was preprocessed by first converting the Bruker file into the mzML format with msconvert (Chambers *et al.*, 2012). We emphasize here that there was no preprocessing done by manufacturer's software in this step. Then the mzML format was processed with the R package for MALDI-MS data processing (Gibb and Strimmer, 2012), and the peak picking was done with another R package (Du *et al.*, 2006). The data here is still count data because the peak picking algorithm is choosing a subset of m/z channels and there are no values changes in the input data. For a grid cell without a particular peak, we extracted the value from the original spectrum data, so there is no zero-inflation. In order to cope with the measurement error, we binned the peaks and chose the maximum value in a bin to represent that bin. This resulted in a tensor (data cube) of size $496 \times 38 \times 59$ for positive mode data, and of size $88 \times 38 \times 60$ for negative mode data. Further details on the preprocessing are given in the Supplementary data.

MOLDL on the Real Data. We constructed the dictionary patterns in both modes according to Section 2.1, except we chose a subset of the elements that had supports in the preprocessed data. We

also removed elements that have unreasonable composition in terms of ion types, e.g. an element with just one ion form [M + 2Na-H] was removed. This results in a dictionary pattern of size 88×108 in the negative mode and of size 496×753 in the positive mode. We initialized the dictionary with NNMF and initialized $\mathbf{W}, \mathbf{W}_0$ to zero. Previous research on the microbe *B. subtilis* (Hoefler *et al.*, 2012; Watrous *et al.*, 2012) showed that it produces the molecule surfactin, and that surfactin is visible in MALDI-IMS data. Surfactin is a bacterial cyclic lipopeptide whose molecular weight is 1008.3 (corresponding to C13 type, Hoefler *et al.*, 2012). Surfactin molecules include a hydrophobic acyl chain of varied length with C13, C14, C15 forms predominant (Atsushi *et al.*, 1969; Bonmatin *et al.*, 1994). Hence, we expect to see different forms of surfactin in our samples as previous studies have shown (Hoefler *et al.*, 2012; Watrous *et al.*, 2012). To prove that we could captured all the peaks associated with a particular form of surfactin in a single dictionary element, experiments with purified surfactin were done using both positive and negative mode.

Note that the purified surfactin consists of different isoforms of surfactins. Though these forms are all called surfactin due to their similar chemical structures, they are considered different in our modeling because different bacterial strains may secrete different composition of isoforms. As our goal is to decompose the secreted compounds under different bacterial strains' interactions, different isoforms of surfactin should be modeled as different compounds. To distinguish between different forms of surfactin, we called surfactin with C13, C14 and C15 acyl chain surfactin-C13, surfactin-C14 and surfactin-C15.

To make a comparison between the computational methods on real datasets (in both positive and negative ionization modes), we applied the same preprocessing steps to the surfactin MALDI data as we used on MALDI-IMS data. We identified each *m/z* signal in surfactin-C13, surfactin-C14 and surfactin-C15 isoforms for each of the different ion types shown in Table 1. We then normalized the *m/z* signals in the spectra as we normalized the dictionary elements in our method. We compared the purified surfactin dictionary signatures and their counterparts from **MOLDL**, sparse-pLSA and sparse-NNMF. To find the matched elements in sparse-NNMF and sparse-pLSA, we used the Hungarian algorithm mentioned before to find the element that has the largest cosine similarity to the surfactin dictionary element. In **MOLDL**, since every element's construction is based on the ion types, one can deduce the underlying molecular weight giving rise to these ion types. For the matched elements, we first removed the entries with intensities lower then 1% as we deemed them as noise. Then we kept the five largest-intensity entries in the elements. We aggregated all these entries of the elements for different isoforms of surfactin and the different computational methods. Surfactin-C14 and surfactin-C15 are shown in Figures 5 and 6, respectively. Among the surfactin forms, these molecules have the most complex ion types. Hence, these molecules provide best illustration of the power of our method.

In both experiments, compared with the other methods, the output of **MOLDL** had the largest cosine similarity with the purified surfactin spectra. In the negative mode, MOLDL learned a dictionary element consisting of only two entries, thus capturing all the peaks of surfactin-C14 (true positive) without introducing any false positives. Moreover, compared with the other computational methods, the relative intensities of the true positive entries from **MOLDL** has the smallest ℓ_1 distance from those from purified surfactin. On the other hand, sparse-NNMF and sparse-pLSA captured the signal of *m/z* 1021.3, but with lower relative intensity. Moreover, they were almost unable to capture the signal of *m/z* 1057.5. Therefore,

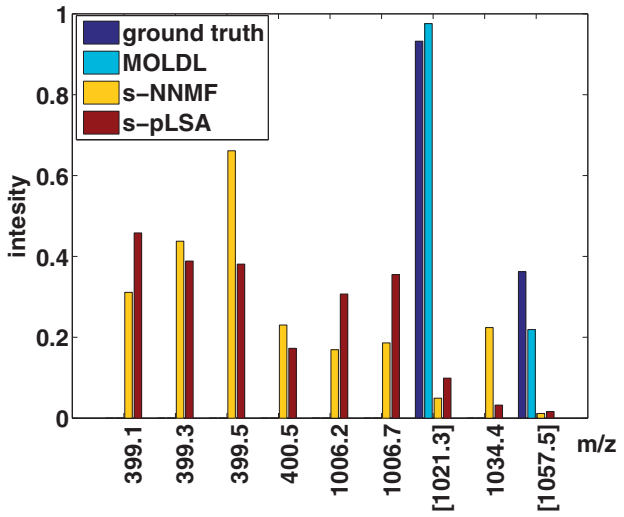

Fig. 5. The surfactin-C14 dictionary entries from ground truth (purified surfactin), **MOLDL**, s-pLSA (sparse-pLSA) and s-NNMF (sparse-NNMF) in the negative mode data. The bars from left to right for each *m/z* are the intensities of that *m/z* in each dictionary element. The *m/z* values with brackets ([1021.3]: [M-H], [1057.5]: [M + Cl]) are the *m/z* values from purified surfactin. They are true positive entries for the surfactin dictionary element. The *m/z* values without brackets are false positive entries

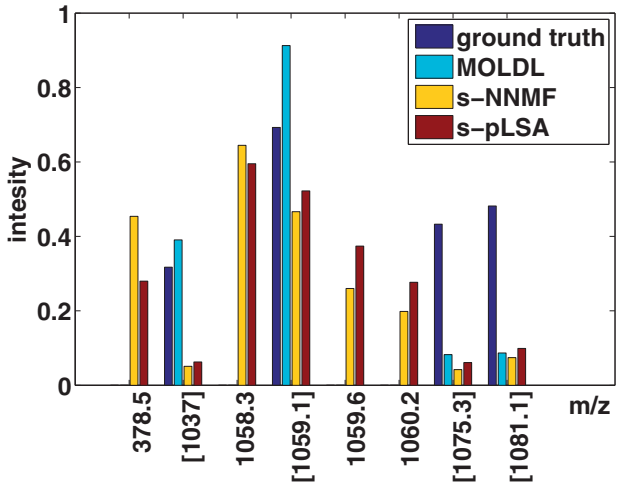

Fig. 6. The surfactin-C15 dictionary entries from ground truth (purified surfactin), **MOLDL**, s-pLSA (sparse-pLSA) and s-NNMF (sparse-NNMF) in the positive mode data. The bars from left to right for each *m/z* are the intensities of that *m/z* in each dictionary element. The *m/z* values with brackets ([1037]: [M + H], [1059.1]: [M + Na], [1075.3]: [M + K], [1081.1]: [M + 2Na-H]) are the *m/z* values from purified surfactin. They are true positive entries for the surfactin dictionary element. The *m/z* values without brackets are false positive entries

in terms of recovery of the molecular signatures in this experiments, MOLDL performed better than other methods both quantitatively and qualitatively. Both sparse-NNMF and sparse-pLSA, although they were made more robust through the use of the ℓ_1 penalty, are still very sensitive to noise. The entries they learned are more likely to be spurious because their *m/z* differences do not correspond to well-known ion types. We emphasize here that by using the dictionary patterns composed of possible ion types and sparsitiy regularization, **MOLDL** not only reduces the false positive significantly compared with other methods, but also that the signals that might

appear as false positives may indeed be real ions with a biochemical hypothesis behind them.

In the positive mode, **MOLDL**, sparse-NNMF and sparse-pLSA all learned four true positives. But sparse-NNMF and sparse-pLSA also learned four false positives. In terms of false positives, **MOLDL** performed better than other methods. Also the relative intensities of the true positive entries from **MOLDL** has the smallest ℓ_1 distance from those from purified surfaction compared to the other computational methods. The results for other surfactin molecules are shown in the Supplementary data. In these cases, **MOLDL** performed better then sparse-NNMF and sparse-pLSA by having less false positive signal, and by recovering more accurate relative intensities of the dictionary element entries.

Note that the relative intensities of some entries in the dictionary element are less accurate. One reason might be due to poor initialization such that **MOLDL** took more iterations to converge. The second reason might be we did not consider isotopic surfactin molecules: they are one or two molecular weight difference and always appear together in the same sample (Pathak and Keharia, 2014). When considering different forms of surfactins, the ion type of an isotopic surfactin variant might have similar molecular weight of the ion type of another form of surfactin. For example, if a isotopic surfactin-C14 has the molecular weight 1024, it has a signal in 1047.29 *m/z* ([M + Na]) that is within the error range of the *m/z* value of ion type [M + K] of surfactin-C13. As we did not group the signal of isotopic surfactin-C14 into a dictionary element, it is possible that it was included in another element that made the deconvolution of signals inaccurate. The third reason might be the modeling of MALDI-IMS data is not accurate, due to distributional assumptions such as Gaussian or Poisson noise. One possible distribution choice that might lead to better results is negative-binomial. Compared with pLSA and NNMF, the framework of dictionary learning makes it easier to incorporate distributions other then the Gaussian distribution into the modeling of MALDI-IMS data, allowing this to be tested easily in future iterations of the method.

In the results section, we compared all the state-of-the-art computational methods of deconvolving MALDI-IMS data on synthetic examples and real datasets. **MOLDL** performed better then other methods both in true positive and true negative rate. It can learn the most complete molecular signature of a molecular species. Also, based on the dictionary pattern used, **MOLDL** can learn the signatures corresponding to molecules based on biochemical knowledge, which other methods cannot. Third, with the framework of dictionary learning, **MOLDL** is extensible to other distributions. Our future work will be based on this framework to improve learning the relative intensities in the dictionary elements more accurately.

4 Conclusion

In this article, we present **MOLDL**, a Dictionary Learning method, aimed at summarizing MALDI-IMS data. This method deconvolves *m/z* signals into components that belong to different molecular species. Our method models the data as Poisson distributed, incorporates the possible ion types of MALDI mass spectrometry, leverages the spatial dependency of IMS data, and can learn the number of molecular species present in the MALDI-IMS data automatically. We implemented this method and develop a straightforward method for choosing its hyperparameters. We tested it on three synthetic and two real experimental datasets, and showed that, compared with prior approaches, our method provides superior recovery of molecular signatures.

Funding

This research is supported by Lineberger Comprehensive Cancer Center and UNC University Cancer Research Fund.

Conflict of Interest: none declared.

References

Alexandrov,T. (2012) Maldi imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics*, **13** (Suppl. 16), S11.

Alexandrov,T. *et al.* (2013a) Analysis and interpretation of imaging mass spectrometry data by clustering mass-to-charge images according to their spatial similarity. *Anal. Chem.*, **85**, 11189–11195.

Alexandrov,T. *et al.* (2013b) Maldi-imaging segmentation is a powerful tool for spatial functional proteomic analysis of human larynx carcinoma. *J. Cancer Res. Clin. Oncol.*, **139**, 85–95.

Atsushi,K. *et al.* (1969) Determination of fatty acid in surfactin and elucidation of the total structure of surfactin. *Agric. Biol. Chem.*, **33**, 973–976.

Balasubramanian,K. *et al.* (2012) Smooth sparse coding via marginal regression for learning sparse representations. *CoRR,* **abs/1210.1121**.

Bonmatin,J. *et al.* (1994) Solution three-dimensional structure of surfactin: a cyclic lipopeptide studied by 1H-NMR, distance geometry, and molecular dynamics. *Biopolymers*, **34**, 975–986.

Cazares,L.H. *et al.* (2009) Imaging mass spectrometry of a specific fragment of mitogen-activated protein kinase/extracellular signal-regulated kinase kinase kinase 2 discriminates cancer from uninvolved prostate tissue. *Clin. Cancer Res.*, **15**, 5541–5551.

Chambers,M. *et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*, **30**, 918–920.

Du,P. *et al.* (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, **22**, 2059–2065.

Gibb,S. and Strimmer,K. (2012) MALDI-quant: a versatile r package for the analysis of mass spectrometry data. *Bioinformatics*, **28**, 2270–2271.

Groseclose,M.R. *et al.* (2008) High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using MALDI imaging mass spectrometry. *Proteomics*, **8**, 3715–3724.

Gross,J.H. (2011) *Mass Spectrometry*, 2nd edn. Springer, Berlin.

Hanselmann,M. *et al.* (2008) Concise representation of mass spectrometry images by probabilistic latent semantic analysis. *Anal. Chem.*, **80**, 9649–9658.

Hoefler,B. *et al.* (2012) Enzymatic resistance to the lipopeptide surfactin as identified through imaging mass spectrometry of bacterial competition. *Proc. Natl Acad. Sci. U.S.A.*, **109**, 13082–13087.

Hong,D. and Zhang,F. (2010) Weighted elastic net model for mass spectrometry imaging processing. *Math. Model. Nat. Phenom.*, **5**, 115–133.

Jones,E. *et al.* (2012) Imaging mass spectrometry statistical analysis. *J. Proteomics*, **75**, 4962–4989.

Klerk,L. *et al.* (2007) Extended data analysis strategies for high resolution imaging ms: New methods to deal with extremely large image hyperspectral datasets. *Int. J. Mass Spectrom.*, **260**, 222–236.

Kobarg,J.H. *et al.* (2013) Numerical experiments with MALDI imaging data. *Adv. Comput. Math.*, **40**, 667–682.

Kuhn,H.W. (1955) The Hungarian method for the assignment problem. *Naval Res. Logist. Q.*, **2**, 83–97.

Lee,H. *et al.* (2007) Efficient sparse coding algorithms. In: Schölkopf,B. *et al* (eds), *Advances in Neural Information Processing Systems*, MIT Press, **Vol. 19**, pp. 801–808.

Lemaire,R. *et al.* (2007) Specific MALDI imaging and profiling for biomarker hunting and validation: fragment of the 11s proteasome activator complex,

reg alpha fragment, is a new potential ovary cancer biomarker. *J. Proteome Res.*, **6**, 4127–4134.

Li,Y. and Ngom,A. (2013) The non-negative matrix factorization toolbox for biological data mining. *Source Code Biol. Med.*, **8**, 10.

Liu,S. *et al.* (2010) Efficient probabilistic latent semantic analysis with sparsity control. In: *Data Mining (ICDM), 2010 IEEE 10th International Conference*, pp. 905–910.

Maurer,A. *et al.* (2013) Sparse coding for multitask and transfer learning. In: Dasgupta, S. and Mcallester, D. (eds.) *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. JMLR Workshop and Conference Proceedings, pp. 343–351.

Mehta,N. and Gray,A.G. (2013) Sparsity-based generalization bounds for predictive sparse coding. In: Dasgupta, S. and Mcallester, D. (eds.) *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. JMLR Workshop and Conference Proceedings, Vol. 28, pp. 36–44.

Olshausen,B.A. and Field,D.J. (1997) Sparse coding with an overcomplete basis set: a strategy employed by v1?. *Vision Res.*, **37**, 3311–3325.

Pathak,K. and Keharia,H. (2014) Identification of surfactins and iturins produced by potent fungal antagonist, bacillus subtilis k1 isolated from aerial roots of banyan (ficus benghalensis) tree using mass spectrometry. *3 Biotech*, **4**, 283–295.

Rauser,S. *et al.* (2010) Classification of her2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. *J. Proteome Res.*, **9**, 1854–1863.

Siy,P. *et al.* (2008) Matrix factorization techniques for analysis of imaging mass spectrometry data. In: *BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference*, pp. 1–6.

Sören-Oliver, D. and Klaus Meyer,A.W. (2015) Application Note # mt-111: Concise Interpretation of MALDI Imaging Data by Probabilistic Latent Semantic Analysis (plsa). http://www.bruker.com/fileadmin/user_upload/8-PDF-Docs/Separations_MassSpectrometry/Literature/literature/ApplicationNotes/MT-111_pLSA_ebook.pdf (10 March 2015, date last accessed).

Trede,D. *et al.* (2012) On the importance of mathematical methods for analysis of MALDI-imaging mass spectrometry data. *J. Integr. Bioinform.*, **9**, 189.

Van de Plas,R. *et al.* (2007) Imaging mass spectrometry based exploration of biochemical tissue composition using peak intensity weighted PCA. In: *Life Science Systems and Applications Workshop, LISA* 2007. IEEE/NIH, pp. 209–212.

Watrous,J. *et al.* (2012) Mass spectral molecular networking of living microbial colonies. *Proc. Natl Acad. Sci.*, **109**, E1743–E1752.

Yang,J. *et al.* (2012) Primer on agar-based microbial imaging mass spectrometry. *J. Bacteriol.*, **194**, 6023–6028.