Classification and Cluster Analysis of Complex Time-of-Flight Secondary Ion Mass Spectrometry for Biological Samples

Xue Tian[a], Stephen E. Reichenbach[a], Qingping Tao[b], Alex Henderson[c]
xtian@cse.unl.edu, reich@cse.unl.edu, qtao@gcimage.com, Alex.Henderson@manchester.ac.uk
[a]*Computer Science and Engineering Department, University of Nebraska-Lincoln, Lincoln, NE, USA*
[b]*GC Image LLC, Lincoln, NE, USA*
[c]*Surface Analysis Research Centre, School of Chemical Engineering and Analytical Science, University of Manchester, Manchester, UK*

Abstract

Identifying and separating subtly different biological samples is one of the most critical tasks in biological analysis. Time-of-flight secondary ion mass spectrometry (ToF-SIMS) is becoming a popular and important technique in the analysis of biological samples, because it can detect molecular information and characterize chemical composition. ToF-SIMS spectra of biological samples are enormously complex with large mass ranges and many peaks. As a result the classification and cluster analysis are challenging. This study presents a new classification algorithm, the most similar neighbor with a probability-based spectrum similarity measure (MSN-PSSM), which uses all the information in the entire ToF-SIMS spectra. MSN-PSSM is applied to automatically classify bacterial samples which are major causal agents of urinary tract infections. Experimental results show that MSN-PSSM is an accurate classification algorithm. It outperforms traditional techniques such as decision trees, principal component analysis (PCA) with discriminant function analysis (DFA), and soft independent modeling of class analogy (SIMCA). This study also applies a modern clustering algorithm, normalized spectral clustering, to automatically cluster the bacterial samples at the species level. Experimental results demonstrate that normalized spectral clustering is able to show accurate quantitative separations. It outperforms traditional techniques such as hierarchical clustering analysis, k-means, and PCA with k-means.

1. Introduction

Classification and cluster analysis are widely used techniques for exploring data. Identification of similar functional groups provides first-stage guidance for data analysis. Classification and cluster analysis of biological samples are difficult because biological samples are complex and similar to one another. ToF-SIMS can detect molecular information and characterize the chemical composition of biological samples. Hence, ToF-SIMS is becoming a popular and important technique in the analysis of biological samples [1] [2] [3].

ToF-SIMS uses a pulsed primary ion beam (e.g. Au_3^+) to remove molecules and fragment ions from the outermost surface of the sample as Figure 1 illustrates. The primary ion beam is carefully controlled to a sufficiently low intensity to ensure that the surface molecules are not completely broken into individual atoms. The fragment ions removed from the surface (secondary ions) are transferred into a "flight tube" and the mass/charge is determined by measuring the time (after the primary pulse) at which they reach the detector (time-of-flight) [4].

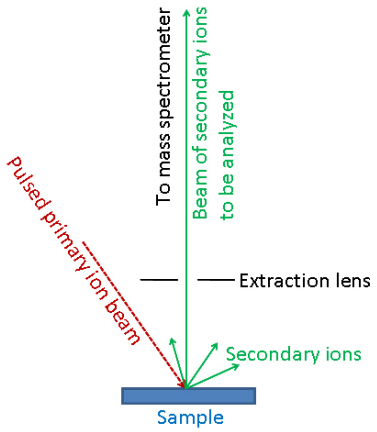


Figure 1. ToF-SIMS scheme

ToF-SIMS spectra of biological samples are enormously complex with large mass ranges (m/z < 5000) and many structurally significant peaks combined with noise peaks (such as contaminants and small or non-diagnostic fragment ions). This complex data contains information about sample composition, molecular orientation, surface order, chemical bonding, etc. The challenge is how to extract useful information from complex ToF-SIMS spectra for classification and cluster analysis. Moreover, the size of a large TOF-SIMS dataset can increase this challenge.

Multivariate analysis (MVA) such as PCA has become the most popular technique in processing of ToF-SIMS data [5] [6]. PCA reduces the dimensionality of ToF-SIMS data, and DFA has been used to identify and discriminate individual strains of bacteria [7] [8] and also prostate cancer cells [9]. Berman *et al*. [10] used SIMCA [11] to classify and characterize sugars, proteins, and mouse embryos. Although MVA is the most popular technique to process ToF-SIMS data, it is widely recognized that the effectiveness of MVA is dependent on appropriate data pretreatment, such as the selection of peaks from the spectra, scaling, centering, and non-linear transformations, and no rules have been established for data pretreatment before MVA.

The decision trees algorithm [12] is one of the most popular classification algorithms in data mining and machine learning. It is a successful algorithm for the description, classification, and generalization of data in many diverse real-world applications [13]. Engrand *et al*. [14] successfully used decision trees to classify ToF-SIMS data of mineral samples. However, there are few applications of decision trees in ToF-SIMS communities.

Hierarchical clustering analysis (HCA) and k-means are the most popular and well-known clustering algorithms in machine learning and pattern recognition [15]. Although HCA and k-means are new techniques in ToF-SIMS communities, Thompson *et al*. [8] and Suzuki *et al*. [16] successfully used HCA to quantitatively determine the degree of similarity and dissimilarity among the TOF-SIMS spectra.

This study presents MSN-PSSM, a new classification algorithm. MSN-PSSM is applied to automatically classify bacterial samples which are major causal agents of urinary tract infection (UTI). UTI is a serious health problem affecting millions of people each year [17]. There is a growing need to identify the causal agent prior to treatment. MSN-PSSM successfully classifies the bacterial samples at the strain level. This study also applies a modern clustering algorithm, normalized spectral clustering, to automatically cluster the bacterial samples at the species level.

2. Bacterial datasets

UTI bacterial species include *Escherichia coli*, *Klebsiella oxytoca*, *Klebsiella pneumoniae*, *Proteus mirabilis*, *Enterococcus spp*, and *Citrobacter freundii*. This study examines 19 strains of UTI bacteria belonging to these six species. The 19 strains are five strains of *Escherichia coli* (Eco), one strain of *Klebsiella oxytoca* (Kox), three strains of *Klebsiella pneumoniae* (Kpn), two strains of *Citrobacter freundii* (Cfr), four strains of *Enterococcus spp* (Esp), and four strains of *Proteus mirabilis* (Pmi). These strains were previously identified by conventional biochemical tests. Each strain has three biological replicates.

Bacterial sample growth, ToF-SIMS instrumentation, and data acquisition parameters are described in detail by Fletcher [7]. Each ToF-SIMS spectrum is from 1 Da to 1000 Da, and binned to 1 Da mass intervals from -0.5 to +0.5 of each nominal mass. Figure 2 (a) shows the ToF-SIMS spectrum of an *Escherichia coli* sample from 1 Da to 1000 Da. Figure 2 (b) shows the spectrum from 1 Da to 200 Da. Figure 2 (c) shows the spectrum from 200 Da to 1000 Da. Figure 3 (a), 3 (b), and 3 (c) show the ToF-SIMS spectrum of a *Proteus mirabilis* sample. These spectra have many peaks with varying intensities over mass range 1 to 1000. Many common peaks make visual inspection and manual identification of spectra an impossible task. Hence, it is necessary to develop automatic techniques to classify and cluster these complex ToF-SIMS spectra. Three ToF-SIMS spectra are generated from three fresh areas of each bacterial sample to make three machine replicates for each biological replicate. So, in total there are nine ToF-SIMS spectra for each of 19 strains of UTI bacteria belonging to six species.

3. Classification and cluster analyses

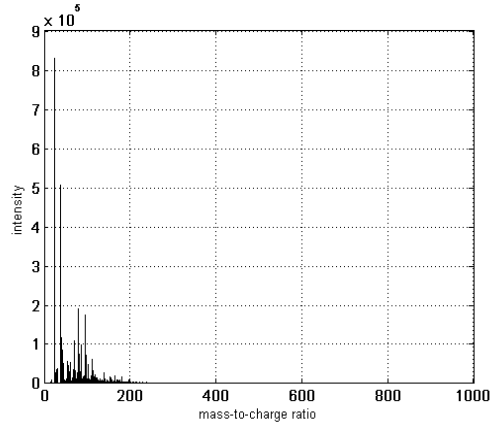
3.1. Pre-processing

Figure 2 and Figure 3 show that the spectra are dominated by Na^+ (m/z=23) and K^+ (m/z=39) ions. Because this salt contamination is apparent and peaks in the low mass region have little discrimination ability, m/z from 1 to 50 are pruned from the ToF-SIMS spectra. Each ToF-SIMS spectrum is then normalized to the most intense peak (the base peak) in the spectrum.

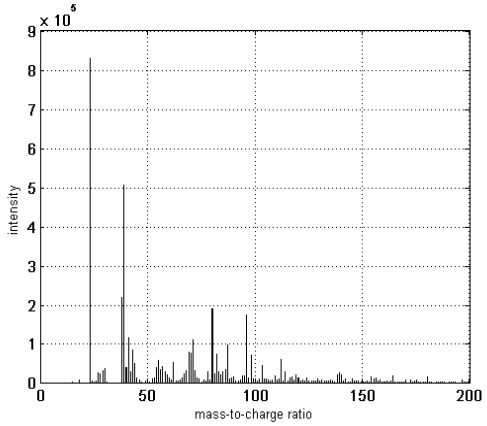
3.2. MSN-PSSM algorithm

Given a query spectrum x_0, p predefined class labels $\{c_1, c_2, ..., c_p\}$, and a set of n labeled spectra $\{x_i, y_i\}$, where i=1, 2, ..., n, and $y_i \in \{c_1, c_2, ..., c_p\}$ is the known class label of each spectrum, the task is to predict the class label of x_0. MSN-PSSM algorithm searches for the most similar spectrum in $\{x_i, y_i\}$; i.e. the spectrum that has the highest probability-based spectrum similarity to x_0, and then predicts the class label of x_0 as the most similar spectrum's class label.

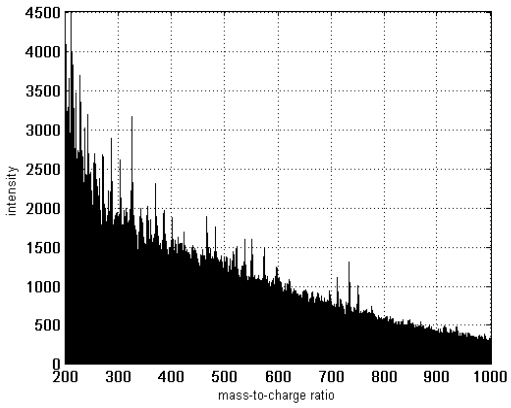
Building upon the recent work of Visvanathan *et al*. [18], who present a new information-theoretic library search technique for comprehensive two-dimensional gas chromatography with mass spectrometry, this study presents a new probability-based spectrum similarity (PSS) measure considering intra-class variability of ToF-SIMS spectra. The PSS between a query spectrum x_0 and



(a) mass-to-charge ratio from 1 to 1000

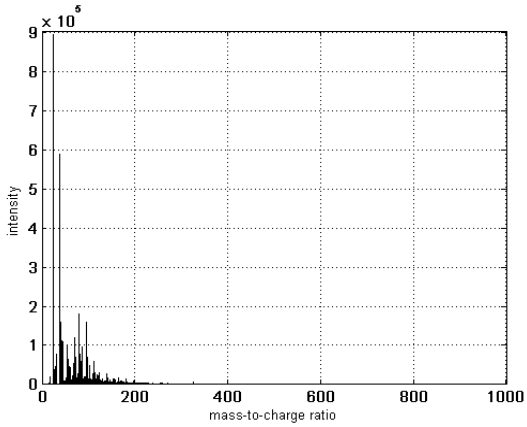


(b) mass-to-charge ratio from 1 to 200

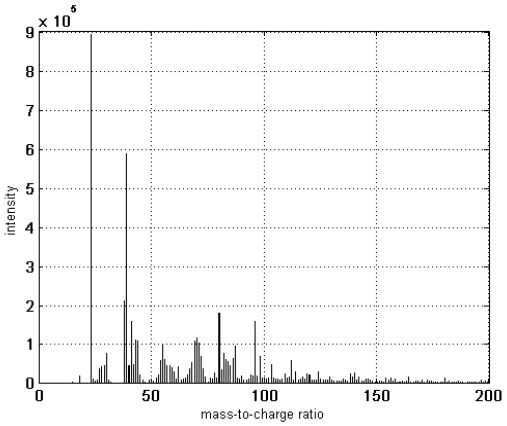


(c) mass-to-charge ratio from 200 to 1000

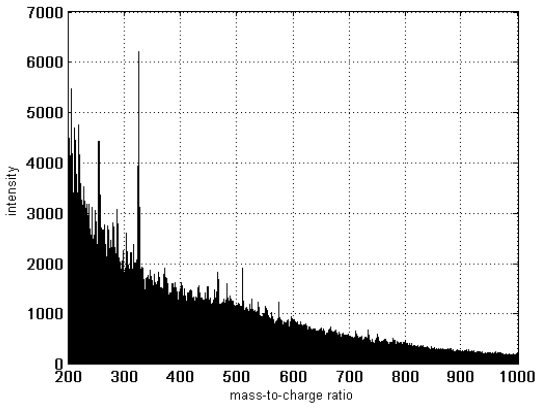
Figure 2. ToF-SIMS spectrum of *Escherichia coli*



(a) mass-to-charge ratio from 1 to 1000



(b) mass-to-charge ratio from 1 to 200



(c) mass-to-charge ratio from 200 to 1000

Figure 3. ToF-SIMS spectrum of *Proteus mirabilis*

one of the labeled spectra $\{x_i, y_i\}$ is:

$$\mathrm{PSS} = \sum_{m=m_{\min}}^{m_{\max}} \log\left(\frac{N_{\varepsilon,m,i}[a_0(m) - a_i(m)]}{(P_m * N_{\varepsilon,m,i})[a_0(m)]}\right). \qquad (1)$$

In equation (1), m is the mass-to-charge ratio (m/z), m_{min} is the minimum mass-to-charge ratio of the query spectrum and the labeled spectra, m_{max} is the maximum mass-to-charge ratio of the query spectrum and the labeled spectra, $a_0(m)$ is the intensity of spectrum x_0 at mass-to-charge ratio m, $a_i(m)$ is the intensity of spectrum x_i at mass-to-charge ratio m, ε represents the intra-class variability parameters, $N_{\varepsilon,m,i}$ is the intra-class variability model for x_i's class, P_m is the intensity probability distribution of all labeled spectra at mass-to-charge ratio

m, and $P_m * N_{\varepsilon,m,i}$ represents convolution of P_m and $N_{\varepsilon,m,i}$. This similarity measure uses all information in the entire ToF-SIMS spectra without any dimensionality reduction of the data.

$N_{\varepsilon,m,i}$ is a Gaussian distribution described by mean (μ_m) and standard deviation (σ_m):

$$N_{\varepsilon,m,i} = \frac{1}{\sigma_m \sqrt{2\pi}} e^{\frac{-(x-\mu_m)^2}{2\sigma_m^2}}. \quad (2)$$

σ_m is estimated by $b_1 \cdot a_{avg}(m)+b_2$, where $a_{avg}(m)$ is the intensity of the average spectrum of x_i's class at mass-to-charge ratio m, and b_1 and b_2 are linear regression parameters. Figure 4 shows the standard deviations at each mass-to-charge ratio versus intensities of the average spectrum at each mass-to-charge ratio for *Citrobacter freundii*. It shows that standard deviations are roughly proportional to intensities. The intra-class variability at a certain mass-to-charge ratio is intensity dependent. Stars in Figure 5 show the intensity distribution of *Citrobacter freundii* at mass-to-charge ratio 351. The curve in Figure 5 shows the Gaussian distribution with the mean and the standard deviation of *Citrobacter freundii*. Figure 5 shows that the Gaussian distribution models the intra-class variability. The intensity difference at mass-to-charge ratio m between two spectra is used as an offset in the Gaussian distribution to measure the similarity between two spectra at mass-to-charge ratio m. When two intensities are the same, the probability that the two spectra being in the same class is the largest (and is upper-bounded by equation (2) with $x=\mu_m$).

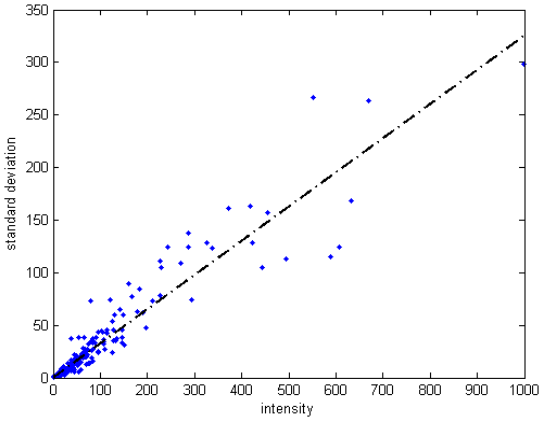


Figure 4. Standard deviations versus intensities plotting of *Citrobacter freundii*

P_m is the intensity probability distribution of all labeled spectra (at mass-to-charge ratio m):

$$P_m[a] = \frac{\#\text{ of spectra with intensity } a}{\#\text{ of labeled spectra}}. \quad (3)$$

$P_m * N_{\varepsilon,m,i}$ is P_m convolved with $N_{\varepsilon,m,i}$. P_m is blurred by the intra-class variability (at mass-to-charge ratio m) of x_i's class. The probability of two spectra with a certain intensity (at mass-to-charge ratio m) being the same or similar is small, if that intensity (at mass-to-charge ratio m) occurs frequently in all labeled spectra. As the intensity probability increases, the similarity between two spectra decreases.

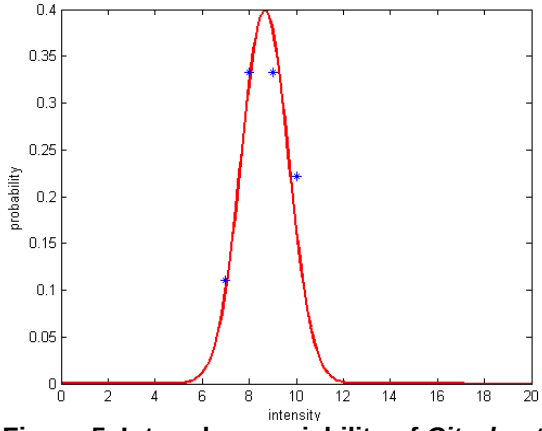


Figure 5. Intra-class variability of *Citrobacter freundii* at mass-to-charge ratio 351

3.3. Normalized spectral clustering algorithm

Cluster analysis gives first-stage guidance for exploratory data analysis before establishing models for data. Spectral clustering is a modern clustering algorithm which uses eigenvectors of a similarity matrix derived from the data. It does not require any models for data. It does not make strong assumptions on the form of clusters. Unlike k-means, for which the resulting clusters form convex sets, spectral clustering can solve general problems such as intertwined spirals.

Spectral clustering has been described in detail by Luxburg [19]. It has many applications in machine learning, exploratory data analysis, computer vision and speech processing [20]. Despite many empirical successes of spectral clustering, it is a new technique to the chemometrics and ToF-SIMS communities.

In this study, normalized spectral clustering [21] is applied to automatically cluster the bacterial samples at the species level. In brief, given a set of n unlabeled spectra $\{x_i\}$, where i=1, 2, ..., n, build a weighted fully connected graph $G=(V, E)$ to represent the spectra such that vertices are spectra $V=\{x_1, x_2, \ldots, x_n\}$, and weights of edges $W=(w_{ij})_{i,j=1,2,\ldots,n}$ are the similarity between spectra. The task is to find a partition of k clusters such that edges within a cluster have high weights and edges between different clusters have low weights.

In this study, similarity between spectra x_i and x_j is determined by the Gaussian similarity function:

$$S(x_i, x_j) = e^{\frac{-\| x_i - x_j \|^2}{2\sigma^2}}, \quad (4)$$

where σ is a parameter which controls the scale of the similarity. To find the partition, first, compute the normalized graph Laplacian:

$$L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}, \quad (5)$$

where D is a diagonal matrix with:

$$d_i = \sum_{j=1}^{n} w_{ij} \quad (6)$$

on the diagonal. Second, compute the eigenvectors u_1, u_2, …, u_k corresponding to the k smallest eigenvalues of L. Third, construct matrix U which takes eigenvectors u_1, u_2, …, u_k as columns. Then, normalize each row of U to have norm 1. Finally, take each row of U as a data point z_i and use k-means algorithm to cluster them into k clusters.

The most important strategy of this algorithm is to change the representation of the original spectra x_i to z_i. This change enhances the cluster properties in the spectra.

4. Experimental results

4.1. Classification analyses

The 19 stains of UTI bacteria are examined as 19 classes. The class labels are Cfr1, Cfr2, Eco1, Eco2, Eco3, Eco4, Eco5, Esp1, Esp2, Esp3, Esp4, Kox, Kpn1, Kpn2, Kpn3, Pmi1, Pmi2, Pmi3, and Pmi4. Each class has nine samples, and together there are 171 samples. This is a challenging multi-class classification task to demonstrate strain-level discrimination of the subtly different bacterial samples.

These 171 samples are classified by four classification algorithms: decision trees, SIMCA, PCA with DFA, and MSN-PSSM. Leave-one-out cross-validation, which is commonly used in chemometrics, is adopted in this study. Overall classification accuracy and Fleiss's kappa statistic [22] are used to quantitatively measure the performance of the different algorithms. Overall classification accuracy is defined as:

$$\text{Accuracy} = \frac{\#\text{ of samples classified correctly}}{\#\text{ of samples in the dataset}}. \quad (7)$$

Fleiss's kappa statistic is a chance-corrected measure of agreement between two sets of categorized data. It tests how agreement exceeds random chance levels. In this study, it is adopted to measure the agreement between samples' true labels and samples' classified labels from different algorithms. The kappa result ranges from -1 to 1 (a negative kappa value occurs when agreement is weaker than expected by chance). Higher kappa values mean stronger agreement. A kappa value of 1 means perfect agreement. Interpretation of the kappa values is based on Landis's categories [23], shown in Table 1.

Table 1. Interpretation of the kappa values

Kappa values	Interpretation
0.00 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

C4.5, designed by Quinlan [24], is employed to build classification trees. The feature selection measure in each node is gain ratio. PCA with DFA is implemented in Matlab (the MathWorks Inc.). PCA considers the first 10 principal components (PCs). Figure 6 shows that the first 10 PCs cover 97% of the variance. DFA uses the linear discriminant function. SIMCA is implemented in Matlab.

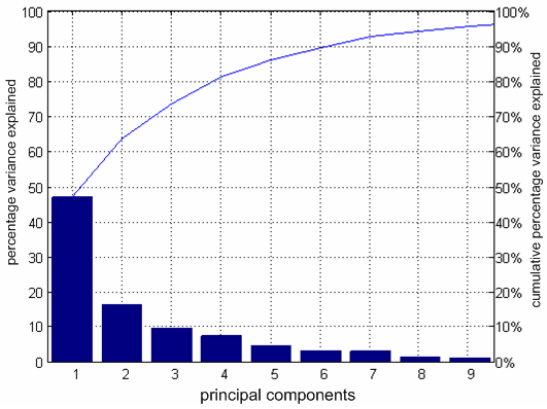


Figure 6. Variance explained by PCs

Table 2 shows the overall classification accuracy and Fleiss's kappa statistic of each classification algorithm. MSN-PSSM outperforms the other three algorithms with the highest overall classification accuracy. The samples' classified labels generated by MSN-PSSM have substantial agreement with the samples' true labels. The other three classification algorithms hold moderate agreement.

Table 2. Performance of classifiers

Classifier	Accuracy (%)	Kappa
Decision trees	45.61	0.43
SIMCA	48.54	0.51
PCA with DFA	59.65	0.57
MSN-PSSM	73.68	0.72

There are two advantages of MSN-PSSM which lead to this outperformance. First, MSN-PSSM models the intra-class variability in the similarity measure. This enhances the robustness of MSN-PSSM. The decision trees algorithm doesn't consider intra-class variability. PCA with DFA captures the variability between all the

samples, but not the intra-class variability. Second, MSN-PSSM uses all the information (all the peaks) in the entire ToF-SIMS spectra. The other three algorithms are based on reduction of the dimensionality which loses information. Table 3 shows how MSN-PSSM outperforms the other three algorithms in classification of the nine samples from one strain of *Enterococcus spp*. Sample 8 and 9 are classified correctly by all algorithms. MSN-PSSM classifies all samples from 1 to 7 correctly. The other three algorithms wrongly classify some samples from 1 to 7 to other strains of *Enterococcus spp*. The decision trees algorithm even classifies sample 7 to one strain of *Klebsiella pneumoniae*. Table 3 also shows bacterial samples from the same species are more similar and difficult to discriminate. Table 4 shows how MSN-PSSM outperforms the other three algorithms in classification of the nine samples from one strain of *Citrobacter freundii*. Sample 9 is classified incorrectly by all algorithms. This sample might be either a difficult sample to classify or an outlier which is contaminated by some kind of noise. MSN-PSSM classifies all samples from 1 to 8 correctly, the other three algorithms incorrectly classify some samples from 1 to 8 to other strains.

Table 3. Classification results for one strain of *Enterococcus spp*

Sample ID	Decision trees	SIMCA	PCA with DFA	MSN-PSSM	True label
1	Esp3	Esp2	Esp3	Esp3	Esp3
2	Esp2	Esp3	Esp2	Esp3	Esp3
3	Esp3	Esp3	Esp2	Esp3	Esp3
4	Esp2	Esp3	Esp2	Esp3	Esp3
5	Esp3	Esp3	Esp4	Esp3	Esp3
6	Esp3	Esp2	Esp4	Esp3	Esp3
7	Kpn3	Esp3	Esp4	Esp3	Esp3
8	Esp3	Esp3	Esp3	Esp3	Esp3
9	Esp3	Esp3	Esp3	Esp3	Esp3
Correct	6	7	3	9	

Table 4. Classification results for one strain of *Citrobacter freundii*

Sample ID	Decision trees	SIMCA	PCA with DFA	MSN-PSSM	True label
1	Cfr1	Kox	Cfr1	Cfr1	Cfr1
2	Cfr1	Kpn2	Esp1	Cfr1	Cfr1
3	Pmi4	Pmi1	Cfr1	Cfr1	Cfr1
4	Eco3	Cfr1	Eco4	Cfr1	Cfr1
5	Eco4	Cfr1	Eco4	Cfr1	Cfr1
6	Cfr2	Cfr1	Eco4	Cfr1	Cfr1
7	Pmi3	Cfr1	Eco4	Cfr1	Cfr1
8	Eco4	Cfr1	Eco4	Cfr1	Cfr1
9	Eco1	Esp3	Kox	Eco1	Cfr1
Correct	2	5	2	8	

4.2. Cluster analyses

Bacterial samples from the same species are more similar than samples from different species. Species-level cluster analyses are presented here to provide first-stage guidance to reveal biological differences between samples. The six species of UTI bacteria are examined as six clusters. The cluster labels are Cfr, Eco, Esp, Kox, Kpn, and Pmi. Cfr has 18 samples, Eco has 45 samples, Esp has 36 samples, Kox has nine samples, Kpn has 27 samples, and Pmi has 36 samples. This is a challenging unbalanced multi-cluster clustering task to quantitatively demonstrate the separation of the bacterial species.

The bacterial samples are clustered by four clustering algorithms: HCA, k-means, PCA with k-means, and normalized spectral clustering algorithm. Average cluster accuracy [25] is used to quantitatively measure the performance of the different algorithms. Given a set of n unlabeled spectra $\{x_i\}$, where i=1, 2, ..., n, k predefined partitions $\{P_1, P_2, ..., P_k\}$, k clusters $\{C_1, C_2, ..., C_k\}$ resulting from a specific clustering algorithm, and a optimal correspondence between $\{C_1, C_2, ..., C_k\}$ and $\{P_1, P_2, ..., P_k\}$, average cluster accuracy is defined as:

$$\text{Accuracy} = \frac{\#\text{ of samples correctly assigned}}{\#\text{ of samples in the dataset}}. \qquad (8)$$

A large value for this measure indicates a high level of agreement between the clusters and the predefined natural partitions.

HCA, k-means, PCA with k-means, and normalized spectral clustering are implemented in Matlab. All algorithms use Euclidean distance as the distance measure. HCA uses the complete linkage. PCA uses the first 10 principal components.

Table 5 shows the average cluster accuracy and the confusion matrix of each clustering algorithm. Normalized spectral clustering outperforms the other three algorithms with the highest average accuracy.

Table 5 also shows that PCA is not able to improve the performance of k-means by reducing the dimensionality, and normalized spectral clustering is able to enhance the cluster-properties in the spectra by changing the representation of the original spectra. To show this in more detail, Figure 7 presents the 2D plot of the 36 *Proteus mirabilis* samples and the 36 *Enterococcus spp* samples in the first two eigenvector (eigenvectors corresponding to the first two smallest eigenvalues) space of normalized spectral clustering. Figure 7 shows clearly two clusters without any overlap after changing the representation of data, and this makes k-means be able to detect clusters accurately.

Figure 8 presents the 3D plot of the same samples in the first three principal components space. The X axis represents the first principal component, the Y axis the second principal component, and the Z axis the third

principal component. Figure 9 shows the corresponding 2D plot in three views. Figure 8 and 9 show that the *Proteus mirabilis* samples in the principal component space form a concave set which k-means can not cluster correctly. PCA is not able to enhance the cluster-properties to improve the performance of k-means. Sample 25 and 27 of *Proteus mirabilis* are incorrectly clustered together with *Enterococcus spp* samples in k-means and PCA with k-means algorithms.

Table 5. Performance of clustering algorithms

Clustering algorithms	Accuracy (%)	Confusion matrix
HCA	38.60	0 0 0 0 18 0 0 2 0 0 20 23 0 0 8 0 12 16 0 0 0 0 1 8 0 0 0 0 27 0 1 0 0 1 5 29
k-means	53.80	18 0 0 0 0 0 2 21 0 2 14 6 0 6 14 2 10 4 0 5 0 0 1 3 4 1 0 8 14 0 2 8 0 0 1 25
PCA with k-means	53.80	18 0 0 0 0 0 2 20 1 0 15 7 2 4 17 0 10 3 0 6 1 0 0 2 12 1 0 0 14 0 0 7 3 2 1 23
Normalized spectral clustering	61.99	16 0 0 0 2 0 2 32 8 2 0 1 2 6 25 0 0 3 0 6 3 0 0 0 3 9 0 0 15 0 0 13 3 0 2 18

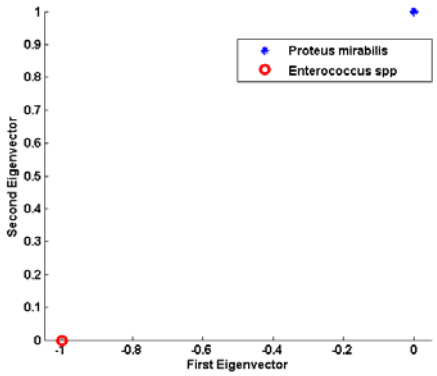


Figure 7. 2D plot of Pmi and Esp samples in eigenvector space of spectral clustering

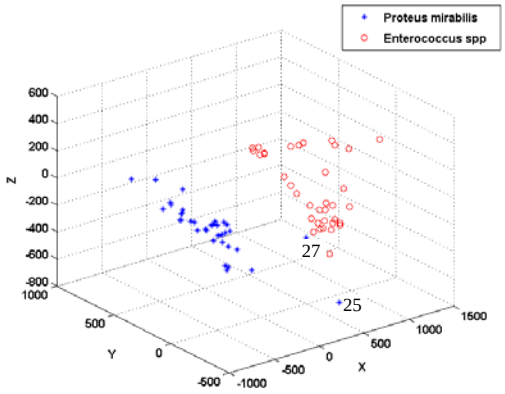


Figure 8. 3D plot of Pmi and Esp samples in principal components space

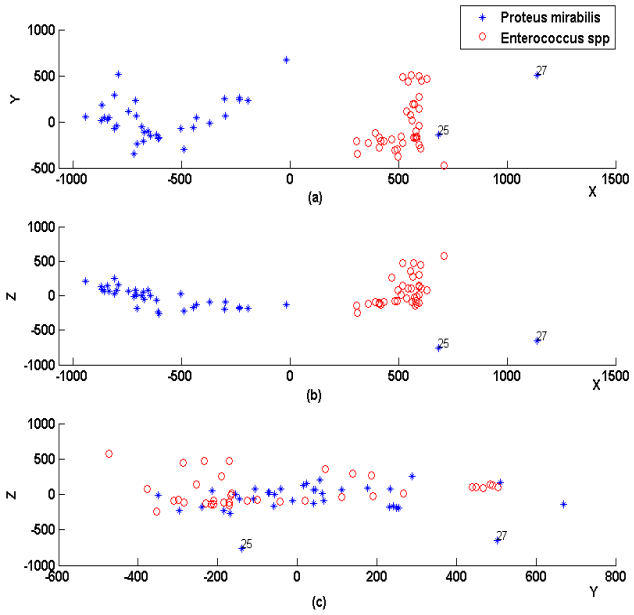


Figure 9. 2D plot of Pmi and Esp samples in principal components space

5. Conclusions

This study presents MSN-PSSM, a new classification algorithm. MSN-PSSM is applied to demonstrate strain-level discriminants of information-rich ToF-SIMS spectra for bacterial samples which are known to be major causal agents of UTI. MSN-PSSM utilizes all the information in the complex ToF-SIMS spectra and considers intra-class variability to build similarity models. These two advantages allow MSN-PSSM to accurately classify the subtly different bacterial samples and outperform traditional classification algorithms such as decision trees, PCA with DFA, and SIMCA. MSN-PSSM provided the best classification result in leave-one-out cross validation experiments. Species-level separation is achieved with normalized spectral clustering to provide first-stage guidance about biological differences between bacterial

samples. Normalized spectral clustering enhances the cluster-properties in the ToF-SIMS spectra. Experimental results demonstrate that normalized spectral clustering successfully separates bacterial samples and outperforms popular clustering algorithms such as HCA, k-means, and PCA with k-means.

Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council's "Collaborating for Success through People" funding to John C. Vickerman (EP/FO12985), by the Faculty Development Leave program of the University of Nebraska-Lincoln, and by the USA National Science Foundation funding to S. E. Reichenbach (IIS-0431119) and Q. Tao (IIP-0741027). The authors gratefully acknowledge the support and data provided by John Vickerman and John Fletcher of the Surface Analysis Research Centre, University of Manchester.

References

[1] O. D. Sanni, M. S. Wagner, D. Briggs, D. G. Castner, and J. C. Vickerman, "Classification of Adsorbed Protein Static ToF-SIMS Spectra by Principal Component Analysis and Neural Networks", *Surface and Interface Analysis*, 33(9), August 2002, pp. 715–728.

[2] A. M. Belu, D. J. Graham, and D. G. Castner, "Time-of-flight Secondary Ion Mass Spectrometry: Techniques and Applications for the Characterization of Biomaterial Surfaces", *Biomaterials*, 24(21), September 2003, pp. 3635–3653.

[3] L. A. McDonnell, and R. M.A. Heeren, "Imaging Mass Spectrometry", *Mass Spectrometry Reviews*, 26, 2007, pp. 606–643.

[4] H. Oechsner, *Thin Film and Depth Profile Analysis*, Springer-Verlag, Berlin, 1984.

[5] M. S. Wagner, D. J. Graham, B. D. Ratner, and D. G. Castner, "Maximizing Information Obtained from Secondary Ion Mass Spectra of Organic Thin Films Using Multivariate Analysis", *Surface Science*, 570(1-2), October 2004, pp. 78–97.

[6] D. J. Graham, M. S. Wagner, and D. G. Castner, "Information from Complexity: Challenges of ToF-SIMS Data Interpretation", *Applied Surface Science*, 252(19), July 2006, pp. 6860–6868.

[7] J. S. Fletcher, A. Henderson, R. M. Jarvis, N. P. Lockyer, J. C. Vickerman, and R. Goodacre, "Rapid Discrimination of the Causal Agents of Urinary Tract Infection Using ToF-SIMS with Chemometric Cluster Analysis", *Applied Surface Science*, 252(19), July 2006, pp. 6869–6874.

[8] C. E. Thompson, J. Ellis, J. S. Fletcher, R. Goodacre, A. Henderson, N. P. Lockyer, and J. C. Vickerman, "ToF-SIMS Studies of Bacillus Using Multivariate Analysis with Possible Identification and Taxonomic Applications", *Applied Surface Science*, 252(19), July 2006, pp. 6719–6722.

[9] M. J. Baker, M. D. Brown, E. Gazi, N. W. Clarke, J. C. Vickerman, and N. P. Lockyer, "Discrimination of Prostate Cancer Cells and Non-malignant Cells Using Secondary Ion Mass Spectrometry", *Analyst*, 133, 2008, pp. 175–179.

[10] E. S. F. Berman, L. Wu, S. L. Fortson, K. S. Kulp, D. O. Nelson, and K. JenWu, "Chemometric and Statistical Analyses of ToF-SIMS Spectra of Increasingly Complex Biological Samples", *Surface and Interface Analysis*, 41(2), December 2008, pp. 97–104.

[11] S. Wold, and M. Sjostrom, "SIMCA: A Method for Analyzing Chemical Data in terms of Similarity and Analogy", *Chemometrics Theory and Application*, American Chemical Society Symposium Series 52, Washington, D.C., 1977, pp. 243–282.

[12] T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[13] S. K. Murthy, "Automatic Construction of Decision Trees from Data: A Multi-disciplinary Survey", *Data Mining and Knowledge Discovery*, 2(4), December 1998, pp. 345–389.

[14] C. Engrand, J. Kissel, F. R. Krueger, P. Martin, J. Silen, L. Thirkell, R. Thomas, and K. Varmuza, "Chemometric Evaluation of Time-of-flight Secondary Ion Mass Spectrometry Data of Minerals in the Frame of Future in situ Analyses of Cometary Material by Cosima onboard ROSETTA", *Rapid Communications in Mass Spectrometry*, 20(8), March 2006, pp. 1361–1368.

[15] S. Theodoridis, and K. Koutroumbas, *Pattern Recognition*, Academic Press, 1999.

[16] N. Suzuki, M. Sarikaya, and F. S. Ohuchi, "Adsorption of Genetically Engineered Proteins Studied by Time-of-flight Secondary Ion Mass Spectrometry (ToF-SIMS). Part B: Hierarchical Cluster Analysis (HCA)", *Surface and Interface Analysis*, 39(5), February 2007, pp. 427–433.

[17] B. Foxman, R. Barlow, H. D'Arcy, B. Gillespie, and J. Sobel, "Urinary Tract Infection: Self-Reported Incidence and Associated Costs", *Annals of Epidemiology*, 10(8), November 2000, pp. 509–515.

[18] A. Visvanathan, and S. Reichenbach, "Information Theoretic Mass Spectral Library Search for Comprehensive Two-Dimensional Gas Chromatography with Mass Spectrometry", *56th ASMS Conference on Mass Spectrometry*, Denver, CO, June 2008.

[19] U. V. Luxburg, "A Tutorial on Spectral Clustering", *Statistics and Computing*, 17(4), August 2007, pp. 395–416.

[20] F. R. Bach, and M. I. Jordan, "Learning Spectral Clustering, with Application to Speech Separation", *Journal of Machine Learning Research*, 7, December 2006, pp. 1963–2001.

[21] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm", *Advances in Neural Information Processing Systems 14*, MIT Press, September 2002, pp. 849–856.

[22] J. L. Fleiss, "Measuring Nominal Scale Agreement among Many Raters", *Psychological Bulletin*, 76(5), 1971, pp. 378–382.

[23] J. R. Landis, and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data", *Biometrics*, 33, 1977, pp. 159–174.

[24] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, California, 1993.

[25] A. Topchy, A. Jain, and W. Punch, "Combining Multiple Weak Clusterings", *Proc. Third IEEE International Conference on Data Mining (ICDM'03)*, November 2003, pp. 331–338.