

Rapid discrimination of the causal agents of urinary tract infection using ToF-SIMS with chemometric cluster analysis

John S. Fletcher^a, Alexander Henderson^a, Roger M. Jarvis^b, Nicholas P. Lockyer^a,
John C. Vickerman^a, Royston Goodacre^b

^aSurface Analysis Research Centre, School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester M60 1QD, UK

^bSchool of Chemistry, The University of Manchester, Manchester M60 1QD, UK

Abstract

Advances in time of flight secondary ion mass spectrometry (ToF-SIMS) have enabled this technique to become a powerful tool for the analysis of biological samples. Such samples are often very complex and as a result full interpretation of the acquired data can be extremely difficult. To simplify the interpretation of these information rich data, the use of chemometric techniques is becoming widespread in the ToF-SIMS community. Here we discuss the application of principal components - discriminant function analysis (PC-DFA) to the separation and classification of a number of bacterial samples that are known to be major causal agents of urinary tract infection. A large data set has been generated using three biological replicates of each isolate and three machine replicates were acquired from each biological replicate. Ordination plots generated using the PC-DFA are presented demonstrating strain level discrimination of the bacteria. The results are discussed in terms of biological differences between certain species and with reference to FT-IR, Raman spectroscopy and pyrolysis mass spectrometric studies of similar samples.

Keywords: ToF-SIMS, chemometrics, multivariate analysis, urinary tract infection.

1. Introduction.

Well established for the analysis of inorganic and organic samples, secondary ion mass spectrometry (SIMS) is now being applied to an increasing range of biological material. Advances in ion beam technologies mean that the use of polyatomic ion beams such as SF₅⁺, Au_n⁺, Bi_n⁺ and C_n⁺ (particularly C₆₀⁺) are now widespread and such beams have been shown to produce significant increases in sputter yield with a particular enhancement in the higher mass region where the more chemically characteristic and biologically relevant molecular mass fragments appear [1,2,3]. ToF-SIMS analysis of complex biological samples has a tendency to produce complex mass spectra that although extremely information rich can be very difficult to interpret. As a result, many analysts are turning to chemometrics to simplify the interpretation. Multivariate analysis has been applied to a range of samples that have been analysed to produce both conventional spectra [4] along with SIMS images [5]. Cluster analysis has also been utilised in the discrimination of yeast strains with ToF-SIMS [6].

Urinary tract infection (UTI), particularly associated with adult women, is a considerable problem in general practice and on average leads to a consultation rate of approximately 63.5 consultations in every 1000 women each year [7]. With this high incident rate of bacteruria (counts of above 10⁵ organisms/ml urine) there is a growing need to identify the causal agent prior to treatment. The bacteria typically associated with UTI include *Escherichia coli* (in over 50% of cases) and *Klebsiella* species that can be resistant to antibiotics. In addition, other

Enterobacteriaceae are implicated including *Proteus mirabilis* and *Citrobacter freundii*, whilst the Gram positive *Enterococcus spp.* also often causes infection [8].

Recently a variety of spectroscopic methods, which fall under the umbrella of ‘whole organism fingerprinting’ have been used to analyse the bacteria associated with UTI and these have included Fourier transform infrared spectroscopy (FT-IR) [9], Raman spectroscopies [10,11] and Curie-point pyrolysis mass spectrometry (PyMS) [9]. The analysis of these high dimensional data has typically been performed by the cluster analysis method of principal components-discriminant function analysis (PC-DFA) which has facilitated the applicability of these spectroscopic methods for high throughput screening. With this method the PC-DFA multivariate algorithm seeks “clusters” in the data, thereby allowing the investigator to group objects together on the basis of their perceived closeness in 2 or 3 dimensional PC-DFA ordination space [12]. As the results from whole-organism fingerprinting with cluster analysis studies have shown to be promising, UTI provides an excellent benchmark with which to compare the application of multivariate statistical analysis of data generated by analysis using ToF-SIMS. All the experiments reported in this study were performed under so-called “static” conditions and so the spectra collected should provide information of the outermost surface chemistry of the bacteria and therefore will probably differ from the information obtainable from the above techniques.

2. Experimental

Bacteria and sample preparation

19 strains of UTI bacteria, previously identified by conventional biochemical tests, were used in this study. These bacteria had previously been analysed by FT-IR, Raman and PyMS. The isolates consisted of *Escherichia coli* (5 isolates coded ‘Eco’), *Klebsiella oxytoca* (1 isolate coded ‘Kox’), *Klebsiella pneumoniae* (3 isolates coded ‘Kp’), *Citrobacter freundii* (2 isolates coded ‘cf’), *Enterococcus spp.* (4 isolates coded ‘Ent’) and *Proteus mirabilis* (4 isolates coded ‘Pm’). The isolates were cultured axenically on LabM Malthus blood agar base (37 mg ml⁻¹) aerobically for 16 h at 37 °C. After sub-culturing three times to ensure purity and phenotypic homogeneity, three ‘biological replicates’ were generated by cultivation of three separate colonies of each isolate on three separate nutrient agar plates. The biomass was carefully collected in distilled H₂O and stored at -80 °C prior to ToF-SIMS analysis.

ToF-SIMS

The BioToF-SIMS instrument used for the analysis has been described in detail elsewhere [13]. In brief, the instrument consists of a two stage reflectron analyser, gold liquid metal and C₆₀ electron impact primary ion sources (Ionoptika Ltd., UK). A sample preparation chamber with an integral fast entry load-lock and freeze-fracture system allow the efficient introduction of frozen material and analysis can be performed at low temperatures by use of a liquid nitrogen cooled sample stage.

3 µl droplets of each sample were transferred to individual pieces of silicon and air dried. Six silicon pieces were attached to each copper sample stub and three sample stubs were introduced into the ToF-SIMS instrument at a time. ToF-SIMS analysis was performed on each of the 18 samples in the instrument using a mass filtered Au₃⁺ primary ion beam with an impact energy of 20 keV, scanned over a 500 × 500 µm² field-of-view, producing a dose density of approximately 7 × 10¹⁰ ions cm⁻². Charge compensation was employed using 25 eV electrons between ion pulses. This process was repeated 3 times to generate 3 ‘machine replicates’, each

spectrum being acquired from a fresh area of the relevant sample. Hence nine spectra were acquired of each strain; 171 spectra in total.

Cluster analysis

The initial step in cluster analysis involved the reduction of the dimensionality of the data by principal components analysis (PCA) [14]. PCA is a well known technique for reducing the dimensionality of multivariate data whilst preserving most of the variance. Matlab (The MathWorks Inc., Natick, MA, USA) was employed to perform PCA according to the NIPALS algorithm [15]. Plots of the first two principal components scores represent the best 2-D representation of natural variance in the data. Discriminant function analysis (DFA) then discriminated between groups on the basis of the retained principal components (PCs) and the *a priori* knowledge of which spectra were acquired from which biological replicates [16]. Therefore this process does not bias the analysis in any way. DFA was programmed according to Manly's principles [17], and was programmed to maximise the Fisher ratio (i.e., the within-class to between-class variance). The spectral similarity between different classes reflects the optimal number of PCs that are fed into the DFA algorithm. Typically 3 PCs were used for DFA and this typically accounted for *ca.* 90-95% of the explained variance. The principle of DFA is similar to that of PCA, but as the objective is to maximise the Fisher ratio, the plot of the first two discriminant function scores will represent the best 2-D representation of group separations. For both PCA and DFA, in addition to the scores matrices mentioned above, a PCA or DFA loadings matrix is produced which indicates which inputs (mass peaks) are most informative either for natural variance in the data (PCA) or for maximal group separations (DFA).

3. Results and Discussion

ToF-SIMS spectra of four of the isolates are reproduced in Fig. 1. Although dominated by the electropositive ions Na^+ and K^+ , a number of peaks are observed of varying intensity over a wide mass range. Although the high levels of sodium and potassium in the sample may introduce matrix effects into the spectra, and result in the formation of cationised fragments, such effects have not been considered in this study. As the aim of the experiment was to demonstrate the ability of computational methods to discriminate between SIMS spectra of bacterial samples in a *pseudo* natural state no effort was made to reduce the salt content of the samples. The individual strains give rise to spectra containing many common ions and identification of an individual specimen by visual inspection of the spectrum would be extremely difficult since separation would be based on the changes in the relative intensities of a number of these common peaks. This data set is therefore highly suited to test the applicability of multivariate analysis to SIMS data.

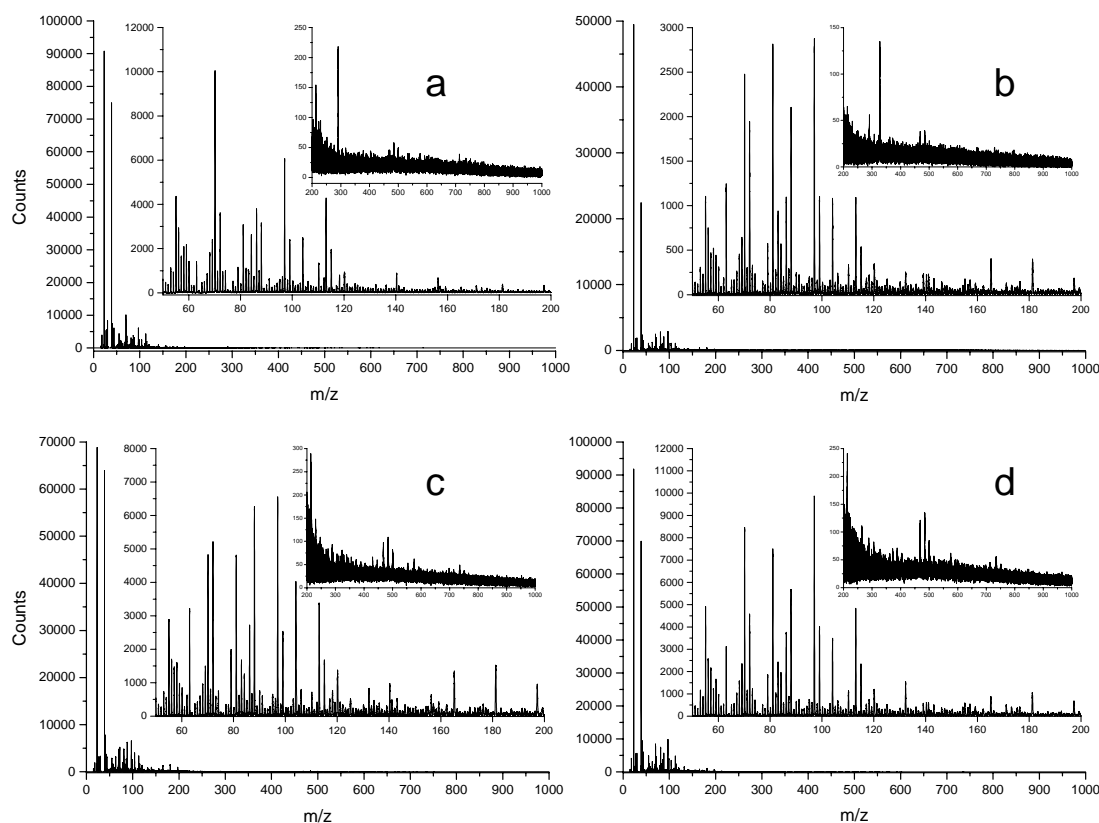


Fig. 1. ToF-SIMS spectra of (a) *C. freundii* [cf102], (b) *E. coli* [Eco13], (c) *Enterococcus* sp. [Ent93] and (d) *K. oxytoca* [Kox105].

In the multivariate analysis of SIMS data it is often common to “peak pick” to a certain extent. This can be performed using statistical analysis to assess the reproducibility or significance of an individual peak in the spectrum so that a decision is made as to its incorporation in the multivariate analysis. Alternatively the analyst themselves may, based on experience, choose to select certain values of m/z for incorporation into the multivariate analysis, a common example of the latter is to ignore the low mass region (e.g. m/z 0-50) of the spectrum where salt contamination is often evident and the peaks are less molecule specific.

Pre-processing in the experiment reported here was kept to a minimum. The only modifications to the spectra were as follows. Prior to multivariate analysis, each spectrum was truncated to 1000 Da. and then bin-summed into 1 Da. mass steps each ranging from -0.5 to +0.5 of each nominal mass. These new data points were then expressed as a proportion (between 0 and 1) of the sum of the spectral intensity. Binning the data to 1 Da. may result in a loss of chemical specificity as there may be more than one species present at each nominal mass, however for the chemometric analysis a fingerprint is produced with the intensities of the nominal masses being characteristic to the sample.

Cluster analysis was performed as described above. The dimensionality of the data was reduced using PCA prior to DFA where *a priori* knowledge of which spectra corresponded to which isolate was used to maximise the Fisher ratio. In each case the *minimum* number of PCs was used to provide maximum separation of each cluster without over-fitting the data. Following the PCA step, no significant clustering of the data was observed and inspection of the loadings plots (Fig. 2) indicates that most of the variance between spectra was related to the changes in the intensities of the Na^+ and K^+ signals that dominated the spectra. However, the subsequent application of DFA, produced loadings that show the inclusion of a large number of higher mass

fragments, with a relatively small influence from Na⁺ and K⁺. The effect of the overwhelming contribution to the PC loadings of the sodium and potassium signal was investigated performing the PCA with the mass range 0-50 removed. Comparison of the ordination plots show that by removing this region of the spectrum the data begins to cluster with some separation of the *Enterococci* and *Proteus mirabilis* (Fig. 3).

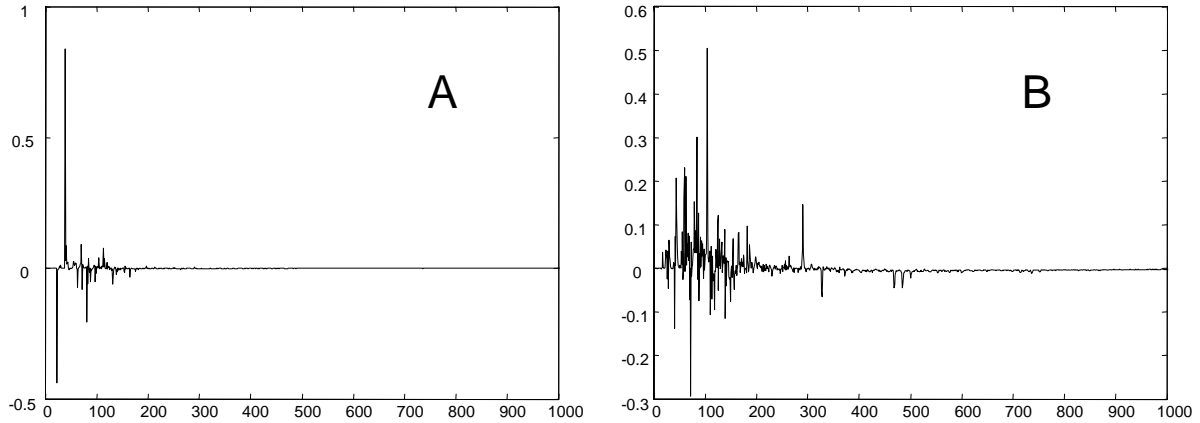


Fig. 2. Loadings on (A) PC1 and (B) DF1. Numbering on the x-axis indicates m/z value. Loadings shown are from the PCA and PC-DFA of the entire data set.

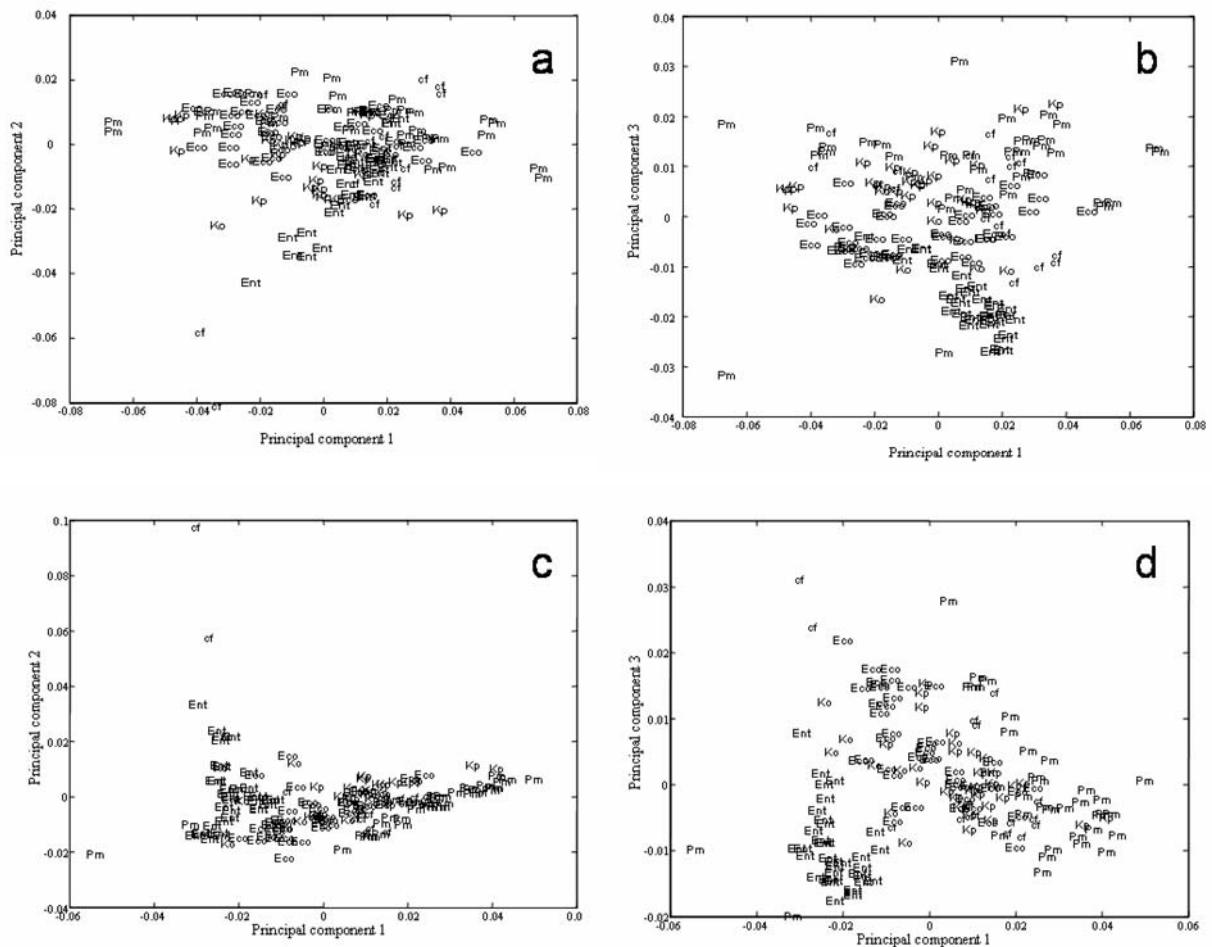


Fig. 3 Ordination plots following the PCA of the data. Plots a and b result from analysis over the entire mass range and plots c and d from analysis performed excluding the mass range 0-50.

PC-DFA of the entire data set, where the classes used in the DFA represented *individual* strains, was performed and the ordination plots are presented in Fig.4. The PC-DFA utilised three principal components. Although clear separation of the individual strains is not evident, the isolates have clustered at the *species* level with the enterococci isolates (highlighted in red) forming a cluster that is well separated from the remainder of the group. The recovery of the *Enterococcus* spp. from the other bacteria is expected, since this is the only species of bacteria in this study to be Gram-positive. These bacteria would therefore have a distinctly different cell membrane in comparison with the Gram-negative *Enterobacteriaceae*, incorporating a thick outer layer of peptidoglycan. Isolates of the *P. mirabilis* also cluster into a well defined group and this result can also be explained by the phenotype of these organisms. The *P. mirabilis* cell walls contain a polysaccharide-rich capsule layer which includes *N*-acetyl-D-glucosamine that is also present in peptidoglycan. As the SIMS analysis is expected to probe only the outer surface of the bacterial sample, it is understandable that such biological differences would have a strong influence on the spectra and hence the clustering observed in this multivariate analysis.

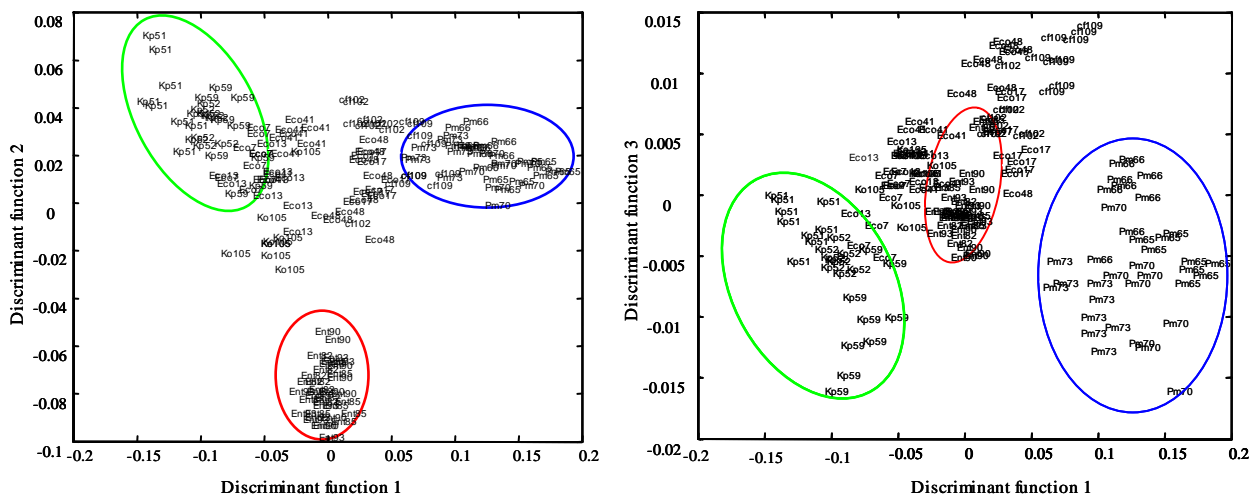


Fig.4. PC-DFA ordination plots of for the entire data set. For clarity three species that are separated from the main cluster have been highlighted; *Enterococcus* spp. (red), *P. mirabilis* (blue), *K. pneumoniae* (green).

For further analysis all the Gram-positive enterococci isolates were removed from the data set; the PC-DFA was repeated; and again, three principal components were used to generate the PC-DFA model. The resulting ordination plots (Fig.5) show extensive clustering of the remaining isolates. The clustering observed is no longer at the species level but now shows good strain level separation. The *P. mirabilis* isolates are separated from the main group, as seen in the analysis of the entire data set, however distinct strain level separation is observed. The Pm70 and Pm73 strains are clearly isolated, although there is a small overlap of the Pm65 and Pm66 strains.

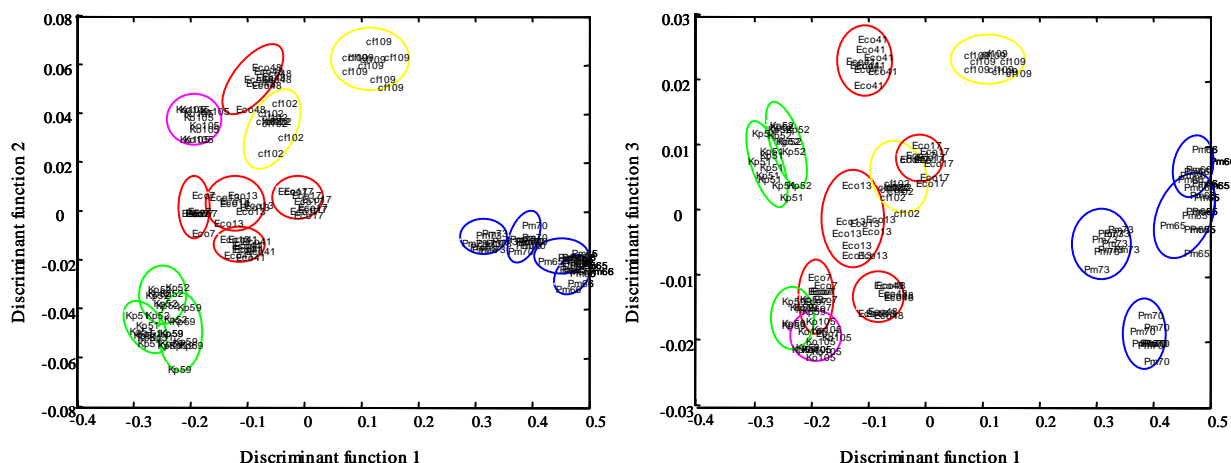


Fig.5. PC-DFA ordination plots of the data set following the removal of the enterococci. Rings have been placed around individual strains as a visual aid only. Rings of the same colour indicate strains belonging to the same species. *Escherichia coli* (red), *Citrobacter freundii* (yellow), *Proteus mirabilis* (blue), *Klebsiella pneumoniae* (blue), *Klebsiella oxytoca* (pink).

The above results clearly show that ToF-SIMS spectra are information-rich and that the chemical data therein allows for the classification of these UTI isolates to the species level using this unsupervised learning approach. The next stage was to assess whether the application of PC-DFA to ToF-SIMS spectra could provide *strain level* discrimination of the bacteria. A single species subset of the data was chosen for this exercise. The *E. coli* spectra were selected since this group contained the highest number of individual isolates ($n=5$). Three spectra of each strain were chosen randomly and removed from the group. PC-DFA was then performed on the remainder ‘training set’ to generate a model as before. The excluded ‘test’ data was then projected into the model (first into the PCA-space and then the resultant PCs projected into DFA-space) and the resulting ordination plot is shown in Fig.6 where the training set is labelled in red and the projected test data appear in blue. As with the previous analyses, and for consistency, three principal components were used for the DFA model. Fig.6 clearly shows that for each of the five isolates, the test data were recovered very close to their corresponding training data clearly showing that ToF-SIMS does provide the sensitivity and reproducibility for sub-species discrimination.

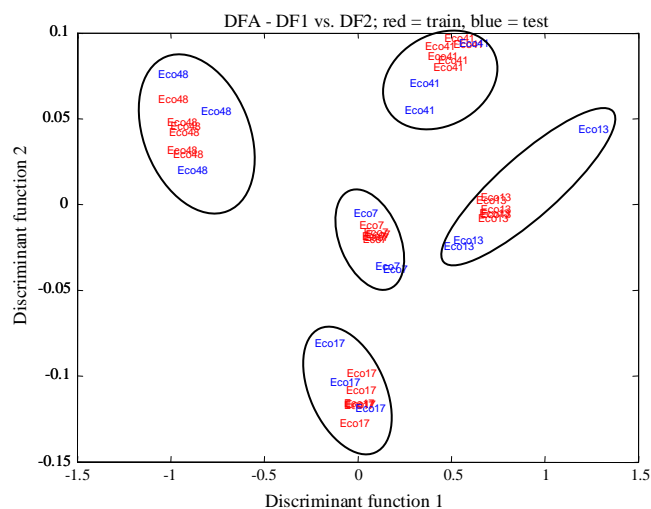


Fig.6. PC-DFA ordination plot of five strains of *E. coli*. The training data is labelled in red and the projected, test data in blue. Three principal components were used to generate the DFA model.

4. Concluding remarks

The use of PC-DFA with ToF-SIMS data demonstrates a clear ability to discriminate between the causal agents of urinary tract infection examined in this investigation. Moreover, strain level discrimination is achieved with minimal manipulation/pre-treatment of the ToF-SIMS data. These results can be compared to those from previous studies of these organisms using PyMS and FT-IR, where separation was observed only at the species level. Another surface analysis method based on surface-enhanced resonance Raman spectroscopy has also demonstrated strain level discrimination of these organisms. This may imply that surface features are particularly discriminating for these UTI bacteria. PCA alone did not achieve the adequate separation of these bacteria from their ToF-SIMS spectra. This is commonly found due to the biological variability seen when bacteria are cultured several times. Therefore the further application of DFA was used to accommodate for those inherent biological differences. Some pre-processing of the ToF-SIMS data was required to account for unavoidable differences in spectral intensity. Bin-summing and normalisation to total spectral area were performed on each spectrum. In conclusion, we believe that ToF-SIMS presents itself as a complementary, whole-organism, fingerprinting approach for the rapid characterisation of bacteria. Further work will attempt to identify key, discriminatory metabolites that may serve as biomarkers for the identification of bacteria.

Acknowledgements

This work was funded by the UK's Biotechnology and Biological Sciences Research Council (BBSRC), under grant BBSB03920.

References

-
- [1] F. Kotter, A. Benninghoven, *Applied Surface Science*, 133 (1998) 47
 - [2] N. Davies, D.E. Weibel, P. Blenkinsopp, N. Lockyer, R. Hill, J.C. Vickerman, *App. Surf. Sci.* 223 (2003) 203
 - [3] D.E. Weibel, S. Wong, N.P. Lockyer, P. Blenkinsopp, R. Hill, J.C. Vickerman, *Anal. Chem.* 75 (2003) 1754
 - [4] M.S. Wagner, B.J. Tyler, D.G. Castner, *Anal. Chem.* 74 (2002) 1824
 - [5] B. Tyler, *App. Surf. Sci.* 203 (2003) 825
 - [6] H. Jungnickel, E.A. Jones, N.P. Lockyer, S.G. Oliver, G.M. Stephens, J.C. Vickerman, *Anal. Chem.* 77 (2005) 1740
 - [7] M.E. Wilkie, M.K. Almond, F.P. Marsh, *British Medical Journal* 303 (1992) 1137
 - [8] R.C.B. Slack, *Urinary Infections*, in: D. Greenwood (ed.), *Antimicrobial Chemotherapy*, Oxford University Press, Oxford, 1995, p. 243
 - [9] R. Goodacre, É.M. Timmins, R. Burton, N. Kaderbhai, A.M. Woodward, D.B. Kell, P.J. Rooney, *Microbiology* 144 (1998) 1157
 - [10] R.M. Jarvis, R. Goodacre, *Anal. Chem.* 76 (2004) 40
 - [11] R.M. Jarvis, R. Goodacre, *FEMS Microbiology Letters* 232 (2004) 127
 - [12] B.S. Everitt, *Cluster Analysis*, Edward Arnold, London (1993)

-
- [13] R.M. Braun, P. Blenkinsopp, S.J. Mullock, C. Corlett, K.F. Willey, J.C. Vickerman, N. Winograd, *Rapid Comm. Mass Spec.* 12 (1998) 1246
- [14] I.T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, 1986
- [15] H. Wold, Estimation of principal components and related models by iterative least squares, in: K.R. Krishnaiah (Ed.), *Multivariate Analysis*, Academic Press, New York, 1966, pp. 391-420.
- [16] W. Windig, J. Haverkamp, P.G. Kistemaker, *Anal. Chem.* 55 (1983) 81
- [17] B.F.J. Manly, *Multivariate Statistical Methods: A Primer*, Chapman & Hall, London, 1994