Explanatory multivariate analysis of ToF-SIMS spectra for the discrimination of bacterial isolates

Seetharaman Vaidyanathan,†*[a] John S. Fletcher,[a] Roger M. Jarvis,[b] Alex Henderson,[a] Nicholas P. Lockyer,[a] Royston Goodacre[b] and John C. Vickerman[a]

Received 15th April 2009, Accepted 26th August 2009
First published as an Advance Article on the web 14th September 2009
DOI: 10.1039/b907570d

Multivariate analysis (PC-CVA and GA-CVA) was carried out on time-of-flight secondary ion mass spectra (ToF-SIMS) derived from 16 bacterial isolates associated with urinary tract infections, with an objective of extracting the spectral information relevant to their species-level discrimination. The use of spectral pre-processing, such as removal of the dominant peaks prior to analysis and analysis of the dominant peaks alone, enabled the identification of 37 peaks contributing to the principal components-canonical variates analysis (PC-CVA) discrimination of the bacterial isolates in the mass range of *m*/*z* 1–1000. These included signals at *m*/*z* 70, 84, 120, 134, 140, 150, 175 and 200. A univariate statistical analysis (Kruskal–Wallis) of the signal intensities at the identified *m*/*z* enabled an understanding of the discriminatory basis, which can be used in the development of robust parsimonious models for predictive purposes. The utility of genetic algorithm (GA)-based feature selection in identifying the discriminatory variables is also demonstrated. A database search of the identified signals enabled the biochemical origins of some these signals to be postulated.

Introduction

Time-of-flight secondary ion mass spectrometry (ToF-SIMS) is a surface technique that yields spectral information useful in discerning chemical changes associated with the surface being analysed.[1] The advent of cluster ion sources, such as C_{60}^+,[2] has enabled the application of the technique to derive molecular information from surfaces.[3] Of particular interest is the development of the technique to analyse biological surfaces.[4,5] The potential of the technique in discriminating bacterial isolates has been demonstrated in earlier investigations.[6,7] Although these investigations demonstrate the ability of the technique to generate spectral information for discriminatory purposes, the value of the technique will be strengthened by seeking explanatory analysis of the discriminating variables, both in terms of chemistry and the associated biology. In this investigation we have sought to explain the discrimination of 16 bacterial isolates associated with urinary tract infections by employing multivariate analysis of the ToF-SIMS spectral data and following it with univariate statistics and bioinformatics.

Multivariate analysis methods such as principal component analysis (PCA) and discriminant function analysis (DFA) have been shown to be useful in exploring the ToF-SIMS spectral information.[6–9] Deconvolution of the spectra with a view to understanding the basis of such analyses is a key challenge that will enable construction of robust models for predictive purposes. An earlier investigation discussed the application of PC-DFA to the discrimination of bacterial isolates associated with urinary tract infection (UTI), based on their ToF-SIMS spectra.[6] UTI, prevalent in adult women, is a considerable problem in general practice with high consultation rates,[10] and there is a growing need for rapid methods to screen for causal agent(s) prior to antibiotic treatment.

In this investigation, we report the application of PC-canonical variates analysis[11,12] (PC-CVA) and genetic algorithms (GAs) on the spectral dataset with a view to extracting the spectral information relevant to the discrimination at the species level. Through the interpretation of spectral loadings plots, where high loadings for a factor indicate spectral components of particular importance to the discrimination, it is possible to discern the chemical basis of the classification model.[13] However, for complex data, such as mass spectra, where many independently measured variables are recorded, spectral loadings plots recovered from PC-CVA can be very difficult to interpret, and it is not always apparent which *combinations* of small numbers of variables have good discriminatory ability. Notwithstanding, it is desirable to reduce the solution to a classification problem down to a handful of spectral variables so that simple, interpretable rules can be achieved. Therefore, in order to determine variable subset combinations that contribute most to the class separation observed with PC-CVA, GA feature selection coupled directly to CVA was examined. GAs belong to a group of evolutionary algorithms that have been shown to be useful in feature selection of spectroscopic data.[14–18] In a feature selection context, the GA is used to select small subsets of spectral variables to assess against a cost function; in this case we seek to maximise the between-group variance and minimise the within-group variance between *a priori* classes in CVA scores space.

[a]*School of Chemical Engineering and Analytical Science, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester, UK M1 7ND. E-mail: S.Vaidyanathan@sheffield.ac.uk*
[b]*School of Chemistry, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester, UK M1 7ND*
† Present address: ChELSI, Department of Chemical & Process Engineering, University of Sheffield, Sheffield, UK S1 3JD.

Experimental

ToF-SIMS spectra of bacterial isolates

The spectral dataset used in the study was acquired in an investigation reported earlier,[6] and consisted of positive ion ToF-SIMS spectra of bacterial isolates associated with urinary tract infection. The growth conditions and ToF-SIMS data acquisition are detailed elsewhere.[6] ToF-SIMS spectra of 16 isolates that belong to six bacterial species were investigated (Table 1). Each bacterial isolate was cultured separately three times (biological replicates) and each of these was analysed three times (instrumental replicates) resulting in 144 spectra.

Principal component-canonical variates analysis (PC-CVA)

Spectral data analysis was carried out in MATLAB (The MathWorks, Natick, MA, USA) using locally written routines, unless mentioned otherwise. Locally written software was used to convert the ToF-SIMS spectra to ASCII format and the spectra were binned to 1 *m*/*z*, centred on integer mass positions, and imported into MATLAB in ASCII format. All spectra were normalised to total ion counts before analysis. Principal components analysis (PCA) was used to reduce the dimensionality of the multivariate data and was carried out using singular value decomposition (SVD). Canonical variates analysis (CVA) was then used on the PC scores to discriminate between the groups on the basis of the retained principal components (PCs) and *a priori* knowledge of the group structure. The number of PCs used accounted for >95% of the explained variance, and varied from 12 to 31. The choice of the number of PCs to use was made after visual inspection of the resultant CVA score biplots. The *a priori* information used was the isolate-wise group structure (16 groups). To optimise the exact number of PCs to use and hence to avoid over-fitting and assure mode generality, cross-validation with independent test sets was adopted. The data were grouped into training and test sets by random assignments within the class structure. Three such training and test sets were created from the same data by random assortments to generate three replicates on which to perform PC-CVA.

Different pre-processing criteria were used to generate data for PC-CVA. These included (a) analysis of all the peaks in the mass range *m*/*z* 1–1000, (b) analysis of all peaks after removal of the 5, or 10, most dominant ones, and (c) analysis of only the 20, 50 or 100 most dominant peaks. Removal of any more than 10 dominant peaks did not result in stable discriminations. Likewise, analysis of any less than 20 most dominant peaks did not result in stable discriminations. The peaks contributing to the discrimination of the species were identified from the corresponding scores and loadings plots, using locally written routines. Only peaks whose loadings contributed to the specific clustering were analysed. Even with these, a threshold (loadings that were 10% of the maximum) was used to select the peaks of interest. The contribution from each replicate analysis was calculated separately and only the peaks found common to all three replicate PC-CV analyses were inspected and analysed.

Table 1 The bacterial isolates whose ToF-SIMS spectra were analysed

Species	Gram stain	Identifier on plots and tables	No. of isolates
Citrobacter freundii	G–	C	2
Escherichia coli	G–	E	4
Enterococcus spp.	G+	N	4
Klebsiella oxytoca	G–	O	1
Klebsiella pneumoniae	G–	K	2
Proteus mirabilis	G–	P	3

Genetic algorithms-canonical variates analysis (GA-CVA)

GA-CVA was carried out using MATLAB and PyChem (http://pychem.sf.net/).[16] With GA-CVA small subsets of variables are extracted from the variable superset, in this case pairs of variables were extracted from the independently measured mass spectra. Two hundred independent GA runs were performed with each run terminating after the optimal solution remained unchanged for 20 consecutive generations. Only the isolates that belonged to four of the species (*Enterococcus* spp., *E. coli*, *P. mirabilis* and *Klebsiella* spp.) were considered for the analysis to demonstrate the utility of the technique.

The GA procedure begins with a random population ($n = 100$ for this experiment) of variable subset (in this case pairs of *m*/*z* values) pairs composed into strings termed chromosomes, of given size. Each chromosome is assessed for its ability to classify the dataset based on a fitness function, which in this case is an estimate of the Fisher ratio (the ratio of within-group to between-group variance). The chromosomes are ranked on the fitness criterion. A new generation of chromosomes with higher classification accuracy is then produced from the fitter individuals of the first set, by mechanisms mimicking natural selection, such as reproduction, selection, mutation, crossover, and migration. These new chromosomes, containing new variable subset combinations replace the initial population, and the progressive improvement of the chromosome population is repeated enough times until a desired level of accuracy is reached.

Results and discussion

PC-CVA

The bacterial isolates analysed were previously identified[19] to belong to six species commonly associated with UTI (Table 1). The ToF-SIMS positive ion spectra of the bacterial isolates were analysed in the mass range of *m*/*z* 1–1000. As detailed above, different spectral pre-processing criteria were adopted to generate the data prior to PC-CVA and assess their influence and combined utility on the analysis.

The discriminatory capability of PC-CVA predominantly arises from the variance in the dataset. In this case, those variables are the signal intensities at different *m*/*z* values, and these are not uniform across the spectrum. The ToF-SIMS spectra (Fig. 1) are characterised by peaks with high intensities in the lower end of the spectrum (usually up to *m*/*z* 200), followed by less intense peaks. Since some variables dominate others, any random variations in these or their variance can potentially skew the multivariate analysis. A consideration of the spectral information that excludes the dominant peaks should account for any such interference. In addition, analysing the discrimination effected by excluding the dominant peaks or considering only the dominant peaks for analysis should give us additional

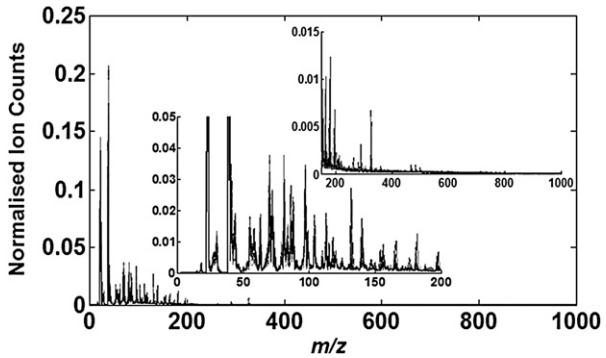

Fig. 1 An overlay of ToF-SIMS positive ion spectra of the bacterial isolates used in the study. The spectra were acquired using a 20 keV C_{60}^+ primary ion source. The different regions of the spectra are expanded in the inset.

information with respect to the variables contributing to the discrimination.

Consequently, the 5 and 10 most dominant peaks were removed and the resulting spectrum analysed sequentially using PC-CVA. Similarly, the 20, 50 or 100 most dominant peaks in the spectra were also independently analysed. In each case, the spectra were normalised to total ion counts after the respective pre-processing, so there is no influence as a function of scaling. Representative pseudo-2D scores biplots of the first two canonical variates for the PC-CVA of the different pre-processed spectra are shown in Fig. 2.

When all the peaks in the spectra from *m*/*z* 1–1000 are used (Fig. 2a) three clusters can be seen to separate out, those corresponding to *Enterococcus* spp., *Klebsiella pneumoniae* and *Proteus mirabilis*. This was also noted with earlier analysis of the data using PC-DFA.[6,20] No significant improvements in the clustering pattern is noticed on removal of the 5 or 10 most dominant peaks, or considering only the 20, 50 or 100 most dominant peaks alone, suggesting that the influence of the random variations are minimal, if any, for this dataset. An analysis of the loadings contributing to the discrimination of the species for the differently pre-processed spectra resulted in the identification of 37 peaks, which are listed in Table 2. The cases (pre-process criteria) where the peaks were identified are also listed in the table. Some of the peaks are present in more than one case, strengthening their contribution to the species-level discrimination. It is noteworthy that some of the identified peaks were also observed in the discrimination of *Enterococcus* spp. by PC-DFA, reported by us earlier.[20] This is especially true of peaks that were observed in more than three cases in the earlier investigation.

In order to understand the contribution of the peaks in these discriminations (Fig. 2), the original spectral data were revisited and a statistical analysis of the signal intensities at the identified peaks performed to obtain a univariate interpretation. The non-parametric Kruskal–Wallis statistical test[21] was used to make an assessment of the differences in the median signals for the replicate spectra for each species. For each of the identified peaks, at least one of the species had a median value that is statistically different from the rest. A pair-wise statistical comparison of the median values between the species for each of the identified peaks is also shown in Table 2. This comparison was done at 95% and 99% confidence levels. As can be seen, a statistically significant difference in the median signal intensities between *Enterococcus* spp. and the other species (except *K. oxytoca*) exists for almost all the identified peaks. *Enterococcus* spp. is the only identified Gram-positive organism amongst the isolates analysed (Table 1), with a cell surface composition distinctly different from the other bacteria analysed. Chemically, the differences are therefore likely to originate from molecular species present in the bacterial cell surfaces. The only exception to this rule is *K. oxytoca*, which on the basis of the ToF-SIMS signals can be differentiated from *Enterococcus* spp. in only four of the 37 identified peaks. *P. mirabilis* and *C. freundii* do not show a statistically significant difference in their median signals, at the identified peaks, nor does *E. coli* when compared to *C. freundii* or *K. oxytoca*. This is reflected in the pseudo-2D scores plots (Fig. 2) where the clustering between the above species cannot be easily discerned.

The signal intensities at selected *m*/*z* values are plotted in Fig. 3, as notched box-whisker plots, from which the isolate-level and species-level differences in the median signals at the representative *m*/*z* can be discerned. The lower and upper lines of the 'box' are the 25th and the 75th percentiles of the sample, the box limits indicating the interquartile range for each setting, with the horizontal bar representing the median for the isolate/species. The 'whiskers' (lines extending above and below the box) show the range (the maximum and minimum values), excluding outliers (values of >1.5 times the interquartile range). A plus sign outside the whiskers indicates the outlier in the data. The notches in the box are a graphic confidence interval about the median. A side-by-side comparison of two notched plots can be considered as a graphical equivalent of a *t*-test. Clear differences in the signal intensities for the different isolates and species can be noticed at the plotted *m*/*z* value. For example, at *m*/*z* 84, the signal intensity for the *P. mirabilis* isolates is the highest and that of *Enterococcus* spp. the least, whilst at *m*/*z* 134, the trend is reversed with respect to these two species. Similarly it is possible to note that *K. oxytoca* dominates at *m*/*z* 175 and *K. pneumoniae* at *m*/*z* 140. The differences noted at the isolate level (Fig. 3a–e) are fairly consistent to allow a species-wise grouping of the results (Fig. 3f–j).

It is thus possible to make univariate inferences, with respect to each species and their relationship to each other, which can be used to build models to discriminate between the isolates. It is also possible to build reliable and robust multivariate models that can be used for predictive purposes. A cross-validated CVA model built by combining the spectral information at the 37 identified *m*/*z* values is shown in Fig. 4. The model was constructed with *a priori* knowledge of the isolate-wise grouping. Even though an isolate-wise group structure was used, a species-wise clustering can be noticed for the training and test sets. For illustrative purposes, the 95% confidence zones for the location of the mean for each group are indicated by circles (blue circles, calculated based on the training set data). Similarly, a circle based on the difference between the maximum and minimum scores for the training set can be defined as the diameter about the median, for each species (black circles). As can be seen, a majority of the test set data (in red) fall within this zone, for three of the species (*Enterococcus* spp, *P. mirabilis*, *K. pneumoniae*) that show maximum discrimination, based on

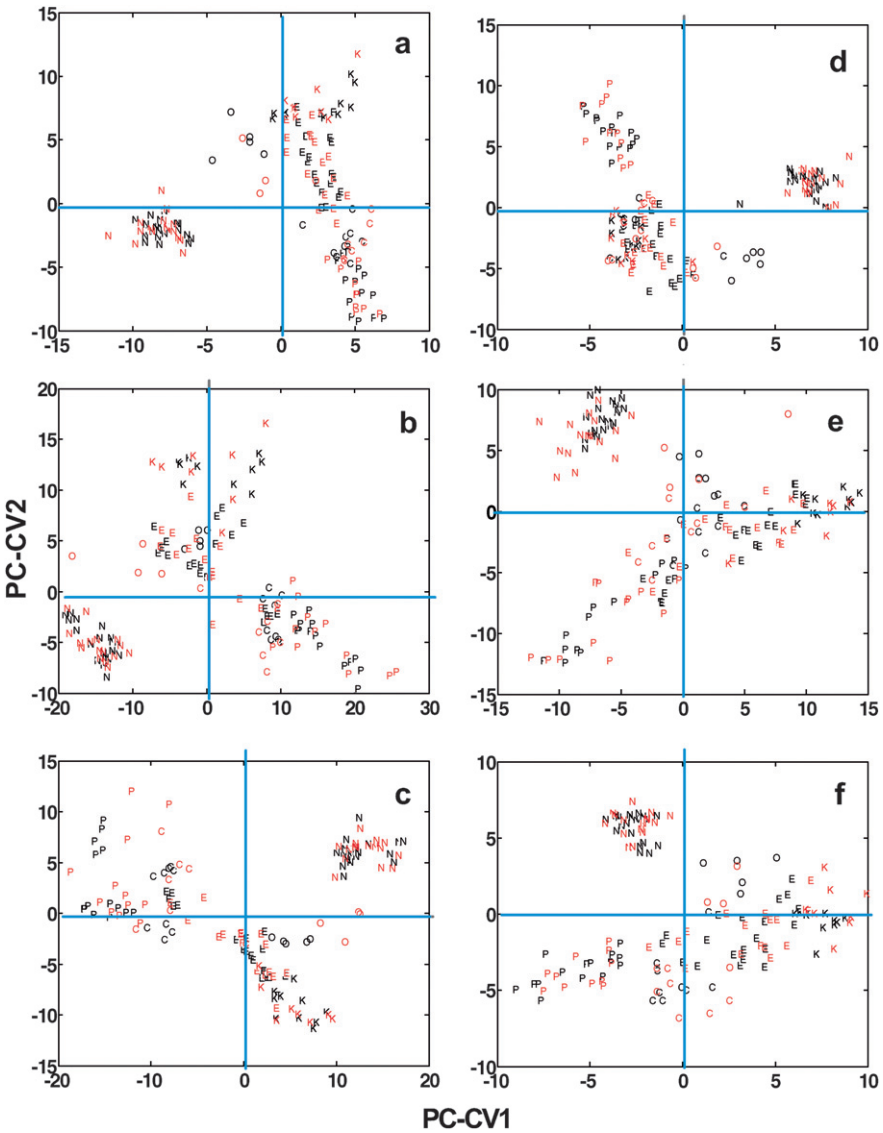

Fig. 2 PC-CVA scores biplots derived from positive ion ToF-SIMS spectra of bacterial isolates associated with urinary tract infections (*cf.* Table 1 for species notation), using different spectral pre-processing options. The peaks considered for analysis are: (a) all peaks in the mass range *m*/*z* 1–1000, (b) all but the 5 most prominent ones, (c) all but the 10 most prominent ones, (d) only the 20 most prominent ones, (e) only the 50 most prominent ones, and (f) only the 100 most prominent ones. The first two canonical variates scores (PC-CV1 and PC-CV2) are plotted against each other, for each case. Training set data are in black and test set data in red.

the identified *m*/*z* values. It is thus possible to define a set zone for discrimination that can be used for predictive purposes.

GA-CVA

Evolutionary algorithms, such as genetic algorithms (GAs) are suitable for mining data to uncover relationships, rules and predictions.[22–24] They can also be employed for feature selection to extract relevant information from spectral data.[14,15,17] Here, we have employed GA to extract variables of relevance in the discrimination of bacterial isolates responsible for causing urinary tract infections, based on their ToF-SIMS spectra. The results of the investigation indicated the spectral information relevant to the discrimination of the bacterial species that were not dissimilar to those obtained from PC-CVA. The power of this technique is that combinations of variables are observed which can account for the discrimination between all of the samples in the analysis.

The results of GA-CVA for the selection of pairs of discriminatory features are shown in Fig. 5. As can be seen from Fig. 5A, it is possible to plot a histogram showing the frequency with which masses are selected by the GA, and this is a useful guide in determining which variables have greatest importance individually for class separation. In addition, the database of pair-wise combinations of variables recovered from the analysis, when ranked by the fitness function score, gives an insight into which collective variables provide discrimination. Given the capability to infer the chemical provenance of the observed masses, this method provides an excellent opportunity for inferring new biochemical knowledge from such studies. Fig. 5B shows a biplot

Table 2 Peaks which are loaded in the discrimination of the species from each other, in the cases analysed. Cases analysed: 1, all peaks; 2, all but the 5 most dominant peaks; 3, all but the 10 most dominant peaks; 4, only the 20 most dominant peaks; 5, only the 50 most dominant peaks; 6, only the 100 most dominant peaks (*cf.* Table 1 for species notation)

		Statistically (Kruskal–Wallis) significant difference in median signal for pair-wise comparison ($\alpha = 0.05$ (x); $\alpha = 0.01$ (**X**))														
m/z	Cases observed	EN	EP	EC	EO	EK	NP	NC	NO	NK	PC	PO	PK	CO	CK	OK
42	PC(4)	**X**	**X**				**X**	**X**		**X**		**X**	**X**			
44	PC(4,5,6)	**X**	**X**				**X**	**X**		**X**		**X**	**X**	x		
51	PC(6)	**X**	**X**				**X**	**X**		x		**X**	**X**			
53	KE(2)	**X**	**X**				**X**	**X**		**X**		**X**				
54	PC(3)	**X**	**X**				**X**	**X**		**X**		**X**	**X**			
55	KE(2)	**X**	**X**				**X**	**X**		**X**		**X**	**X**			
56	PC(1,6)	**X**	**X**				**X**	**X**		**X**		**X**	**X**			
57	N(1), P(4)		**X**				**X**	x		x		x	**X**			
59	PC(6)	**X**	**X**				**X**	**X**		**X**		**X**				
60	PC(6)	**X**	**X**	**X**			**X**	**X**		**X**		**X**	**X**	**X**	x	
65	PC(3,6)	**X**	**X**				**X**	**X**		x		**X**	**X**			
67	PC(3)	**X**	**X**				**X**	**X**		**X**		**X**	**X**			
68	KE(2)	**X**	**X**				**X**	**X**		**X**		**X**		x		
69	PC(3,6)	**X**	**X**				**X**	**X**		**X**		**X**	**X**			
70	K(4,5,6)	**X**	**X**				**X**	**X**		**X**		**X**		x		
77	PC(3)	**X**	**X**				**X**	**X**		x		**X**	**X**			
80	PC(3)	**X**	**X**				**X**	**X**		**X**		**X**	**X**			
84	PC(1,2,4,5,6)	**X**	**X**				**X**	**X**		**X**		**X**	**X**	x		
85	PC(6)	**X**	**X**				**X**	**X**		**X**		**X**	**X**	x	**X**	**X**
87	PC(6)	**X**	**X**	x			**X**	**X**		**X**		x	**X**		x	
88	KE(2)	**X**	**X**				**X**	**X**		**X**		**X**		x		x
93	PC(3,6)	**X**	**X**				**X**	**X**		**X**	x	**X**	x			
104	PC(2)	**X**				x	**X**	x	x	**X**	x				**X**	x
110	KE(2,3), OE(2)	**X**					**X**	**X**		**X**						**X**
118	KE(2,3), OE(2)		x			**X**	**X**	**X**					**X**		**X**	**X**
120	N(1,3)	**X**					**X**	x	**X**	**X**					x	
129	PC(6)	**X**	**X**				**X**	**X**		**X**		**X**	**X**			
134	N(1,3,6)	**X**	**X**				**X**	**X**		**X**		**X**	**X**	**X**		
135	N(1,3,6)	**X**	x		x		**X**	**X**		**X**						
140	KE(6)		**X**		x	**X**	**X**	x	**X**	**X**			**X**		**X**	**X**
150	N(6)	**X**	**X**	**X**			**X**	**X**		**X**		**X**	**X**	**X**	**X**	
156	KE(5,6)				x	**X**	x		**X**				**X**		**X**	**X**
175	KE(2)	**X**			**X**		**X**	**X**		**X**		**X**		**X**		x
181	N(4), KE(5)	**X**	x				**X**	**X**		**X**		**X**	**X**	**X**	x	
197	N(1,5)	**X**					**X**	**X**		**X**		**X**				
468	KE(3)	**X**	**X**		**X**	x	**X**	**X**				**X**	**X**	**X**	**X**	
484	KE(3)	**X**	**X**		**X**	x	**X**	**X**				**X**	**X**	**X**	x	

of the normalised intensities for *m/z* 54 and *m/z* 200 (the combination of variables used most by the GA across the 200 independent runs to generate a two-variable subset) and illustrates excellent separation between all of the bacterial species analysed.

The spectral information at *m/z* 54 was previously identified from PC-CVA (Table 2), although *m/z* 200 represents discriminatory information that was not highlighted by the other data-mining methods. However, it is clear from Fig. 5B that separation between *E. coli* and *Klebsiella* spp., two species that are genetically and phenotypically related closely, is only achieved through the combination of these two variables Finally, an inspection of the spectral intensities (data not shown) showed that *Enterococcus* spp. showed maximum signal intensity at *m/z* 200 and minimum signal intensities at *m/z* 54, whilst *P. mirabilis* showed the exact opposite trend.

In this investigation we implemented a selective GA-CVA analysis in that we only sought to arrive at variable pairs that contribute to the discrimination. The idea was to see if GA-CVA throws up the same variables earlier identified by PC-CVA. Whilst *m/z* 54 was picked up by both PC-CVA and GA-CVA, *m/z* 200 was not observed with PC-CVA. PC-CVA clearly identified several other variables, which would have possibly been the case had we made a more elaborate search with GA-CVA. Nevertheless, it is interesting to note that the variables selected by GA-CVA are not the two most discriminating in PC-CVA. Whilst additional variables may be of value, it requires further validation to assess the usefulness of using more than one approach to analyse the data.

Postulated biological origins of identified signals

The likely biological origins of the discriminating peaks were investigated by performing an extensive search in metabolite databases and mass spectroscopic literature. Likely candidates for some of the peaks are listed in Table 3. Four metabolite databases, namely METLIN (http://metlin.scripps.edu/), KEGGS (http://www.genome.jp/kegg/pathway.html), MetaCyc (http://metacyc.org/) and NIST Chemistry WebBook (http://webbook.nist.gov/chemistry/) were queried for masses corresponding to the observed *m/z* − 1, assuming that the

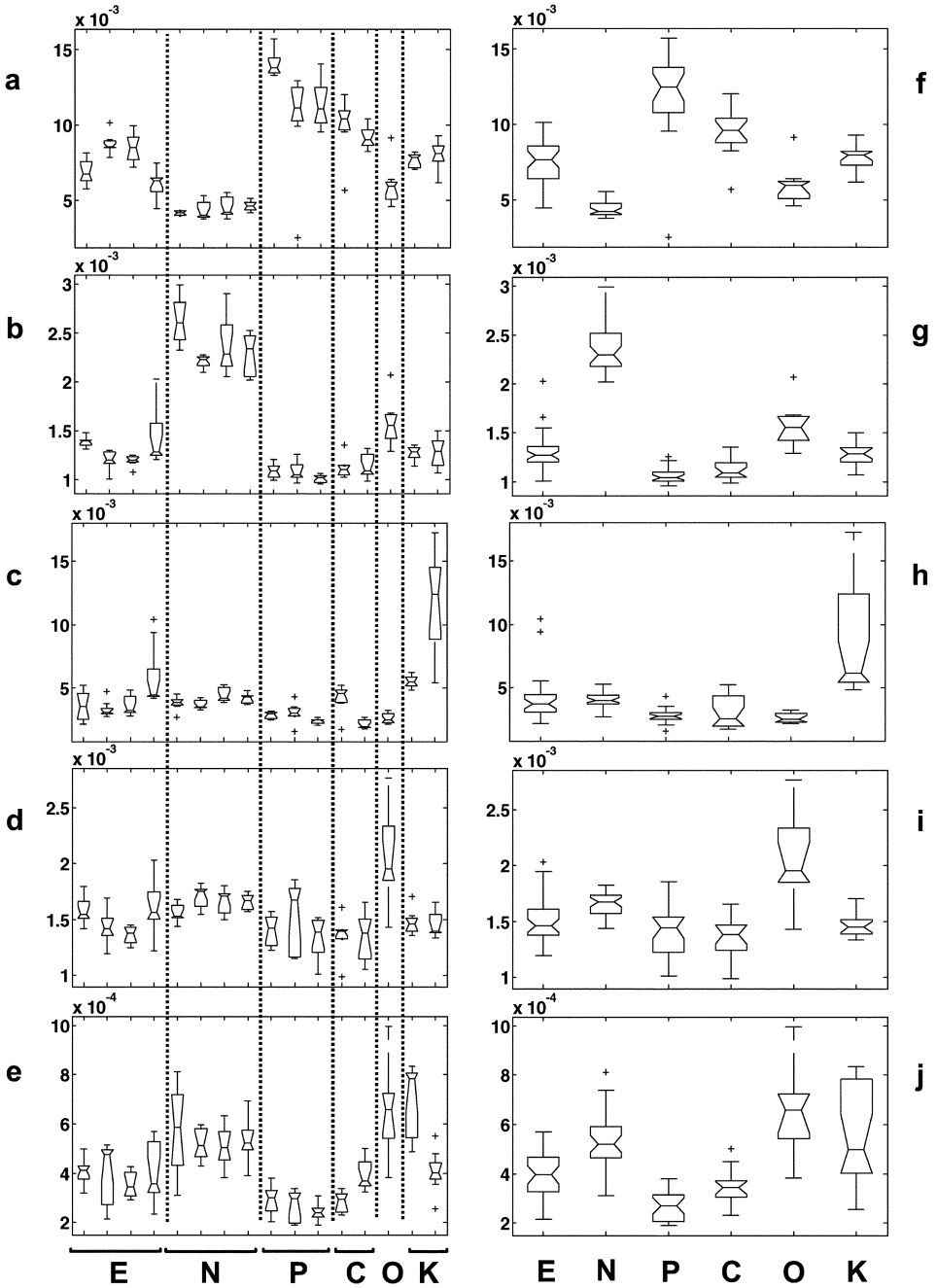

Fig. 3 Median signal intensities for the different bacterial isolates at different *m/z*: (a,f) *m/z* 84; (b,g) *m/z* 134; (c,h) *m/z* 140; (d,i) *m/z* 175; (e,j) *m/z* 468. The isolate-wise distribution of the signal intensities can be seen in panels a–e, and the species-wise distribution from panels f–j (*cf.* Table 1 for species notations, and text for interpretation of the box-whisker plots).

protonated pseudo-molecular ions are detected. Matches within a tolerance of ±0.5 mass units were considered. The results are summarised in Table 3. In addition, a literature search for potential fragments and adducts was also conducted.

Two potential candidates for the peak at *m/z* 54 are the benzoquinone fragment ($C_3H_2O^+$) and protonated butadiene or cyclobutene ($C_4H_6^+$). The latter are volatile organics known to be emitted by bacteria,[25] whilst the former are plasma membrane components that have been studied as chemotaxonomic markers.[26] This peak was observed to be the highest in *P. mirabilis* and the least in the Gram-positive *Enterococcus* spp., and hence contributes to the discrimination of *Enterococcus* from the rest. Whilst the Gram-negative species employed in this study have been listed to contain quinones,[26] the Gram-positive *Enterococcus* spp. are notably absent.

There are several candidates for the peak at *m/z* 70. Quite like *m/z* 54, this peak was also found to be the least in intensity the Gram-positive *Enterococcus* spp. The protonated pseudo-molecular ion of 1-pyrroline[27] is a candidate that has been reported to be emitted by *K. pneumoniae* and *C. freundii*.[28]

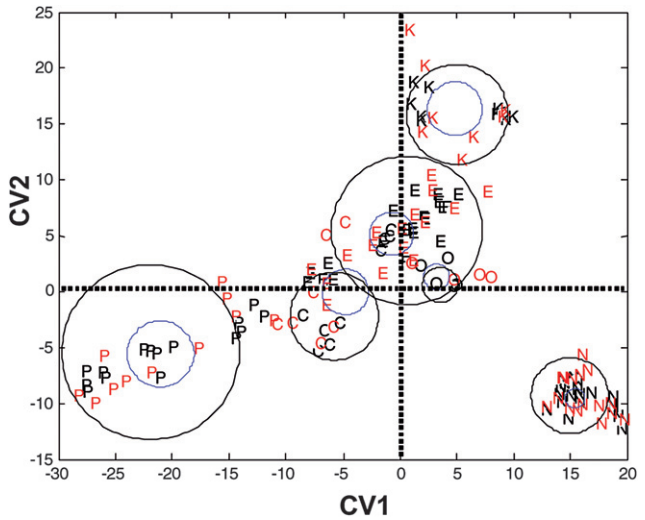

Fig. 4 Scores biplot for the discrimination of the bacterial species derived from CVA on the signal intensities at the identified peaks only. Training set data are in black and test set data in red. The 95% confidence zones (calculated on the training set) for the location of the mean are shown (for illustrative purposes) as blue circles, for each species. The black circle is about the median for each group, with the difference between the maximum and minimum values for each group, as its diameter (calculated on the training set data).

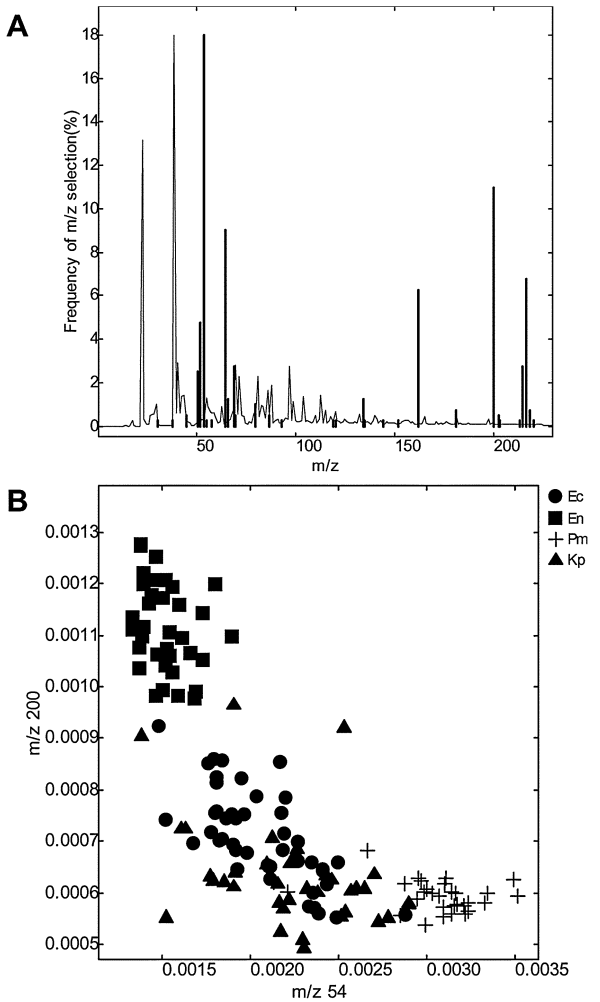

Fig. 5 (A) A histogram showing the frequency with which masses were selected by the GA over 200 independent runs. (B) A biplot of normalised intensities for *m/z* 54 *vs.* *m/z* 200, two variables selected by the GA. Ec = *E. coli*; En = *Enterococcus* spp.; Pm = *P. mirabilis*; Kp = *K. pneumoniae/oxytoca*.

Another candidate is a fragment of *N*-(3-hydroxy-octanoyl)homoserine lactone, with a loss of C_4H_6O. These are quorum sensing signals secreted by Gram-negative bacteria and have been reportedly detected in *Pseudomonas fluorescens*, where the fragment ion has been observed.[29] A third possible origin is 3-methyl-1-butyl acetate ($C_5H_{10}^+$ fragment) which has been reported in mass spectrometric investigations of volatile organics secreted by the Gram-negative *K. pneumoniae* and *P. mirabilis*, among others.[30] Yet another possibility is the fragment ion ($C_4H_8N^+$) from the amino acids such as proline, asparagine or ornithine, originating from proteins secreted by or present in the outer membranes of the Gram-negative bacteria and absent in the Gram-positive *Enterococcus* spp.

The peak at *m/z* 84 is found to be the highest in the *Proteus* samples and is also present in *K. pneumoniae* and *C. freundii*, but relatively low in *Enterococcus* spp. (Fig. 3). A possible biochemical origin for this signal is the oxonium ion from *O*-acetylated and *O*-carbamoylated *N*-acetylglucosamine, a collision induced dissociation fragment of which has been reportedly detected at *m/z* 84, in plant symbiotic bacteria.[31] Another candidate is the protonated pseudo-molecular ion of tetrahydropyridine, which like 1-pyrroline has also been detected in the head space of *K. pneumoniae* and *C. freundii* cultures.[28] The other likely origins are the amino acids glutamine, lysine or methylated proline.

Protonated ions from homoserine and threonine are likely matches to the peak at *m/z* 120. This peak is found to be higher in *Enterococcus* spp. Threonine is present in several of the signalling peptides secreted by enterococci and related species, for example, PrkC,[32] and bacteriocins.[33,34] It is also present in peptidoglycan hydrolases, such as AltA, secreted by *Enterococcus* spp.[35] Peptide signalling is specific to Gram-positive bacteria, whilst Gram-negative bacteria use homoserine lactones as signalling molecules. Lanthionine, a modified amino acid is known to produce a fragment at *m/z* 120.[36] These are known to be present in peptidoglycans of anaerobic bacteria[36] and have been reported to be present in enterococcal cytolysins.[37] Another candidate is the immonium ion of the amino acid phenylalanine.

The two peaks at *m/z* 134 and 135 are markedly higher in the Gram-positive *Enterococcus* spp. Primary protonated species that match *m/z* 134 are the amino acids aspartic acid and ornithine. Aspartate exists in the peptide bridges of the peptidoglycan in *Enterococcus* spp.[38,39] Peptidoglycan in Gram-positive bacteria are thicker than those in Gram-negatives. Although the deprotonated negative ion of aspartate is more readily detected in mass spectrometry, protonated positive ions have been reported in earlier investigations.[40] Aspartate has also been monitored by SIMS, based on its fragment $C_3H_6NO_2^+$ at *m/z* 88.[41] The signal intensity at *m/z* 88 is also found to be higher in *Enterococcus* spp. resulting in its discrimination from the other Gram-negative bacteria (Table 2). Ornithine is another possible candidate for the

Table 3 Database match of peaks, using exact mass information (with a tolerance of ±0.5 mass units), indicating possible biochemical origins of signals identified to contribute to the discrimination of the bacterial isolates associated with urinary tract infections. The listed candidates are potential sources of the signals, assuming them to be originating from protonated pseudo-molecular ions. The search was done on METLIN, KEGG, MetaCyc and NIST compound databases. Candidates relevant to the bacterial isolates are listed

m/z	Database match for protonated pseudo-molecular ion	Exact mass of matched metabolite	Protonated ion	Likely source
70	1-Pyrroline	69.0578	$C_4H_8N^+$	Volatile organics
84	Tetrahydropyridine	83.07		Volatile organics
120	Threonine	119.0582	$C_4H_{10}NO_3^+$	Signalling peptides
	Homoserine	119.0582	$C_4H_{10}NO_3^+$	
134	Aspartic acid	133.0375	$C_4H_8NO_4^+$	Peptidoglycan
	D-Ornithine	133.17	$C_5H_{14}N_2O_2^+$	Excreted metabolite
135	2-Methyl-3-(2-propenyl)pyrazine	134.08		Volatile organics
140	6-Hydroxynicotinic acid	139.0269	$C_6H_6NO_3^+$	Metabolite
	Carbamoyl phosphate	139.004	$CH_3NO_5P^+$	Metabolite
150	Methionine	149.0511	$C_5H_{12}NO_2S^+$	Membrane sugars
175	Arginine	174.1117	$C_6H_{15}N_4O_2^+$	Metabolite
	Indole-3-acetate	174.179	$C_{10}H_9NO_2^+$	Excreted metabolite
200	Clavulanate	199.0481	$C_8H_{10}NO_5^+$	Antibiotic

signal at *m/z* 134, the release of which by *Enterococcus* spp. has been documented.[42,43] Two potential candidates for the peak at *m/z* 135 could be matched: the protonated ion of 2-methyl-3-(2-propenyl)pyrazine and a fragment of caffeic acid sulfates. The former group is known to be released by bacteria, though not specifically by *Enterococcus* spp.,[28,44,45] and the latter are chlorogenic acid metabolites in coffee catabolised by colonic bacteria, such as *Enterococcus* spp.[46] It can be argued this is a result of the coffee intake by the UTI patients prior to sampling and the resultant metabolism by the human commensal organisms. However, both these candidates do not appear to be specific for *Enterococcus* spp. to sufficiently explain the higher signal intensity for these species.

The peak at *m/z* 140 is specific for *K. pneumoniae* and has two potential matches that correspond to protonated pseudo-molecular ions: 6-hydroxynicotinic acid and carbamoyl phosphate. The former has been implicated in the diagnosis of *Pseudomonas aeruginosa* induced UTI as compared with others.[47] However, this metabolite is reportedly not detected with *K. pneumoniae*. The latter is a common intermediate of arginine and pyrimidine biosynthesis,[48] but is not specific to *K. pneumoniae*. Two other candidates for the peak at *m/z* 140 that are more likely are the fragment ion of 3-*O*-carbamoyl-1-*O*-methyl *N*-acetylglucosamine (oxonium ion fragments have been detected in plant symbiotic bacteria[31]), and the sodium adduct of glycine betaine (betaines are known to serve as organic osmolytes, substances synthesised or taken up from the environment by cells for protection against osmotic stress, drought, high salinity or high temperature, and have been specifically observed in *K. pneumoniae*[49]).

A potential match for the peak at *m/z* 150 is the protonated methionine ion. Methionine is implicated in the synthesis of membrane associated sugars in *Enterococcus* spp.[50] It is possible that this metabolite is sequestered by *Enterococcus* for synthetic purposes, hence its higher signal and discrimination from the Gram-negative bacteria (Table 2). The peak at *m/z* 175 has a database match with the protonated ions of indole-3-acetate and arginine. Indole-3-acetic acid is a plant growth hormone that is also reportedly produced by bacteria, especially the Gram-negative enterobacteriaceae, such as *K. pneumoniae*,[51,52] in which the signal is higher. The protonated pseudo-molecular ion match for *m/z* 200 is clavulanate, an antibacterial used in combination with other antibiotics (such as amoxicillin) to overcome antibiotic resistance.[53] It is possible that the bacterial cells were metabolising this compound when they were harvested for analysis. Another possible candidate for this signal is a glycan oxonium ion, reported in the analysis of flagellin glycan from *Methanococcus voltae*.[54]

Conclusion

Multivariate data relevant to the discrimination of bacterial isolates based on their ToF-SIMS spectra have been analysed to seek explanation for the basis of the discrimination. A combined consideration of the spectral information in the absence of dominant peaks, in the presence of dominant peaks alone, and of all the peaks, enabled identification of the spectral information contributing to the PC-CVA discrimination of bacterial isolates. A statistical analysis of the spectral intensities at the identified *m/z* enabled an understanding of the discriminatory basis towards development of robust models that can be used for predictive purposes. The application of GAs for feature selection helped in substantiating the findings and derive additional information for explaining the discrimination. The biochemical origins of the identified peaks have been postulated based on database search and literature data. Many potential candidates that are likely to contribute to the identified signals have been listed. Future investigations would seek to verify these signals experimentally to obtain a robust basis for the discrimination for potential diagnostic use in translational medicine.

Acknowledgements

The authors are grateful to Dr David Broadhurst for help with data analysis and gratefully acknowledge the funding of the work by BBSRC, UK. R. J. and R. G. also thank the Home Office for financial support. S. V. wishes to acknowledge his current employers at the EPSRC funded ChELSI for the academic freedom to complete the work.

References

1 *ToF-SIMS: Surface Analysis by Mass Spectrometry*, ed. J. C. Vickerman and D. Briggs, IM Publications, Chichester and Surface Spectra, Manchester, UK, 2001.
2 D. Weibel, S. Wong, N. Lockyer, P. Blenkinsopp, R. Hill and J. C. Vickerman, *Anal. Chem.*, 2003, **75**, 1754–1764.
3 N. Winograd, *Anal. Chem.*, 2005, **77**, 142A–149A.
4 J. S. Fletcher, N. P. Lockyer, S. Vaidyanathan and J. C. Vickerman, *Anal. Chem.*, 2007, **79**, 2199–2206.
5 S. Vaidyanathan, J. S. Fletcher, R. Goodacre, N. P. Lockyer, J. Micklefield and J. C. Vickerman, *Anal. Chem.*, 2008, **80**, 1942–1951.
6 J. S. Fletcher, A. Henderson, R. M. Jarvis, N. P. Lockyer, J. C. Vickerman and R. Goodacre, *Appl. Surf. Sci.*, 2006, **252**, 6869–6874.
7 H. Jungnickel, E. A. Jones, N. P. Lockyer, S. G. Oliver, G. M. Stephens and J. C. Vickerman, *Anal. Chem.*, 2005, **77**, 1740–1745.
8 O. D. Sanni, M. S. Wagner, D. Briggs, D. G. Castner and J. C. Vickerman, *Surf. Interface Anal.*, 2002, **33**, 715–728.
9 M. S. Wagner and D. G. Castner, *Langmuir*, 2001, **17**, 4649–4660.
10 M. E. Wilkie, M. K. Almond and F. P. Marsh, *Br. Med. J.*, 1992, **305**, 1137–1141.
11 J. C. Evans, *Biometrics*, 1978, **34**, 170–170.
12 H. J. H. Macfie, C. S. Gutteridge and J. R. Norris, *J. Gen. Microbiol.*, 1978, **104**, 67–74.
13 W. J. Krzanowski, *Principles of Multivariate Analysis: A User's Perspective*, Oxford University Press, Oxford, 2000.
14 D. Broadhurst, R. Goodacre, A. Jones, J. J. Rowland and D. B. Kell, *Anal. Chim. Acta*, 1997, **348**, 71–86.
15 R. Cavill, H. C. Keun, E. Holmes, J. C. Lindon, J. K. Nicholson and T. M. Ebbels, *Bioinformatics*, 2009, **25**, 112–118.
16 R. M. Jarvis, D. Broadhurst, H. Johnson, N. M. O'Boyle and R. Goodacre, *Bioinformatics*, 2006, **22**, 2565–2566.
17 R. M. Jarvis and R. Goodacre, *Bioinformatics*, 2005, **21**, 860–868.
18 H. E. Johnson, D. Broadhurst, R. Goodacre and A. R. Smith, *Phytochemistry*, 2003, **62**, 919–928.
19 R. M. Jarvis and R. Goodacre, *Anal. Chem.*, 2004, **76**, 40–47.
20 S. Vaidyanathan, J. S. Fletcher, A. Henderson, N. P. Lockyer and J. C. Vickerman, *Appl. Surf. Sci.*, 2008, **255**, 1599–1602.
21 M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*, John Wiley& Sons, Inc., NY, 1999.
22 D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Kluwer Academic Publishers, Boston, MA, 1989.
23 J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, 1975.
24 J. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, Cambridge, MA, 1992.
25 M. O'Hara and C. A. Mayhew, *J. Breath Res.*, 2009, **3**, 027001.
26 M. D. Collins and D. Jones, *Microbiol. Rev.*, 1981, **45**, 316–354.
27 R. I. Zalewski, *Org. Mass Spectrom.*, 1981, **16**, 328–329.
28 D. C. Robacker and R. J. Bartelt, *J. Chem. Ecol.*, 1997, **23**, 2897–2915.
29 X. Cui, R. Harling, P. Mutch and D. Darling, *Eur. J. Plant Pathol.*, 2005, **111**, 297–308.
30 T. S. Wang, D. Smith and P. Spanel, *Int. J. Mass Spectrom.*, 2004, **233**, 245–251.
31 M. Treilhou, M. Ferro, C. Monteiro, V. Poinsot, S. Jabbouri, C. Kanony, D. Prome and J. C. Prome, *J. Am. Soc. Mass Spectrom.*, 2000, **11**, 301–311.
32 C. J. Kristich, C. L. Wells and G. M. Dunny, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 3508–3513.
33 B. Batdorj, M. Dalgalarrondo, Y. Choiset, J. Pedroche, F. Metro, H. Prevost, J. M. Chobert and T. Haertle, *J. Appl. Microbiol.*, 2006, **101**, 837–848.
34 T. Nilsen, I. F. Nes and H. Holo, *Appl. Environ. Microbiol.*, 2003, **69**, 2975–2984.
35 C. Eckert, M. Lecerf, L. Dubost, M. Arthur and S. Mesnage, *J. Bacteriol.*, 2006, **188**, 8513–8519.
36 S. Satyanarayana, J. S. Grossert, S. F. Lee and R. L. White, *Amino Acids*, 2001, **21**, 221–235.
37 W. Haas and M. S. Gilmore, *Med. Microbiol. Immunol.*, 1999, **187**, 183–190.
38 S. Bellais, M. Arthur, L. Dubost, J. E. Hugonnet, L. Gutmann, J. van Heijenoort, R. Legrand, J. P. Brouard, L. Rice and J. L. Mainardi, *J. Biol. Chem.*, 2006, **281**, 11586–11594.
39 G. J. Patti, S. J. Kim and J. Schaefer, *Biochemistry*, 2008, **47**, 8378–8385.
40 P. Chaimbault, K. Petritis, C. Elfakir and M. Dreux, *J. Chromatogr., A*, 1999, **855**, 191–202.
41 J. S. Robach, S. R. Stock and A. Veis, *J. Struct. Biol.*, 2006, **155**, 87–95.
42 C. Collar, A. F. Mascaros, J. A. Prieto and C. B. Debarber, *Cereal Chem.*, 1991, **68**, 66–72.
43 T. Yamamoto-Osaki, S. Kamiya, S. Sawamura, M. Kai and A. Ozawa, *J. Med. Microbiol.*, 1994, **40**, 179–187.
44 N. Camu, T. De Winter, K. Verbrugghe, I. Cleenwerck, P. Vandamme, J. S. Takrama, M. Vancanneyt and L. De Vuyst, *Appl. Environ. Microbiol.*, 2007, **73**, 1809–1824.
45 P. Hashim, J. Selamat, S. K. S. Muhammad and A. Ali, *J. Sci. Food Agric.*, 1998, **78**, 543–550.
46 A. Stalmach, W. Mullen, D. Barron, K. Uchida, T. Yokota, C. Cavin, H. Steiling, G. Williamson and A. Crozier, *Drug Metab. Dispos.*, 2009, **37**, 1749–1758.
47 A. Gupta, M. Dwivedi, G. A. N. Gowda, A. Ayyagari, A. A. Mahdi, M. Bhandari and C. L. Khetrapal, *NMR Biomed.*, 2005, **18**, 293–299.
48 H. Nicoloff, J. C. Hubert and F. Bringel, *Dairy Sci. Technol.*, 2001, **81**, 151–159.
49 D. Le Rudulier and L. Bouillard, *Appl. Environ. Microbiol.*, 1983, **46**, 152–159.
50 R. C. Wood, *J. Bacteriol.*, 1994, **176**, 6131–6133.
51 A. Karadeniz, S. F. Topcuoglu and S. Inan, *World J. Microbiol. Biotechnol.*, 2006, **22**, 1061–1064.
52 W. Zimmer, B. Hundeshagen and E. Niederau, *Can. J. Microbiol.*, 1994, **40**, 1072–1076.
53 R. Daza, J. Gutierrez and G. Piedrola, *Int. J. Antimicrob. Agents*, 2001, **18**, 211–215.
54 S. Voisin, R. S. Houliston, J. Kelly, J. R. Brisson, D. Watson, S. L. Bardy, K. F. Jarrell and S. M. Logan, *J. Biol. Chem.*, 2005, **280**, 16586–16593.

