Discrimination of bacteria using pyrolysis-gas chromatography-differential mobility spectrometry (Py-GC-DMS) and chemometrics

William Cheung,[ab] Yu Xu,[a] C. L. Paul Thomas[bc] and Royston Goodacre*[a]

Received 23rd July 2008, Accepted 25th November 2008
First published as an Advance Article on the web 16th December 2008
DOI: 10.1039/b812666f

Discrimination of bacteria was investigated using pyrolysis-gas chromatography-differential mobility spectrometry (Py-GC-DMS). Three strains belonging to the genus *Bacillus* were investigated and these included two strains of *Bacillus subtilis* and a single *Bacillus megaterium*. These were chosen so as to evaluate the possibility of bacterial strain discrimination using Py-GC-DMS. The instrument was constructed in-house and the long-term reproducibility of the instrument was evaluated over a period of 60 days using a Scotch whisky quality control. To assess the reproducibility further each bacterium was cultured six times and each culture was analysed in replicate to give three analytical replicates. The DMS data were generated in both positive and negative modes, and the data in each mode were analysed independently of each other. The Py-GC-DMS data were pre-processed *via* correlation optimised warping (COW) and asymmetric least square (ALS) to align the DMS chromatograms and to remove any unavoidable baseline shifts, prior to normalisation. Processed chromatograms were analysed using principal component analysis (PCA) followed by supervised learning methodology using partial least squares for discriminant analysis (PLS-DA). It was found that the separations between *B. subtilis* and *B. megaterium* can be readily observed by PCA; however, strain discrimination within the two *B. subtilis* was only possible using supervised learning. As multiple biological replicates were analysed an exhaustive splitting of the training and test sets was undertaken and this allowed correct classification rates (CCRs) to be assessed for the 3375 test sets. It was found that with PLS-DA the negative ion mode DMS data were more discriminatory than the positive mode data.

Introduction

In just about every area of microbiology the rapid identification of bacteria is desirable. For example, being able to identify a pathogen from a patient admitted into hospital would allow targeted antimicrobial therapy and accurate epidemiology studies to be conducted. Physicochemical methods are constantly being investigated and these have focussed mainly on vibrational spectroscopy- and mass spectrometry-based techniques.[1–5] For the latter technique a variety of sample introduction and ionisation methods have been employed and these included fast atom bombardment, pyrolysis, matrix assisted laser desorption ionisation and electrospray ionisation[1,6–9]

In general, MS is vacuum-based which has implications for portability of the instrument. By contrast, differential mobility spectrometry (DMS) is a gaseous phase ionic separation technique operating at ambient pressure, where the separation of ions is achieved by exploiting the difference in the ion mobilities between alternating high and low electric fields within the DMS drift cell.[10–16] During the last decade DMS has been primarily employed for detecting volatile organic compounds (VOCs). The low power consumption, compactness of DMS, coupled to ambient pressure operation with minimal maintenance makes it an attractive alternative to MS for VOC analysis where portability is required. Recently, DMS has been successfully coupled with GC for the analysis of human breath, bacterial odours and for jet fuel analysis.[11–17]

Eiceman *et al.*[11,12,14,15] have extensively studied the suitability of using GC-DMS as an alternative method to MS for bacterial characterisation where the non-volatile bacterial components are introduced into the GC using pyrolysis; a method that has been routinely coupled to MS.[1,7,18] In a series of experiments Eiceman's group has shown that biomarkers can be discovered which are specific to sporulated *Bacillus* and these included crotonic acid (a pyrolysis product of 3-hydroxybutyric acid) from *Bacillus megaterium*[11] and pyridine for *Bacillus subtilis*.[15] Pyridine is a pyrolysis product from dipicolinic acid which is found within bacterial spores and readily identified using Py-MS.[1,3] In addition, these authors have explored the pyrolysis conditions used[14] and found that these are consistent with those adopted for Py-GC-MS.[3] Finally, they have investigated the phenotypic changes that bacteria undergo when cultured at different temperatures, and have shown that the Py-GC-DMS signal is dependent on the organism's phenotype.[16] However, to date,

[a]*School of Chemistry, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester, UK M1 7DN. E-mail: roy.goodacre@manchester.ac.uk*
[b]*School of Chemical Engineering and Analytical Science, The University of Manchester, P.O. Box 88, Sackville St., Manchester, UK M60 1QD*
[c]*Department of Chemistry, Loughborough University, Loughborough, Leicestershire, UK LE11 3TU*

bacterial discrimination at the sub-species (*i.e.* strain) level has not been successfully reported.

Therefore the purpose of the present study was to investigate whether it is possible to distinguish between different bacteria at the strain level using Py-GC-DMS. Three strains from the genus *Bacillus* were selected, and these included two different strains from *B. subtilis* and *B. megaterium* as a closely related but phylogenetically different species. In order, to assess spectral reproducibility each strain was cultured six times and three analytical replicates recorded from each culture. Chemometric analysis was used to assess reproducibility and the ability to classify these bacteria.

Methods and materials

Culture and harvesting methodology

Three stains belonging to the genus *Bacillus* were studied; these included *B. subtilis* B0014, *B. subtilis* B1382 and *B. megaterium* B0010.[1,19] These strains were cultivated axenically on LabM blood agar base plates at 37 °C for 16 h. Growth was performed independently six times to generate six biological replicates per strain. This was because we are measuring the phenotype of the organism with Py-GC-DMS and since the phenotype = genotype + environment we need to control the latter else the former will be variable and the ability to characterise the different bacteria impaired, a phenomenon noted by all whole organism fingerprinting methods.[1,7,11] After incubation the vegetative biomass was carefully collected using a sterile plastic loop and suspended in 1 mL of physiological saline (0.9% NaCl in H_2O). The bacterial suspensions were centrifuged at 15 871 *g* for 3 min, the supernatants were then discarded, and the pellets were then resuspended in 1 mL of saline solution and centrifuged for 3 min; this process was repeated twice to remove any medium components from the agar. In the final resuspension the biomass concentration was adjusted to lie within an optical density of 2.5–2.7 AU at 600 nm (Biomate 5, Thermo Electron Corporation). As the analytical equipment was housed in a category 1 environment the resulting bacterial pellets were sterilised by autoclaving at 70 °C at 4 psi for 45 min. These were then stored at −80 °C until analysed.

In order to compensate for any systematic drift it is essential that the analysis was randomized by injections rather than by batches (biological specimens). Each bacterial injection was followed by a system blank to check for instrumental and/or environmental artefacts. In addition, after every three bacterial injections, a QC sample (see below) was run to assess the reproducibility and performance of the system. Sample sizes of 1.5 and 2 μL injections were chosen for the bacterial and QC samples respectively, this is equivalent to 3.9 and 2 mg of dry matter.

Quality control (QC) samples

In addition to the bacterial samples as described above, a number of QC samples were also analysed in order to monitor the performance of the system with respect to time. This was to determine whether or not there was any systematic drift within the instrumental response over the course of the experiment. The QC was an in-house whisky mix, 10 mL of a single malt whisky (Glen Moray Classic, Elgin, Spreyside, Scotland) was rotary evaporated down to dryness and left under a vacuum system (5×10^{-2} Torr) overnight to ensured the removal of all volatiles. After this the brown ‘slurry’ residue was then dissolved in 2 mL of ethanol (analytical grade; Fischer Scientific).

This QC sample had a complex matrix, which upon pyrolysis produced a complicated mixture of pyrolysates that can be used to assess instrumental drift; whisky has been used for this application before and yielded complicated Py-GC and Py-MS spectra.[20,21] The data obtained from these QC samples were treated using the same methodology that is described in the data analysis section below. Principal component analysis (PCA) was also performed on these QC samples to assess whether there was any systematic drift that correlated with injection time.

Instrumentation

As shown in Fig. 1 a CDS 5200 analytical pyrolysis unit (CDS Analytix Ltd., Unit 9 Seaview Workshop, Timber Rd, Horden, Peterlee, Durham, UK) was connected to a HP5890 gas chromatography unit (Mass Spec UK, Regal House, Highfield St., Oldham, UK) *via* a digitally controlled heated transfer line (an insulated silco steel coated inner core) into the front injector port A. The GC was also modified to enable a DMS unit (Sionex® Corp, 8-A Preston Court, Bedford, MA, USA) to be fitted directly to the existing flame ionisation detection housing using an annular heat pipe (30 cm × 0.635 cm inner diameter) maintained at 170 °C.

Pyrolysis unit. A Pt coil pyroprobe and quartz fire tube were used with the walls of the pyrolysis chamber maintained at 150 °C to minimise condensation of the pyrolysate. The pyrolysis chamber was continuously purged with He (5 mL min^{-1}) prior to sealing the chamber. The pyrolysis chamber was sealed and allowed to equilibrate for 60 s when the internal temperature reached 150 °C. Next the chamber temperature was increased from 150 to 300 °C at a rate of 20 °C ms^{-1}. After equilibration the temperature of the Pt coil pyroprobe was increased to 530 °C at a rate of 20 °C ms^{-1}. Pyrolysis took place for 5 s. During this sequence the Py-GC transfer line was maintained at 300 °C with a flow rate of 1 mL min^{-1} of He.

GC. The front injector port A was maintained at 300 °C with a split ratio of 10 : 1. The analytical column was a Restek RTx 5 Sil MS analytical column (30 m × 0.25 mm × 0.25 μm film thickness, with a stationary phase composition of 5% diphenyl/ 95% dimethylsiloxane). The He flow rate was maintained at approximately 1 mL min^{-1}. The analytical column was connected directly to the DMS unit through an annular heat pipe with the following temperature program: initial temperature: 60 °C (held for 2 min); then increased to 280 °C at a rate of 8 °C min^{-1}; the final temperature of 280 °C was held for 2 min.

DMS. The DMS unit was a Sionex SVAC-1 unit (Sionex® Corp) maintained at 100 °C with a N_2 flow rate of 270 mL min^{-1}, see Table 1 for the instrument parameters. Dispersion field programming methodology was used which increased the maximum dispersion field strength from 10 to 26 kV cm^{-1}. The

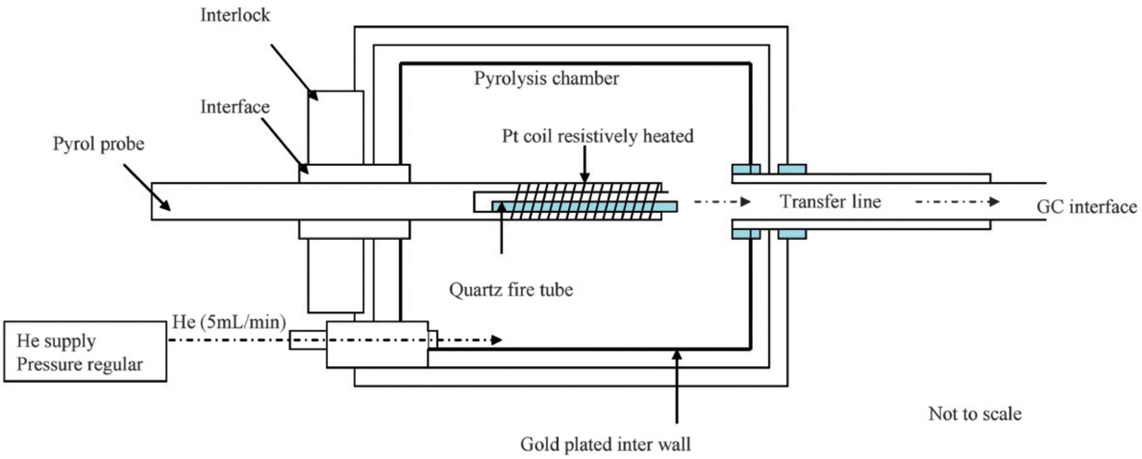

Fig. 1 Schematic diagram of the pyrol probe and pyrolysis chamber interface.

Table 1 DMS settings and parameters

Starting CV/V	−42
End CV/V	15
No. of steps	100
Step duration/mS	10
Step settle time/mS	3
Step to blank	1
Positive gain	High
Negative gain	High
RF step size/V	1
RF steps	0
CV step/V	0.56565

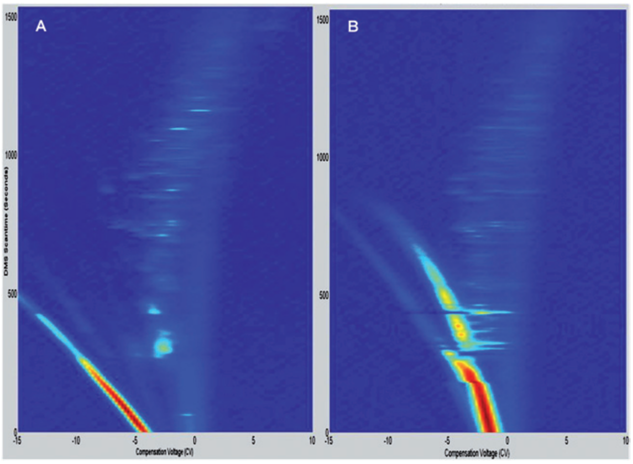

Fig. 2 DMS responses from *B. subtilis* B0014 in the (A) negative and (B) positive modes.

rationale for adopting this methodology has been reported recently.[17] Typical DMS profiles for *B. subtilis* are shown in Fig. 2.

Data analysis

The overall scheme used for data pre-processing and data analysis is detailed in Fig. 3. For each strain there were six biological replicates and each biological replicate was analysed three times (machine/analytical replicates) creating 18 samples per strain. Therefore the data matrix analysed consisted of three strains, containing 54 Py-GC-DMS spectra in total.

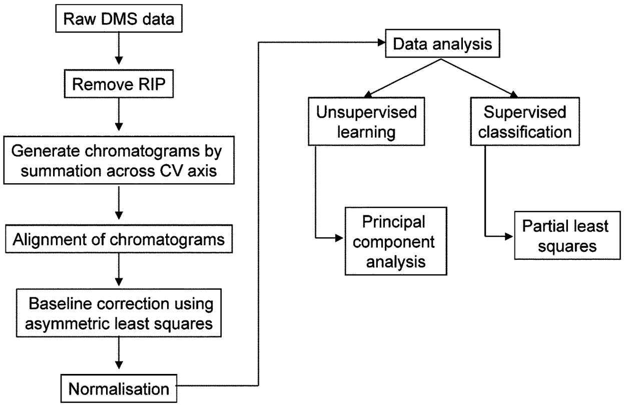
Raw DMS data
Remove RIP
Generate chromatograms by summation across CV axis
Alignment of chromatograms
Baseline correction using asymmetric least squares
Normalisation
Data analysis
Unsupervised learning
Supervised classification
Principal component analysis
Partial least squares

Fig. 3 Flow diagram summarising the data pre-processing (LHS) and data analysis (RHS) methodology.

Signal processing. The unprocessed DMS data were generated in Microsoft Excel Worksheet format. The resultant files were catalogued into biological, and then analytical replicates, with the positive and negative modes data-processed separately. Preliminary visual inspection of the data enabled dominating and potentially non-reproducible features to be identified and we chose to exclude the reactant ion peak (RIP) from the all Py-GC-DMS data. The original DMS matrix sizes were as follows: the voltage compensation was from −15 to +10 CV and DMS scan time was 0–1534 s^{-1}. After RIP removal the cropped matrix for negative mode included the GC eluent from 9 to 22.5 min retention time and the DMS ranged from −6 to +6 CV with a 500–1350 s^{-1} scan time. Whilst for the positive mode the GC included the retention times from 10 to 23.3 min and the DMS was from −5 to +5 CV and the scan time was 600–1400 s^{-1}. The data were then summed across the compensation voltage (CV) axis producing two DMS chromatograms; for the negative mode this was summed from −5 to +5 CV, and −6 to +6 CV for the positive mode data. For consistency all Py-GC-DMS data were pre-processed in an identical manner.

The chromatograms were aligned by using correlation optimised warping (COW[22]) to correct the drifting of the peaks between different Py-GC-DMS runs. The segment size was set to 10% of the number of the data points within each chromatogram, and the slack variable was set to 10% of the segment size (*e.g.* as the Py-GC-DMS chromatogram contained 500 data points the segment size was therefore set to 50 and the slack variable was set to 5).

After COW alignment the baseline was corrected by using asymmetric least square (ALS[23]), an adaptive baseline estimation algorithm. Finally, the chromatograms were normalised by using min–max normalisation, whereby each chromatogram was divided by the absolute difference between the maximal and minimal intensities.

Pattern recognition. MATLAB version 2007a (MathWorks, Nantwich, USA) was employed for the data analysis. Principal component analysis (PCA)[24–27] was performed on the processed chromatograms (detailed in Fig. 3) containing 500 data points. PCA is an unsupervised vector space transform method, often used to reduce high dimensional data sets to lower dimensions for analysis, *e.g.* modelling or visualisation. The variance of the data set was captured by a few latent variables called principal components (PCs). The variance captured by each PC is in a nested fashion. The first PC always captures the largest variance of the whole data set and the second PC captures the largest variance of the residues (the unexplained variance of the previous PC) and so on for the later PCs. The data set was column centred before the PCA. The scores of the first few PCs were plotted against each other in order to visualise any natural clustering trends within the data sets. After PCA, the next step was to perform a supervised classification upon the data set to investigate if it was possible to discriminate between the three bacterial strains. Unlike PCA, which is an unsupervised method, supervised classification attempts to build a predictive model based on a subset of samples with known origin (training set). If there are sufficient chemical differences between the bacteria that are detected by Py-GC-DMS the model should be able to predict the class membership of unknown samples. The accuracy of such prediction was assessed by using an independent data set (test set) which was not used during the training stage. In this study, we used partial least squares-discriminant analysis (PLS-DA)[24,26–29] as the supervised classifier.

PLS-DA is a commonly used supervised classification method which is based on a well known regression model called partial least squares or project to latent structure (PLS). Similar to PCA, PLS is also a latent variable-based model but in a supervised manner. Instead of finding a smaller set of latent variables capturing as many variations as possible, PLS finds a linear model describing some predicted variables (*e.g.* concentration levels, class membership, *etc.*) in terms of a set of other observable variables, *e.g.* the Py-GC-DMS data in our case. Similar to PCA, the observable variables were also 'compressed' into a few latent variables, called PLS components, and the fundamental relations between the predicted variables and observable variables were established based on the PLS components. PLS can model one predicted variable, which is usually called a PLS1 model, while it can also model several predicted variables simultaneously, which is usually called a PLS2 model. Although it was originally designed as a regression model, various applications as well as some theoretical studies have proved that it can be a very effective classification model.[24,26–29] For two-class separation problems, both PLS1 and PLS2 models can be employed, while for multiple class classification problems, in general a PLS2 model should be used (although it is also possible to combine multiple PLS1 models). In this study, there are three classes to be separated; hence PLS2 modelling was used. The class memberships of the samples were represented by a Y matrix with three columns and each column corresponds to one different class. Binary encoding was used such that class 1 would be encoded as 1, 0, 0, class 2 as 0, 1, 0, and class 3 as 0, 0, 1.

The PLS model was built on the training set and the number of significant PLS components were determined by using k-fold cross-validation while k is the number of biological replicates (see below). This model was then applied to the test set and the class membership of each sample was determined by using the procedure described by Wu *et al.*[30] For each sample, the predicted vector of $\boldsymbol{y}$ was calculated. The sample was assigned to the class for which the predicted value is the only one higher than 0.5. For instance, if the predicted $\boldsymbol{y}$ is [0.1, 0.9, 0.2], the sample is assigned to class 2. When the prediction is, for example, [0.1, 0.9, 0.8] or [0.1, 0.4, 0.2], the sample was regarded as a misclassified sample, *i.e.* the class membership of that sample cannot be confidently determined.

The training set was created by selecting 66.7% (2/3) of the samples from the original data set for training and the remaining 1/3 were used as the test set. During training set selection each of three machine replicates per biological replicate were considered as a 'whole' rather than as independent samples, the sample selection must account for this else one is merely measuring the reproducibility of the analytical instrument rather than the biological differences. In our case, there were six biological replicates for each strain, so four were used for training and two were used for testing. Since the number of different combinations of selecting four samples out of six is 15, this splitting of training and test set procedure had been repeated exhaustively $15^3 = 3375$ times. The predictions were averaged to give an estimation of the expected accuracy of the classification model. The reason for such exhaustive resampling is to minimise the influence of selecting samples for training or testing on the final outcome of the classification models and avoid the chance that seemingly good results were in fact due to a 'lucky' set of samples being chosen as the training set and another 'lucky' set of samples being chosen as the test set. If the separation is genuine, it should be relatively insensitive to the splitting of training and test set and the results of these models should be similar to each other. There might be some 'fortunate' or 'unfortunate' occasions which yield extremely good or poor results, but such circumstances should be rare.

Results and discussion

In order to show that Py-GC-DMS could be a useful analytical approach for the characterisation and identification of bacteria we designed a robust experiment where multiple biological replicates of three bacteria (*viz.* two strains of *B. subtilis* and one

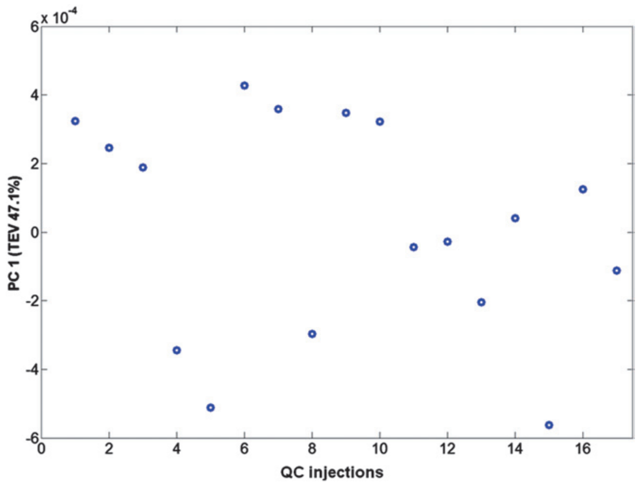

Fig. 4 Plot of injection number for the QC standards *versus* the first principal component score (total explained variance = 47.1%). This shows that there was no systemic time-related trend within the Py-DMS data.

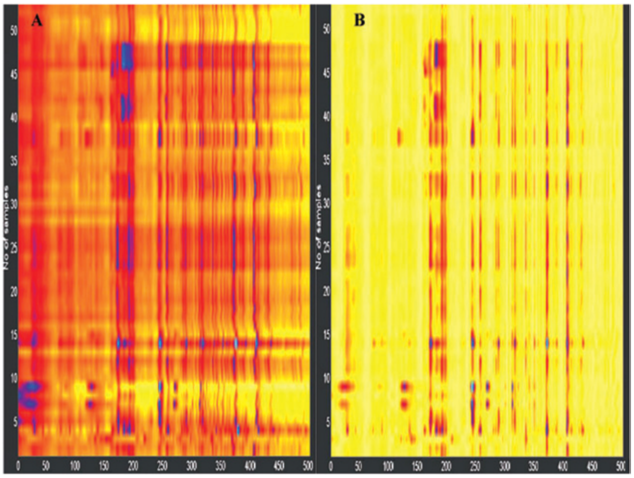

Fig. 5 (A) Chromatograms before alignment, and (B) after alignment using the COW algorithm. In this two-step alignment process the (i) global alignment uses one replicate from each of the three machine replicates for each sample ($n = 18$) aligned to the global mean, followed by (ii) a local second alignment where the remaining two chromatograms were aligned to the above alignment for each of the 18 groups. In the figure the x-axis represents the chromatograms (DMS scantime) and the y-axis the 54 chromatograms (3 strains × 6 biological replicates × 3 machine replicates).

B. megaterium) were analysed in triplicate using Py-GC-DMS over a period of 60 days. During this analysis analytical blanks were collected and used to assess artefacts in terms of carryover of pyrolysate from one sample to another and inspection of these blanks showed that there were no 'memory effects' observed (data not shown). In addition, after three bacterial analyses a QC sample from dried whisky was analysed so as to assess reproducibility during data acquisition. Following the pre-processing as detailed in Fig. 3 and above the QC samples were analysed using PCA. A plot of the injection number (which is relative to time) *versus* the first PC (which contained 47.1% total explained variance) is shown in Fig. 4 where no correlation with respect to sample injection time is seen, and the same was true for other PCs as well (data not shown). This result suggests that as no systematic trends can be observed in PCA space that the instrument was running in a stable and reproducible manner. This gave us confidence that the analysis of the bacteria by the same system would not be overtly influenced by any analytical artefacts.

Following data collection from the bacterial samples the Py-GC-DMS data were processed as detailed above (and Fig. 3). Initially the Py-GC-DMS data (Fig. 2 for examples) were summed across the CV axis after RIP removal and aligned using a two-step COW alignment. The data before and after the results of the COW alignment are shown in Fig. 5, where it can be clearly seen that the peak drifting is significantly reduced after the alignment, indicating the utility of running this step in the pre-processing sequence of procedures.

Once aligned the data were baseline corrected using ALS and were normalised to min–max per chromatogram. These data were then ready for chemometric analysis. The initial stage of the data analysis strategy was to use unsupervised exploratory data analysis and PCA was employed to discover any natural groups within the data and is also a useful algorithm for discovering any outliers. The results of the PCA are shown in Fig. 6 where it can be seen that both negative and positive modes generated similar trends and there was an obvious separation between *B. megaterium* and the two strains of *B. subtilis* in both PC1 and PC2 for both positive mode and negative mode. However, no obvious separation between the two strains of *B. subtilis* (B0014 *vs.* B1382) can be observed, also for both ionisation modes. Whilst all three bacteria belong to the *Bacillus* genus, in terms of phylogenetics *B. megaterium* is different from *B. subtilis* at the 16S rDNA sequence level,[31] and these genotypic differences are manifest in the organism's phenotype which is what we are measuring using Py-GC-DMS. By contrast the two strains of *B. subtilis* are very closely related and so using this unsupervised learning algorithm cannot be separated; indeed a finding we have previously observed with Raman spectroscopy[31,32] but not electrospray ionisation mass spectrometry,[19] presumably because the latter generates analyte specific information. This therefore presents a true test for Py-GC-DMS. The question arises as to whether there are any significant differences which lie in less obvious variance of the Py-GC-DMS data that can be discovered using supervised learning.

Therefore we chose to use a supervised classifier PLS-DA which was programmed as described above. As it was trained with training data – *i.e.* data from Py-GC-DMS from known bacterial origin – it is important that the classification model is tested independently and we used a third of the data as the independent hold out test set. These training and test sets were selected from the Py-GC-DMS profiles in an exhaustive fashion. As detailed above, this resulted in 3375 splits of the data. The 3375 predictions for the test set were then averaged to give an estimation of the expected accuracy of the classification model.

In most cases, three PLS components appeared to be optimal *via* k-fold cross-validation. The average correct classification rates (CCRs) on the test set along with their standard deviation as well as the minimum and maximum of all 3375 runs are shown in Table 2. In addition to the CCR, the averaged confusion matrices are also shown in Table 3. The results suggested that the

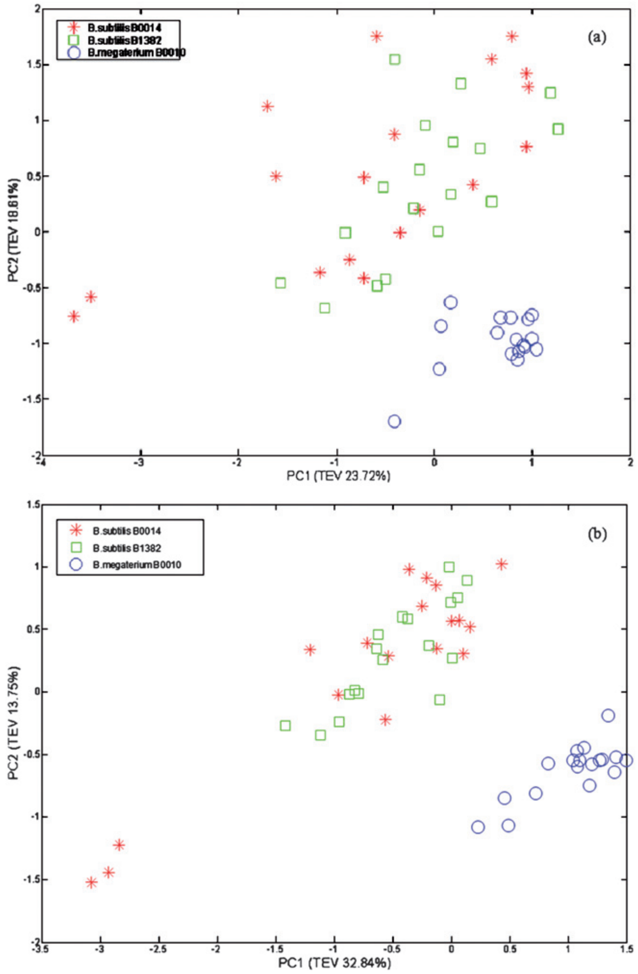

Fig. 6 PCA scores plot of PC1 *vs.* PC2: (a) negative mode; (b) positive mode. See inset legend for which symbols refer to *B. subtilis* strains B0014 and B1382 and *B. megaterium* B0010.

Table 2 Summary of the correct classification rate (CCR) for PLS-DA

	Mean CCR	Standard deviation	Minimum CCR	Maxmum CCR
PLS-DA				
Positive mode	91.90%	0.073%	61.00%	100.00%
Negative mode	97.39%	0.036%	77.78%	100.00%

positive prediction accuracy of *B. megaterium* is 100% and so consistently better than those of the two strains of *B. subtilis*. For the data obtained using the negative mode data from the DMS, the error rate of the prediction of *B. megaterium* is also always 0 for all 3375 runs, and for positive mode data *B. subtilis* B1382 is rarely mis-identified as *B. megaterium*.

Inspection of the confusion matrix indicates that by using PLS-DA it is possible to separate the two strains of *B. subtilis* with very high accuracy. The prediction accuracies are higher in the negative mode where 98.83% on average for *B. subtilis* B0014 and 95.50% average for *B. subtilis* B1382 are predicted. This is particularly encouraging as in PCA (Fig. 6) there were no obvious separations between these two *B. subtilis* strains; by contrast PLS-DA suggests that there is enough information in the Py-GC-DMS data to allow supervised classification methods to separate these two strains.

Table 3 Confusion matrices for the supervised classification algorithm

	Predicted strain		
Actual strain	*B. subtilis* B0014	*B. subtilis* B1382	*B. megaterium* B0010
PLS-DA on positive mode data			
B. subtilis B0014	85.83%	14.17%	0.00
B. subtilis B1382	9.67%	89.83%	0.50%
B. megaterium B0010	0.00	0.00	100.00%
PLS-DA on negative mode data			
B. subtilis B0014	98.83%	1.17%	0.00
B. subtilis B1382	4.50%	95.50%	0.00
B. megaterium B0010	0.00	0.00	100.00%

It is also evident in Table 2 and Table 3 that the negative mode data yielded better prediction accuracy of the bacterial class than those of the positive mode data.

Conclusions

It has been successfully demonstrated that discrimination at the bacterial strain level is possible using Py-GC-DMS as the analytical technique, but only when coupled with supervised learning methods. The separations between different species can be readily observed by PCA; however, strain discrimination requires more powerful chemometric methods using supervised classifiers such as PLS-DA.

In addition, the data analysis suggests that the positive mode data contained less strain-specific information than the negative mode data in DMS (Table 2 and Table 3) and this is likely to be due to the difference in ionisation chemistry with negative reactant ions, and such a finding is consistent with the work carried out by other researchers.[11] However, rather than disregard the chemical information obtained from the positive mode this should be viewed as an additional orthogonal response to that of the negative mode, and in the future we shall investigate this further as the positive mode in DMS also offers the added option of derivatization to be employed to obtain greater chemical information.

In conclusion, we believe that Py-GC-DMS presents itself as a complementary analytical approach for the rapid characterisation of bacteria, and this is the first study to apply advanced chemometrics for the separation of bacteria at the sub-species level and will be investigated further.

Acknowledgements

Y. X. and R. G. are grateful to the European Commission's BIOTRACER Integrated Project (www.biotracer.org) for financial support. We are grateful to Sionex Corporation for funding W. C. and technical support and maintenance of the DMS SVAC-1 unit, and Dr Ian Fleet for his guidance and support.

References

1 R. Goodacre, B. Shann, R. J. Gilbert, E. M. Timmins, A. C. McGovern, B. K. Alsberg, D. B. Kell and N. A. Logan, *Anal. Chem.*, 2000, **72**, 119–127.

2 D. I. Ellis, D. Broadhurst, D. B. Kell, J. J. Rowland and R. Goodacre, *Appl. Environ. Microbiol.*, 2002, **68**, 2822–2828.

3 A. P. Snyder, J. P. Dworzanski, A. Tripathi, W. M. Maswaden and C. H. Wick, *Anal. Chem.*, 2004, **76**, 6492–6499.

4 J. P. Dworzanski, A. Tripathi, A. P. Snyder, W. M. Maswadeh and C. H. Wick, *J. Anal. Appl. Pyrolysis*, 2005, **73**, 29–38.

5 R. Jarvis, S. Clarke and R. Goodacre, *Topics in Applied Physics*, 2006, **103**, 397–408.

6 B. Michael, R. S. Bordoli, R. D. Sedgwick and A. N. Tyler, *J. Chem. Soc., Chem. Commun.*, 1981, 325–327.

7 H. L. C. Meuzelaar, J. Haverkamp and F. D. Hileman, *Pyrolysis mass spectrometry of recent and fossil biomaterials*, Elsevier, Amsterdam, 1982.

8 S. Vaidyanathan, D. B. Kell and R. Goodacre, *J. Am. Soc. Mass Spectrom.*, 2002, **13**, 118–128.

9 S. Vaidyanathan, D. Jones, D. I. Broadhurst, J. Ellis, T. Jenkins, W. B. Dunn, A. Hayes, N. Burton, S. G. Oliver, D. B. Kell and R. Goodacre, *Metabolomics*, 2005, **1**, 243–250.

10 I. A. Buryakov, E. V. Krylov, A. L. Makas, E. G. Nazarov, V. V. Pervukhin and U. K. H. Rasulev, *J. Anal. Chem.*, 1993, **48**, 114–121.

11 S. Schmidt, F. K. Tadjimukhamedov, K. M. Douglas, S. Prasad, G. B. Smith and G. A. Eiceman, *Anal. Chem.*, 2004, **76**, 5208–5217.

12 P. H. Rearden and P. Harrington, *Hyphenated separations*, LabPlus international, February/March, 2006.

13 P. Rearden, P. B. Harrington, J. J. Karnes and C. E. Bunker, *Anal. Chem.*, 2007, **79**, 1485–1491.

14 S. Schmidt, F. K. Tadjimukhamedov, K. M. Douglas, S. Prasad, G. B. Smith and G. A. Eiceman, *J. Anal. Appl. Pyrolysis*, 2006, **76**, 161–168.

15 S. Prasad, H. Schmidt, P. Lampen, M. Wang, R. Guth, J. Rao, G. B. Smith and G. A. Eiceman, *Analyst*, 2006, **131**, 1216–1225.

16 S. Prasad, K. M. Pierce, H. Schmidt, J. V. Rao, R. Robert Guth, S. Bader, R. E. Synovec, G. B. Smith and G. A. Eiceman, *Analyst*, 2007, **132**, 1031–1039.

17 M. Basanta, D. Singha, S. Fowler, I. Wilson, R. Dennis and C. L. P. Thomas, *J. Chromatogr., A*, 2007, **1173**, 129–138.

18 E. M. Timmins and R. Goodacre, in *Identification of Microorganisms by Mass Spectrometry*, ed. C. L. Wilkins, J. O. Lay and J. D. Winefordner, John Wiley & Sons, New Jersey, 2006, pp. 319–343.

19 S. Vaidyanathan, J. J. Rowland, D. B. Kell and R. Goodacre, *Anal. Chem.*, 2001, **73**, 4134–4144.

20 K. J. G. Reid, J. S. Swan and C. S. Gutteridge, *J. Anal. Appl. Pyrolysis*, 1993, **25**, 49–62.

21 R. I. Aylott, A. H. Clyne, A. P. Fox and D. A. Walker, *Analyst*, 1994, **119**, 1741–1746.

22 N. P. Vest Nielsen, J. M. Carstensen and J. Smedegarrd, *J. Chromatogr., A*, 1998, **805**, 17–35.

23 H. F. M. Boelens, R. J. Djikstra, P. H. C. Eilers, F. Fitzpatrick and J. A. Westerhuis, *J. Chromatogr., A*, 2004, **1057**, 21–30.

24 H. Wold, in *Multivariate Analysis*, ed. P. R. Krishnaiah, Academic Press, New York, 1966, pp. 391–420.

25 R. C. Beavis, S. M. Colby, R. Goodacre, P. Harrington, J. P. Reilly, S. Sokolow and C. W. Wilerson, in *Encyclopedia of Analytical Chemistry*, ed. R. A. Meyers, John Wiley & Sons, Chichester, 2000, pp. 11558–11597.

26 R. G. Brereton, *Chemometrics Data Analysis for the Laboratory and Chemical Plants*, Wiley, Chichester, 2002.

27 Y. Xu, *Chemometrics pattern recognition with applications to Genetics and Metabolomics data*, PhD Thesis, University of Bristol, 2006.

28 B. K. Alsberg, D. B. Kell and R. Goodacre, *Anal. Chem.*, 1998, **70**, 4126–4413.

29 S. J. Dixon, Y. Xu, R. G. Brereton, H. A. Soini, M. V. Novotny, E. Oberzaucher, K. Grammer and D. J. Penn, *Chemom. Intell. Lab. Syst.*, 2007, **87**, 161–172.

30 W. Wu, Q. Guo, D. Jouan-Rimbau and D. L. Massart, *Chemom. Intell. Lab. Syst.*, 1999, **45**, 39–53.

31 E. C. López-Díez and R. Goodacre, *Anal. Chem.*, 2004, **76**, 585–591.

32 R. M. Jarvis, A. Brooker and R. Goodacre, *Faraday Discuss.*, 2006, **132**, 281–292.