Detection of the Dipicolinic Acid Biomarker in *Bacillus* Spores Using Curie-Point Pyrolysis Mass Spectrometry and Fourier Transform Infrared Spectroscopy

Royston Goodacre,*[,†] Beverley Shann,[†] Richard J. Gilbert,[†] Éadaoin M. Timmins,[†] Aoife C. McGovern,[†] Bjørn K. Alsberg,[†,‡] Douglas B. Kell,[†] and Niall A. Logan[§]

Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion, SY23 3DD, Wales, U.K., Department of Computer Sciences, University of Wales, Aberystwyth, Ceredigion, SY23 3DB, Wales, U.K., and School of Biological and Biomedical Sciences, Glasgow Caledonian University, Cowcaddens Road, Glasgow, G4 0BA, Scotland, U.K.

Thirty-six strains of aerobic endospore-forming bacteria confirmed by polyphasic taxonomic methods to belong to *Bacillus amyloliquefaciens*, *Bacillus cereus*, *Bacillus licheniformis*, *Bacillus megaterium*, *Bacillus subtilis* (including *Bacillus niger* and *Bacillus globigii*), *Bacillus sphaericus*, and *Brevi laterosporus* were grown axenically on nutrient agar, and vegetative and sporulated biomasses were analyzed by Curie-point pyrolysis mass spectrometry (PyMS) and diffuse reflectance−absorbance Fourier-transform infrared spectroscopy (FT-IR). Chemometric methods based on rule induction and genetic programming were used to determine the physiological state (vegetative cells or spores) correctly, and these methods produced mathematical rules which could be simply interpreted in biochemical terms. For PyMS it was found that *m*/*z* 105 was characteristic and is a pyridine ketonium ion ($C_6H_3ON^+$) obtained from the pyrolysis of dipicolinic acid (pyridine-2,6-dicarboxylic acid; DPA), a substance found in spores but not in vegetative cells; this was confirmed using pyrolysis-gas chromatography/mass spectrometry. In addition, a pyridine ring vibration at 1447−1439 cm^{-1} from DPA was found to be highly characteristic of spores in FT-IR analysis. Thus, although the original data sets recorded hundreds of spectral variables from whole cells simultaneously, a simple biomarker can be used for the rapid and unequivocal detection of spores of these organisms.

The genus *Bacillus* consists of Gram-positive bacteria, which can respond to slowed growth or starvation by initiating the process of sporulation. These organisms as well as exhibiting morphological variation of vegetative cells can undergo differentiation to distinct resting bodies or spores, and after a period of time, these spores can germinate to produce a single vegetative cell. As sporulation proceeds, there are striking morphological and biochemical changes in the developing spore. It becomes encased in two novel layers, a peptidoglycan layer (the spore cortex) and a number of layers of spore coats that contain proteins unique to spores.[1] The spore also accumulates a substantial deposit (5−14% of dry weight) of pyridine-2,6-dicarboxylic acid (dipicolinic acid; DPA), which is unique to spores, as well as a large amount of divalent cations.[2]

Members of the genus *Bacillus* are widely distributed in soil, water, and air, and because their spores are so resistant their control is of considerable importance in the food processing industry and in the preparation of sterile products.[3] In addition, the rapid identification of *Bacillus anthracis* spores is of importance because of its potential use as a biological warfare agent.[4] Therefore, there is a need for a generic characterization system that can be used to carry out the large-scale and rapid detection of bacterial spores.

Although spores can be observed by simple microscopy, this approach does not lead itself to automation, and hence, analytical techniques are being employed to detect bacterial spores much more rapidly and with more specificity. The theoretical fluorescence of dipicolinic acid and its anion have been studied,[5] while terbium dipicolinate photoluminescence has been used to detect the presence of sporulating bacteria.[6] The vibrational spectroscopic methods of UV resonance Raman spectroscopy[7] and Fourier transform infrared spectroscopy[8] have also been used to differenti-

* Corresponding author: (telephone) +44 (0)1970 621947; (telefax) +44 (0)1970 621947; (e-mail) rrg@aber.ac.uk.

[†] Institute of Biological Sciences, University of Wales.

[‡] Department of Computer Sciences, University of Wales.

[§] Glasgow Caledonian University.

[1] Buchanan, C. E.; Henriques, A. O.; Piggot, P. J. In *Bacterial Cell Wall*; Hakenbeck, J.-M. A. R., Ed.; Elsevier Science Publishers: New York, 1994; pp 167−186.

[2] Murrell, W. G. In *The Bacterial Spore*; Gould, G. W., Hurst, A., Eds.; Academic Press: London, 1969; pp 215−273.

[3] Doyle, M. P.; Beuchat, L. R.; Montville, T. J. *Food Microbiology: Fundamentals and Frontiers*; Amercian Society of Microbiology Press: Washington, DC, 1997; p 768.

[4] Barnaby, W. *The Plague Makers: The Secret World of Biolgoical Warfare*; Vision Paperbacks: London, 1997.

[5] Hameka, H. F.; Jensen, J. O.; Jensen, J. L.; Merrow, C. N.; Vlahacos, C. P. *J. Mol. Struct. (THEOCHEM)* **1996**, *365*, 131−141.

[6] Pellegrino, P. M.; Fell, N. F.; Rosen, D. L.; Gillespie, J. B. *Anal. Chem.* **1998**, *70*, 1755−1760.

[7] Ghiamati, E.; Manoharan, R.; Nelson, W. H.; Sperry, J. F. *Appl. Spectrosc.* **1992**, *46*, 357−364.

[8] Helm, D.; Naumann, D. *FEMS Microbiol. Lett.* **1995**, *126*, 75−80.

10.1021/ac990661i CCC: $19.00
Published on Web 11/19/1999

ate between spores and vegetative bacteria. Flow cytometry[9] has also been explored for the rapid detection of spores. Finally, a number of workers have investigated methods based on mass spectrometric analyses. However, bacterial spores are nonvolatile and so to introduce these and their component chemical species to a mass spectrometer they first have to be pyrolyzed. The volatile fragments (or pyrolysate) from bacterial spores have then been measured following electron ionization using a quadrupole mass analyzer,[10] a tandem mass spectrometer,[11] or gas chromatography/ion mobility spectrometry;[12] however, none of these has yet gained acceptance as the analytical method of choice for spore detection in a military context.

The above studies used only a handful of *Bacillus* species, and some merely a single strain. Most studies for detecting *B. anthracis* spores that have been deliberately released for military or terrorist actions have concentrated on *Bacillus subtilis*, and in particular its pigmented variant "*Bacillus globigii*" (*B. subtilis* var. *niger*), which is perhaps surprising since phylogenetic analysis of their 16S rRNA sequences shows that *B. subtilis* and *B. globigii* are distinct from *B. anthracis*,[13] which itself is phylogenetically indistinguishable from *Bacillus cereus*. Therefore the present study investigates a wide variety of genetically distinct bacteria (36 representatives of seven mesophilic *Bacillus* species; *B. cereus* was represented by five strains and was used to draw analogy to *B. anthracis*; the latter was omitted from this study on grounds of safety).

The rapid, fully automated, analytical methods that were employed in this study included Curie-point pyrolysis mass spectrometry (PyMS), and diffuse reflectance−absorbance Fourier transform infrared (FT-IR) spectroscopy, in the mid-infrared range. PyMS and FT-IR are physicochemical methods that measure predominantly the bond strengths of molecules and the vibrations of bonds within functional groups, respectively.[14,15] Therefore they give quantitative information about the total biochemical composition of a sample. However, the interpretation of these multidimensional spectra, or what are known as hyperspectral data,[16] has conventionally been by the application of "unsupervised" pattern recognition methods such as principal component (PCA), discriminant function (DFA), and hierarchical cluster (HCA) analyses. With "unsupervised learning" methods of this sort, the relevant multivariate algorithms seek "clusters" in the data, thereby allowing the investigator to group objects together on the basis of their perceived similarity.[17] This process is often subjective because it relies upon the interpretation of complicated scatterplots and dendrograms. More recently, various related but much more powerful approaches, most often referred to within the framework of chemometrics, have been applied to the "supervised" analysis of hyperspectral data;[18] arguably the most significant of these is the application of intelligent systems based on artificial neural networks (ANNs).[19,20]

Although ANNs are excellent at identifying unknown bacterial isolates to the correct genera and species, the information in terms of which masses in the mass spectrum or wavenumbers in the infrared spectrum are important is not *readily* available. The use of ANNs therefore is often perceived as a "black box" approach to modeling spectra and so has limited use for the deconvolution of hyperspectral data in (bio)chemical terms. Therefore, in this study a number of rule induction methods[21] and methods involving evolutionary computation[22,23] were employed to aid in the deconvolution of these hyperspectra.

EXPERIMENTAL SECTION

Bacteria and Cultivation. Identities of the 36 *Bacillus* species studied were confirmed by use of a polyphasic approach using conventional biochemical (API tests) and nucleic acid technologies.[24] Full details are shown in Table 1.

Vegetative cells were collected by incubating the 36 bacteria on Lab M blood agar base plates (without blood) at 37 °C for 10 h. Spores were prepared by incubation on Lab M blood agar base plates + 5 mg/L^{-1} $MnSO_4$ at 30 °C for 7 days.[25] After incubation, the biomass was collected in physiological saline (0.9% NaCl). These bacterial slurries contained $\sim 10^9$ cells/mL and were stored at −20 °C until analysis. To assess the level of sporulation, all 72 cultures were examined microscopically. The vegetative cells contained no spores, while the sporulated biomass comprised >80% spores; no attempt was made to remove any cell debris or vegetative cells from the spore preparations.

Pyrolysis Mass Spectrometry. Five-microliter aliquots of the above bacterial samples were evenly applied to iron−nickel foils to give a thin uniform surface coating. Prior to pyrolysis, the samples were oven-dried at 50 °C for 30 min. Each sample was analyzed in triplicate. The pyrolysis mass spectrometer used for this study was a Horizon Instrument PYMS-200X (Horizon Instruments, Heathfield U.K.). For full operational procedures, see ref 26. The sample tube carrying the foil was heated, prior to pyrolysis, at 100 °C for 5 s. Curie-point pyrolysis was at 530 °C for 3 s, with a temperature rise time of 0.5 s. The data from PyMS were collected over the m/z (mass) range 51−200 (Figure 1). These conditions were used for all experiments. Data were

[9] Davey, H. M.; Kell, D. B. *Microbiol. Rev.* **1996**, *60*, 641−696.

[10] Beverly, M. B.; Basile, F.; Voorhees, K. J.; Hadfield, T. L. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 455−458.

[11] Voorhees, K. J.; Deluca, S. J.; Noguerola, A. *J. Anal. Appl. Pyrolysis* **1992**, *24*, 1−21.

[12] Dworzanski, J. P.; McClennen, W. H.; Cole, P. A.; Thornton, S. N.; Meuzelaar, H. L. C.; Arnold, N. S.; Synder, A. P. *Field Anal. Chem. Technol.* **1997**, *1*, 295−305.

[13] Goodacre, R.; Shann, B.; Gilbert, R. J.; Timmins, É. M.; McGovern, A. C.; Alsberg, B. K.; Logan, N. A.; Kell, D. B. In Proc. 1997 ERDEC Scientific Conference on Chemical and Biological Defense Research, Aberdeen Proving Ground, ERDEC-SP-063; 1998; pp 257−265.

[14] Meuzelaar, H. L. C.; Haverkamp, J.; Hileman, F. D. *Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials*; Elsevier: Amsterdam, 1982.

[15] Griffiths, P. R.; de Haseth, J. A. *Fourier transform infrared spectrometry*; John Wiley: New York, 1986.

[16] Goetz, A. F. H.; Vane, G.; Solomon, J.; Rock, B. N. *Science* **1985**, *228*, 1147−1153.

[17] Everitt, B. S. *Cluster Analysis*; Edward Arnold: London, 1993.

[18] Goodacre, R.; Timmins, É. M.; Burton, R.; Kaderbhai, N.; Woodward, A. M.; Kell, D. B.; Rooney, P. J. *Microbiology* **1998**, *144*, 1157−1170.

[19] Bishop, C. M. *Neural networks for pattern recognition*; Clarendon Press: Oxford, 1995.

[20] Despagne, F.; Massart, D. L. *Analyst* **1998**, *123*, 157R−178R.

[21] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and regression trees*; Wadsworth, Inc.: Pacific Grove, CA, 1984.

[22] Koza, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*; MIT Press: Cambridge, MA, 1992.

[23] Bäck, T.; Fogel, D. B.; Michalewicz, Z. *Handbook of Evolutionary Computation*; IOP Publishing/Oxford University Press: Oxford, U.K., 1997.

[24] Heyndrickx, M.; Vandemeulebroecke, K.; Scheldeman, P.; Kersters, K.; DeVos, P.; Logan, N. A.; Aziz, A. M.; Ali, N.; Berkeley, R. C. W. *Int. J. Syst. Bacteriol.* **1996**, *46*, 988−1003.

[25] Shute, L. A.; Gutteridge, C. S.; Norris, J. R.; Berkeley, R. C. W. *J. Gen. Microbiol.* **1984**, *130*, 343−355.

[26] Goodacre, R.; Neal, M. J.; Kell, D. B. *Anal. Chem.* **1994**, *66*, 1070−1085.

Table 1. *Bacillus* Strains Studied

species[a]	strain no.	training/ test set	identifier on plots
B. sphaericus	7134^T	test	1a
	B0219	test	1b
	B0408	test	1c
	B0769	training	1d
	B1147	training	1e
B. subtilis	B0014^T	test	2a
	B0044	test	2b
(var. *B. niger*)	B0098	test	2c
(var. *B. niger*)	B0099	test	2d
	B0410	test	2e
	B0501	training	2f
(var. *B. globigii*)	B1382	training	2g
B. licheniformis	B0242	test	3a
	B0252^T	test	3b
	B0755	test	3c
	B1081	training	3d
	B1379	training	3e
Br. laterosporus	B0043	test	4a
	B0115	training	4b
	B0262	training	4c
	B0616	test	4d
B. cereus	B0002^T	training	5a
	B0550	test	5b
	B0702	test	5c
	B0712	test	5d
	B0851	training	5e
B. amyloliquefaciens	B0168	test	6a
	B0175	test	6b
	B0177^T	test	6c
	B0251	training	6d
	B0620	training	6e
B. megaterium	B0010^T	training	7a
	B0056	training	7b
	B0057	test	7c
	B0076	test	7d
	B0621	test	7e

[a] var., variants of *B. subtilis* that produce pigmented spores. [b] Superscript "T" indicates type strain.

normalized as a percentage of the total ion count to remove the influence of sample size per se.

Diffuse Reflectance−Absorbance Fourier Transform Infrared Spectroscopy. Ten-microliter aliquots of the above bacterial suspensions were evenly applied onto a sand-blasted aluminum plate. Prior to analysis, the samples were oven-dried at 50 °C for 30 min. Samples were run in triplicate. The FT-IR instrument used was the Bruker IFS28 FT-IR spectrometer (Bruker Spectrospin Ltd., Coventry, U.K.) equipped with a mercury−cadmium−telluride (MCT) detector cooled with liquid N_2. The aluminum plate was then loaded onto the motorized stage of a reflectance TLC accessory.[27] The IBM-compatible PC used to control the IFS28, was also programmed (using OPUS version 2.1 software running under IBM O/S2 Warp provided by the manufacturers) to collect spectra over the wavenumber range 4000−600 cm^{-1}. Spectra were acquired at a rate of 20 s^{-1}. The spectral resolution used was 4 cm^{-1}. To improve the signal-to-noise ratio, 256 spectra were coadded and averaged. Each sample was thus represented by a spectrum containing 882 points, and spectra were displayed in terms of absorbance as calculated from the reflectance−absorbance spectra using the Opus software (which is based on the Kubelka−Munk theory[15]) (Figure 2). These conditions were used for all experiments. To minimize problems arising from baseline shifts the following procedure was implemented: (i) the spectra were first normalized so that the smallest absorbance was set to 0 and the highest to +1 for each spectrum; (ii) these spectra then had their baselines removed using a fast Fourier transform

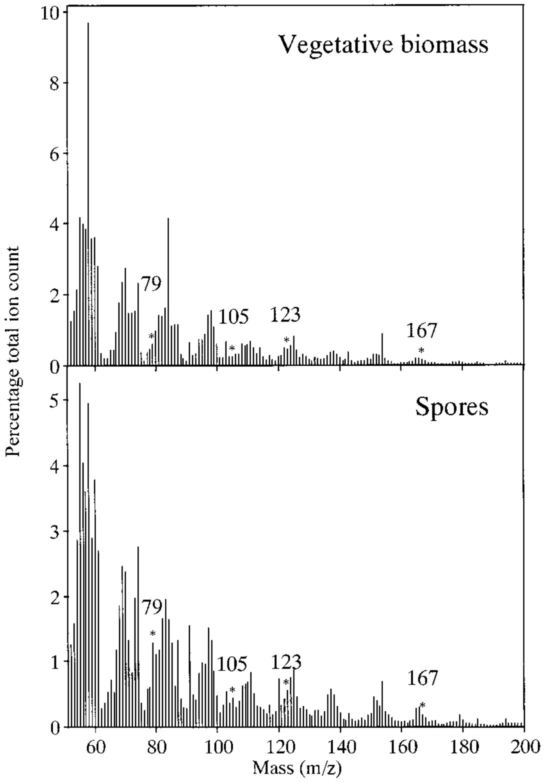
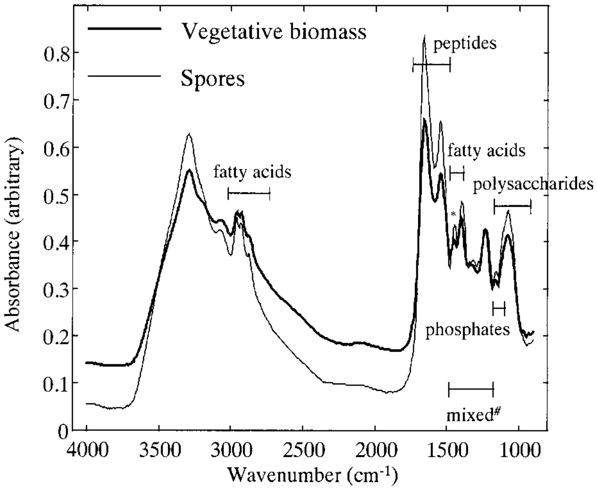


Figure 1. Curie-point pyrolysis-MS spectra of *Bacillus subtilis* B0014^T in its vegetative and sporulated states. *, *m/z* 79, 105, 123, and 167 peaks from dipicolinic acid labeled (see Figure 4 and text for details).



├──┤ = various 'spectral windows'

#proteins, fatty acids, phosphate rich

Figure 2. FT-IR spectra of *B. subtilis* B0014^T in its vegetative and sporulated states. * is the label for the pyridine ring vibrations between 1470 and 1435 cm^{-1} from dipicolinic acid (see Figure 6 and see text for details). Also shown are the 'spectral biochemical windows' for FT-IR spectra of bacteria based on studies by Naumann and colleagues.[42]

[27] Timmins, É. M.; Howell, S. A.; Alsberg, B. K.; Noble, W. C.; Goodacre, R. *J. Clin. Microbiol.* **1998**, *36*, 367−374.

baseline routine developed in-house by Dr. A. M. Woodward. Briefly, these IR spectra were transformed from the spectral wavelength domain into the Fourier domain spectra (FDS),[28] where the signal is concentrated into the low-delay region and the baseline information into the very-low-delay domain of FDS (while any noise is located in the high-delay domain). To reduce any spectral distortions, seen when a simple windowing approach was employed, an inverted Gaussian filter was deconvolved from the FDS spectra to taper off the low-frequency bins. These FDS were then inversely transformed back to the wavenumber domain.

Cluster Analysis. The initial stage involved the reduction of the multidimensional PyMS and FT-IR data by PCA.[29] PCA is a well-known technique for reducing the dimensionality of multivariate data while preserving most of the variance, and Matlab was employed to perform PCA according to the NIPALS algorithm.[30] DFA (also known as canonical variates analysis; CVA) then discriminated between groups on the basis of the retained principal components (PCs) and the a priori knowledge of which spectra were replicates, and thus this process does not bias the analysis in any way.[31] Finally, the Euclidean distance between a priori group centers in DFA space was used to construct a similarity measure, with the Gower general similarity coefficient S_G,[32] and these distance measures were then processed by an agglomerative clustering algorithm to construct a dendrogram.[31] These methods were implemented using Matlab version 5.0.0.4069 (The Math Works, Inc., Natick, MA), which runs under Microsoft Windows NT on an IBM-compatible PC.

Common Supervised Analysis Methods. When the desired responses (targets) associated with each of the inputs (spectra) are known, then the system may be supervised. The goal of supervised learning is to find a model that will correctly associate the inputs with the targets; this is usually achieved by minimizing the error between the target and the model's response (output).[33]

The input data sets for all supervised learning methods contained the full PyMS spectra (150 m/z intensities) and the full FT-IR spectra (882 wavenumber absorbances), and these were partitioned into training and test sets. The training set contained the replicate spectra from 14 of the *Bacillus* species, as both vegetative cells and sporulated biomass, chosen randomly (84 spectra), and the test set comprised the 132 remaining spectra (details are given in Table 1). The output data were binary encoded such that vegetative biomass was coded as 0 and spores as 1.

Two artificial neural network-based methods, viz. standard back-propagation multilayer perceptrons (MLPs)[19,34] and radial basis functions (RBFs),[35,36] were used. Both ANNs were carried out with a user-friendly, neural network simulation program, NeuFrame version 3,0,0,0 (Neural Computer Sciences, Totton, Southampton, Hants, U.K.), which runs under Microsoft Windows NT on an IBM-compatible personal computer.

The multivariate linear regression method of partial least squares (PLS)[37] was also exploited. All PLS analyses were carried out using an in-house program, developed by Dr. Alun Jones[38] following the pseudocode given in ref 37, which runs under Microsoft Windows NT on an IBM-compatible PC.

Rule Induction. Rule induction methods produce *if-then-else* decision trees, which are often very easy to interpret. These decision trees are found by attempting to partition the space of sample objects into regions of single class memberships. The data set is *recursively split* into smaller subsets where each subset contains objects belonging to as few different classes as possible.[21]

There are two main strategies for finding the best object partitioning which in general can be described as *univariate* and *multivariate* rule induction. In *univariate rule induction* a single variable x_i at each recursion step is found that gives rise to the purest subsets. A univariate rule is interpreted as a decision plane *parallel* to the original coordinate axes. In *multivariate rule induction* methods, however, each recursion step uses a linear (or nonlinear) combination of the original variables and thus the decision planes can point in any direction in the multidimensional space.

Three different methods of rule induction were used: (1) univariate (classification and regression trees), (2) (CART) multivariate OC1 rule induction, and (3) multivariate Breiman rule induction. All rule induction techniques were performed using the OC1 program[39] (Department of Computer Science, Johns Hopkins University, Baltimore, MD). All programs run under Windows NT 4.0.

Genetic Programming. A genetic algorithm (GA) is an optimization method based on the principles of Darwinian selection.[23,40] A population of individuals, each representing the parameters of the problem to be optimized as a string of numbers or binary digits, undergoes a process analogous to evolution in order to derive an optimal or near-optimal solution. The parameters stored by each individual are used to assign it a *fitness*, a single numerical value indicating how well the solution using that set of parameters performs. New individuals are generated from members of the current population by processes analogous to asexual and sexual reproduction.

In the steady-state version of the GA used herein, "asexual reproduction", or *mutation*, is performed by randomly selecting a parent with a probability proportional to its fitness and then randomly changing one (or occasionally more) of the parameters it encodes. The new individual then replaces a less fit member of the population, if one exists. "Sexual reproduction", or *crossover*, is achieved by randomly selecting two parents with a probability proportional to fitness, generating two new individuals by copying

[28] Mattu, M. J.; Small, G. W. *Anal. Chem.* **1995**, *67*, 2269−2278.

[29] Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.

[30] Wold, H. In *Multivariate Analysis*; Krishnaiah, K. R., Ed.; Academic Press: New York, 1966; pp 391−420.

[31] Manly, B. F. J. *Multivariate Statistical Methods: A Primer*; Chapman & Hall: London, 1994.

[32] Gower, J. C. *Biometrika* **1966**, *53*, 325−338.

[33] Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; DeJong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part A*; Elsevier: Amsterdam, 1997.

[34] Rumelhart, D. E.; McClelland, J. L.; The PDP Research Group *Parallel Distributed Processing, Experiments in the Microstructure of Cognition*; MIT Press: Cambridge, MA, 1986.

[35] Moody, J.; Darken, C. J. *Neural Comput.* **1989**, *1*, 281−294.

[36] Saha, A.; Keller, J. D. In *Advances in Neural Information Processing Sytems*; Touretzky, D., Ed.; Morgan Kaufmann Publishers: San Mateo, CA, 1990; Vol. 2, pp 482−489.

[37] Martens, H.; Næs, T. *Multivariate Calibration*; John Wiley: Chichester, U.K., 1989.

[38] Jones, A.; Shaw, A. D.; Salter, G. J.; Bianchi, G.; Kell, D. B. In *Lipid Analysis of Oils and Fats*; Hamilton, R. J., Ed.; Chapman & Hall: London, 1998; pp 317−376.

[39] Murthy, S. K.; Kasif, S.; Salzberg, S. *J. Artif. Intell. Res.* **1994**, *2*, 1−32.

[40] Holland, J. H. *Adaption in natural and artifcial systems*, 2 ed.; MIT Press: Cambridge, MA, 1992.

parameters from one parent, and switching to the other parent after a randomly selected point. The two new individuals then replace less fit members of the population as before. The above procedure is repeated, with the overall fitness of the population improving at each generation, until an acceptably fit individual is produced.

A genetic program (GP) is an application of the GA approach to derive mathematical equations, logical rules, or program functions automatically.[22,41] Rather than representing the solution to the problem as a string of parameters, as in a conventional GA, a GP uses a tree structure. The leaves of the tree, or *terminal*s, represent input variables or numerical constants. Their values are passed to *nodes*, at the junctions of branches in the tree, which perform some numerical or program operation before passing on the result further toward the root of the tree. Mutations are performed by selecting a parent and modifying the value or variable returned by a terminal or changing the operation performed by a node. Crossovers are performed by selecting two parents and swapping subtrees at randomly selected nodes within their trees. The new individuals so generated replace less fit members of the population chosen probabilistically on the basis of their unfitness.

For the GP implementations used here, three types of GP were used: (1) arithmetic GP, which used the node operator functions “add”, “subtract”, “multiply”, and “protected divide” (where $n/0 = 1$); (2) transcendental GP, which used the four arithmetic functions plus “inverse ($1/x$)”, “negate ($-x$)”, “square”, “square root”, “absolute value ($|x|$)”, “exponent (e^x)”, “natural logarithm”, “sine”, “cosine”, and “tangent”; (3) conditional GP, which used the four arithmetic functions plus the function “if-then-else”. All GP analyses were carried out using an in-house program,[41] which runs under Microsoft Windows NT on an IBM-compatible PC; on a Pentium 133, with 128 MB of RAM, a typical run took 1 (PyMS) and 5 min (FT-IR). The GP analyses used the following reproductive strategy: when a parent is chosen, there was a 0.7 probability of a crossover, 0.2 probability of mutation, and a 0.1 probability of direct duplication.

The GP used five independent subpopulations (demes) with a 5% migration every 10 generations. The deme size was set to 300 individuals (therefore the population size was 1500). The maximum number of generations was set to 5000 (therefore the total number of allowed operations was 5000 × 1500). Convergence was taken to have been achieved when the fitness of the best individual, defined as the root-mean-squared (rms) error between the training set estimates, and the true values was within 0.1%. In order for relatively simple rules to be developed, the tree complexity was constrained by setting the maximum number of nodes used to 100 and the maximum depth of the trees to only 8 layers; in addition, a penalty of 0.001 multiplied by the number of nodes in the individual’s function tree was implemented to reduce verbose trees. Finally, to assess the GPs reproducibility each type of GP was run 10 times using different randomly chosen starting populations.

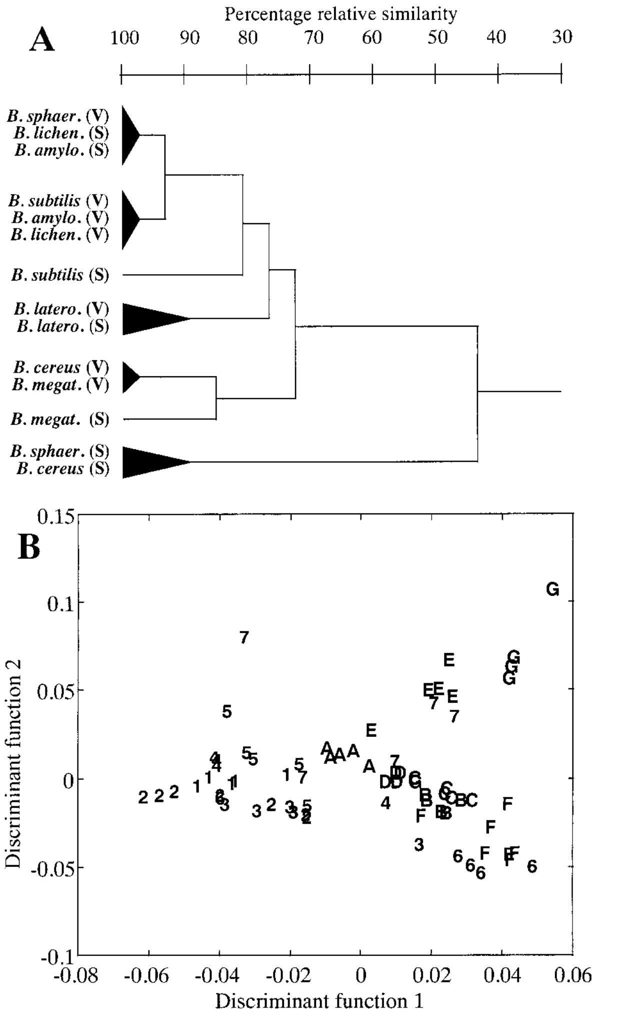


Figure 3. Cluster analyses (as detailed in the text) of both the *Bacillus* spp. vegetative cells (V) and spores (S) by PyMS (A) and FT-IR (B) spectra. In part B: 1, A = *B. sphaericus*, 2, B = *B. subtilis*, 3, C = *B. licheniformis*, 4, D = *Br. laterosporus*, 5, E = *B. cereus*, 6, F = *B. amyloliquefaciens*, and 7, G = *B. megaterium* (where numbers are vegetative cells and letters are spores).

RESULTS AND DISCUSSION

Raw Data. Typical normalized PyMS and FT-IR spectra of *B. subtilis* in its vegetative and sporulated states are shown in Figures 1 and 2, respectively. Visual inspection of these spectra indicates that there was very little qualitative difference between the spectra (and indeed between the others collected), although at least some complex quantitative differences between them were observed. Such spectra, essentially uninterpretable by the naked eye, readily illustrate the need to employ multivariate statistical techniques for their analyses.

Cluster Analyses. PyMS analyses of all the vegetative cells (Figure 3A) showed that *Bacillus amyloliquefaciens, Bacillus licheniformis,* and *B. subtilis* comprised one tight group, while *B. cereus* and *Bacillus megaterium* formed another cluster, and *Brevi laterosporus* and *Bacillus sphaericus* were recovered separately. Similar groupings from the cluster analysis of FT-IR spectra from these bacilli was also observed (Figure 3B). Moreover, the discrimination found by both these hyperspectral methods, which measure the total biochemistry of these bacterial cells, was in agreement with discrimination based on their DNA homologies

[41] Gilbert, R. J.; Goodacre, R.; Woodward, A. M.; Kell, D. B. *Anal. Chem.* **1997**, *69*, 4381−4389.

[42] Naumann, D.; Helm, D.; Labischinski, H.; Giesbrecht, P. In *Modern techniques for rapid microbiological analysis*; Nelson, W. H., Ed.; VCH Publishers: New York, 1991; pp 43−96.

Table 2. Percentage of Correct Estimations from Each of the Supervised Analysis Methods, PyMS and FT-IR

	PyMS		FT-IR	
	training set	test set	training set	test set
MLPs	100	100	100	100
RBFs	100	100	100	100
PLS	100	100	100	100
CART (univariate rule induction)	96.4	95.5	97.6	93.9
Breiman (multivariate induction)	100	97.7	97.6	96.2
OC1 (multivariate induction)	96.4	95.5	97.6	93.9
GP using arithmetic functions[a]	85.7−100	86.4−100	92.9−100	87.1−97.7
GP using transcendental functions[a]	95.2−100	90.2−100	96.4−100	84.1−93.9
GP using arithmetic and conditional functions[a]	100	100	95.2−100	84.1−93.2

[a] The minimum and maximum are shown from 10 different GP rules.

as judged by phylogenetic analysis of 16S rRNA sequences from representatives of these organisms.[13]

Also shown in Figure 3 are the same PyMS and FT-IR analyses of the vegetative cells with the sporulated bacilli. The taxonomic relationship shown by the PyMS analyses (Figure 3A) of the spores was very different from that afforded by the vegetative cells, a phenomenon observed previously with *Bacillus*.[25] However, the phenotypic extent to which the sporulated biomass is different from the vegetative cells was inconsistent. To illustrate this, both physiological types of *B. megaterium* are recovered in the same cluster with the vegetative cells of *B. cereus*, at an 85% relative similarity, but the spores from *B. cereus* are grouped with the spores from *B. sphaericus*, and this group had only 45% relative similarity with vegetative *B. cereus*. The phenotypic relationship between vegetative and sporulated biomass by FT-IR was also inconsistent, although the DFA plot (Figure 3B) did show that there was an overall trend in discriminant function 1, which may explain whether the bacteria were sporulated or not. A similar trend was seen in the DFA plot based on PyMS data (not shown), suggesting that this difference in physiology might be detected automatically.

Supervised Analyses. To distinguish between spectra from spores or vegetative cells, supervised learning was implemented using training and test sets detailed in Table 1. The output was encoded as either “0” for vegetative biomass or “1” for spores. MLPs, RBFs, and PLS were calibrated on the training set as detailed above and then interrogated with both the training and test sets. All three methods predicted all 216 spectra correctly for both the PyMS and FT-IR spectra (Table 2). Therefore it was possible to distinguish the different bacterial physiological states, but only after having used supervised analyses. However, to find a good biomarker for bacterial sporulation, an additional question that needs to be answered is, “which inputs (either PyMS m/z intensities or FT-IR wavenumber absorbances) were indicative of spores?”. Although MLPs, RBFs, and PLS are excellent methods of supervised learning, the information of which masses in the mass spectrum or wavenumbers in the infrared spectrum are important is not *readily* available. For MLPs and RBFs, the information used by these neural networks can nominally be found in their weights; however, this information is very abstract and almost impossible to extract realistically, especially when these ANNs are *interconnected*, and for the MLPs trained with the PyMS and FT-IR spectra these contain 1217 and 8841 weights, respectively. For PLS, the interpretation is theoretically simpler as the PLS model is a summation of the dot products of linear weighting vectors (latent variable loadings) and the original data; but when plotted (data not shown) they are as complex as the original spectra, and no single m/z intensity or IR absorbance was seen to be especially important.

Spore Biomarker in PyMS. Rule induction and GP techniques were performed on the same training sets as before and the results are shown in Table 2. For the unseen test set, the best GPs assess the physiological status of the bacilli more accurately than any of the rule induction techniques. However, *some* of the GPs are not as accurate as rule induction; this is presumably because the initial starting populations are chosen randomly and by chance some of the 10 GPs had converged to local minimums. Each of the rule induction methods produce only a single rule because they are deterministic and do not use random starting points.

The next stage was to inspect the rules created by each of the methods. Both the univariate CART and multivariate OC1 rule induction methods produced the same single rule

$$\text{IF } m/z\ 105 < 0.3985, \text{ THEN vegetative cells ELSE spores}$$

while the multivariate Breiman rule induction method produced the single rule

$$\text{IF multivariate rule} < 0.4282, \\ \text{THEN vegetative cells ELSE spores}$$

where the multivariate rule only consisted of the intensity of m/z 105.

The genetic programs all produced different function trees, again indicative of local minimums being found rather than a global one, but these also used predominantly m/z 105, and an example of one of the rules from a GP trained with transcendental functions was

$$\text{output} = \{\log(2 \times m/z\ 105) + \cos[(m/z\ 154 \times m/z\ 84) - (2 \times m/z\ 105)]\}^{1/16}$$

where the output was 1 for spores and 0 for vegetative cells. While

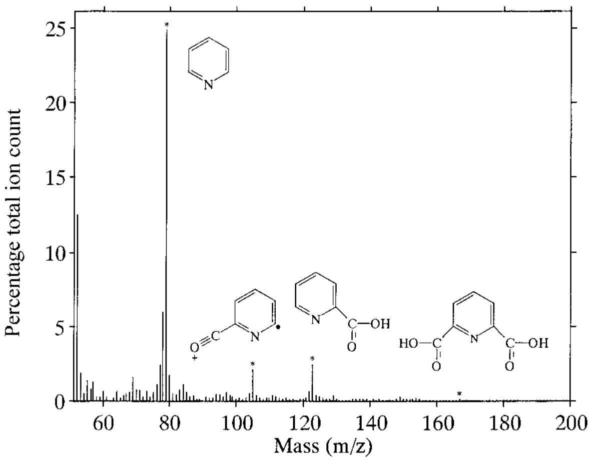


Figure 4. Curie-point pyrolysis-MS spectrum of dipicolinic acid.

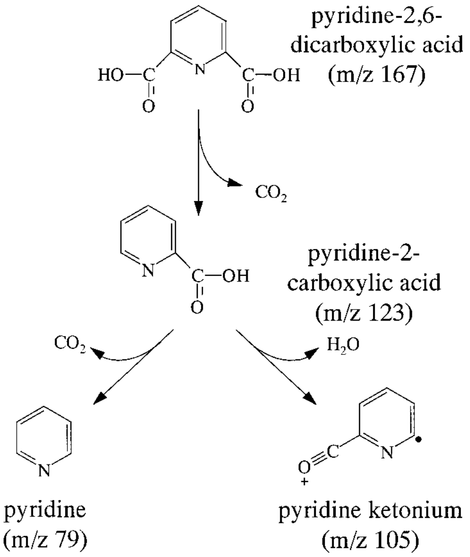


Figure 5. Pyrolytic degradation and electron impact fragmentation of dipicolinic acid.

one from a GP using arithmetic and conditional functions was

$$\text{IF } [m/z\ 65 \times (m/z\ 105 + m/z\ 76)] \geq m/z\ 121, \quad \text{THEN spores ELSE vegetative cells}$$

where the output was 1 for spores and 0 for vegetative cells.

The next stage was to find the biochemical substance characteristic of the physiological difference between vegetative cells and spores, which was characterized by *m*/*z* 105 in the PyMS spectra. It is known that *Bacillus* spores contain 5−14% (dry weight) DPA and have elevated protein levels in the spore coat. Therefore, DPA (Sigma) was analyzed by PyMS and its spectrum is shown in Figure 4. It can be seen that DPA has very intense peaks at *m*/*z* 79, 105, and 123; the molecular ion at *m*/*z* 167 is also just visible. Under thermal degradation in vacuo, one would expect DPA to decarboxylate (lose one or two CO_2 molecules at either the 2 or 6 position on the pyridine ring), and this was indeed found to be the case because *m*/*z* 123 and 79 are from pyridine-2-carboxylic acid (MPA; monopicolinic acid) and pyridine, respectively. Initially it was difficult to assign the molecular fragment responsible for the *m*/*z* 105 peak with a known structure, so Py-GC/MS was used to analyze pure DPA and the MS was programmed to scan for fragments with a molecular weight of 105. It was found that the *m*/*z* 105 fragment came from pyridine-2-carboxylic acid rather than directly from pyridine-2,6-dicarboxylic acid, and this allowed us to elucidate that the pyrolytic degradation and electron impact fragmentation of dipicolinic acid proceeds as shown in Figure 5; DPA decarboxylates to MPA, and then this (and any other MPA in the spore coat) undergoes electron impact fragmentation and hydrogen abstraction, and hydrolysis occurs to produce a pyridine ketonium ion ($C_6H_3ON^+$).

Subtraction spectra of each strain in both physiological states confirmed that *m*/*z* 105 was very significant, and more so than *m*/*z* 79 or 123 (data not shown). Other plots showed that *m*/*z* 91 was also elevated in spores (shown in Figure 1), which is not surprising since this ion is from toluene, the base peak of phenylalanine found in proteins. It might be supposed that the rule induction and GPs should have indicated that *m*/*z* 91 was important for the differentiation between spores and vegetative cells, but while the protein content in spores is *quantitatively* higher than in vegetative cells, this is not as characteristic a difference as the *qualitative presence* of an entirely new biochemical substance.

Spore Biomarker in FT-IR. For the analysis of the FT-IR spectra from vegetative cells and spores, CART and OC1 rule induction methods produced the same rule:

$$\text{IF } 1443\ \text{cm}^{-1} < 0.0135 \text{ AND } 3827\ \text{cm}^{-1} < 0.0005, \quad \text{THEN vegetative cells}$$

$$\text{IF } 1443\ \text{cm}^{-1} < 0.0135 \text{ AND } 3827\ \text{cm}^{-1} > 0.0005, \quad \text{THEN spores}$$

$$\text{IF } 1443\ \text{cm}^{-1} > 0.0135, \text{ THEN spores}$$

while the multivariate Breiman rule was more complex and was

$$\text{IF multivariate rule} < 0.0099, \quad \text{THEN vegetative cells ELSE spores}$$

where the multivariate rule was dominated by wavenumbers 3896, 3861, 3854, 3842, 3811, 3541, 3533, 1443, and 933.5 cm^{-1}.

Again the GPs produced many different complex function trees, but many of these used the vibration at peak 1443 cm^{-1} (as well as the two wavenumbers at 1447 and 1439 cm^{-1} collected either side of this vibration) as the discriminating variable(s). A typical function tree in reverse Polish notation from an arithmetic GP was

```
(+ (+ (+ (+ (+ p1389
                (+p1389
                   (*p1539 5.09451)))
             (+ (+ (* -9.44825 p903) p1443) p1443))
          (/ (* -6.63661 p3838)
             (* 4.55398 p3306)))
       (/ p2287 p2855))
    (+p1443
       (+ (* p609 9.79268)
          (+p1443
             (+p1443
                (+ (/ (* -6.63661 p3838)
                      (* 4.44827 p3306))
                   (+ (* -9.44825 p903) p1443)))))))
```

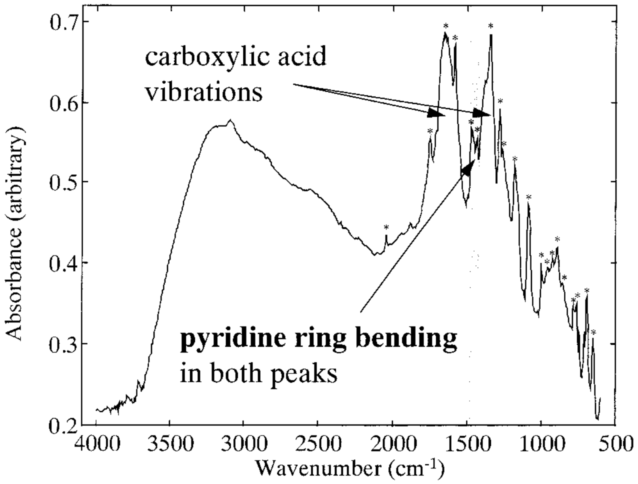


Figure 6. Diffuse reflectance−absorbance FT-IR spectrum of dipicolinic acid. *, vibrations from dipicolinic acid at 2044, 1747, 1643, 1582, 1470, 1435, 1342, 1281, 1261, 1180, 1088, 999, 922, 891, 783, 760, 744, 710, 694, and 652 cm^{-1}.

this complex tree has been chosen to illustrate that the vibration at 1443 cm^{-1} is used many times. To simplify this rule, the following procedure was adopted: first the program “unpolish”, written in-house by Alun Jones was used to turn this function tree into normal algebra; next Maple (Waterloo Maple Inc.), also running on a UNIX workstation, which uses symbolic mathematical calculations was used to simplify this algebra furthur to

$$\text{output} = [(6 \times \text{p1443}) + (2 \times \text{p1389}) + (5.09 \times \text{p1539}) + (9.79 \times \text{p609}) + (\text{p2287}/\text{p2855})]$$

where the output was 1 for spores and 0 for vegetative cells. While an example of a relatively simple rule generated from using transcendental functions which used only two wavenumbers was

$$\text{output} = \{\tan[(\text{p1443} + \text{p602})^{1/16}]\}^{1/4}$$

where the output was 1 for spores and 0 for vegetative cells.

Pure DPA was analyzed by FT-IR and the resultant spectrum is shown in Figure 6. An AM1 semiempirical force field method using Hyperchem version 5.1 (HyperCube Inc.) was used to elucidate which parts of the vibrational normal mode of the DPA (as the dipicolinate ion solvated in H_2O) occurred at around 1443 cm^{-1}, and it was found that this peak area was dominated by two vibrations (indicated in Figure 6) from two different flexing modes of the pyridine ring. It is noteworthy that the rule induction and GP homed in on this specific small pyridine vibration rather than the large broad vibrations from the two carboxylic acids (also shown in Figure 6), since the later will have arisen from many other biochemicals present in cells and spores. The large amide I vibration from proteins at 1666 cm^{-1} shown in Figure 2 was not highlighted as being important for differentiation of spores and vegetative cells; subtractions of spore and vegetative cell spectra using integration under this amide I vibration showed no significant elevation of proteins in spores compared with those of vegetative cells.

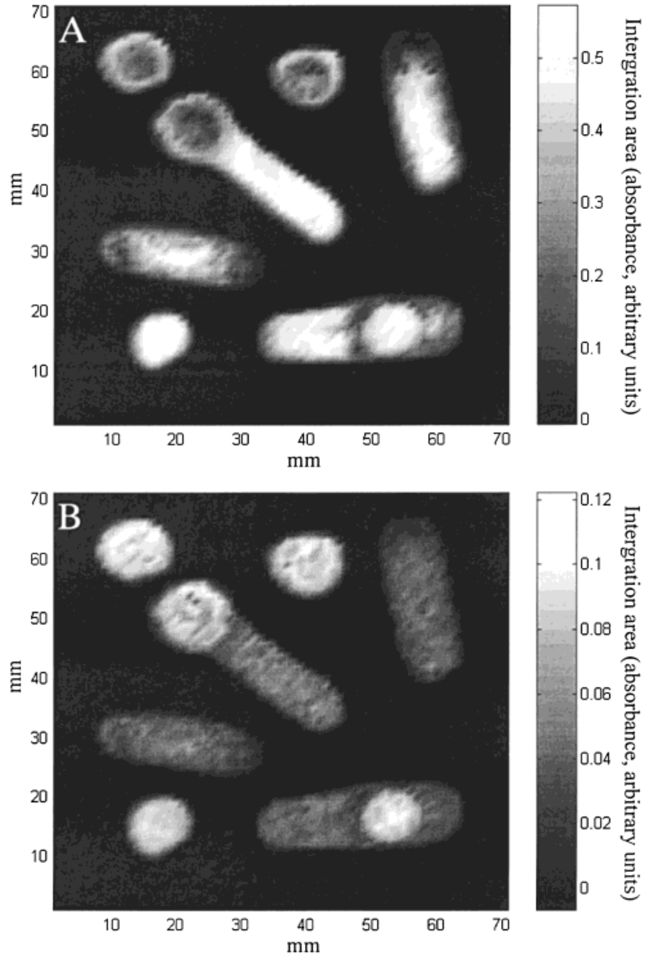


Figure 7. Chemical images of (A) amide I vibration at 1666 cm^{-1} and (B) pyridine vibration at 1443 cm^{-1}. Vegetative and sporulated biomass from *B. cereus* B0002^T was applied to the surface of a 7 cm by 7 cm metal plate (∼200 μg/cm^2). Data were acquired at a resolution of 1 mm (therefore, these maps are 71 by 71 pixels; 5041 spectra).

To illustrate the significance of this biomarker for spores, vegetative and sporulated biomass from *B. cereus* B0002 was applied to the surface of a 7 cm by 7 cm metal plate at a concentration of ∼200 μg/cm^2 (dry weight); various cartoons of vegetative cells, vegetative cells containing spores, and free spores were drawn with the biomass. FT-IR spectra were acquired at a spatial resolution of 1 mm (therefore these maps are 71 pixels by 71 pixels, by 882 wavenumbers). Figure 7 shows two slices from this data cube, the first slice (Figure 7A) is from the simple integration from 1682 to 1651 cm^{-1} under the amide I band at 1666 cm^{-1}. This chemical image indicates where proteins are found on the plate, and shows, as expected, that proteins are found in both the vegetative and sporulated biomass, although this is to a variable degree. The second chemical image (Figure 7B) is of DPA from the integration from 1458 to 1427 cm^{-1} under the pyridine vibration at 1443 cm^{-1} and clearly shows where the sporulated biomass has been applied to the surface of the plate. This highlights the point that now that the spore biomarker DPA has been detected at a single reproducible vibration that simple integration is sufficient for the detection of spores compared to vegetative cells; indeed, there is now no need to use complex mathematical calculations.

CONCLUDING REMARKS

The “whole-organism fingerprinting” techniques of Curie-point PyMS and diffuse reflectance−absorbance FT-IR were used to analyze a diverse group of 36 bacterial strains which belong to 1 of 7 species. Unsupervised cluster analyses were used to reduce the dimensionality of these hyperspectral data, and while from vegetative biomass the discrimination found was in agreement with phylogenetic data, this discrimination changed when spores were analyzed, highlighting that for these seven species the biochemistry of the physiological change was not wholly consistent.

Neural network analyses and PLS regression of the PyMS and FT-IR spectra showed that these supervised analyses could discriminate easily between spores and vegetative cells; however, the information in terms of which mass intensities or wavenumber absorbances in the MS and IR were important was not available. Therefore, three rule induction methods and three GPs with varying levels of complexity (arithmetic, transcendental, to conditional) were used to classify the physiological state (vegetative biomass versus spores) correctly, with the added benefit that they all produced mathematical rules that could be interpreted biochemically.

In the mass spectrometric studies, it was found that m/z 105 was highly characteristic for spores and is a pyridine ketonium ion ($C_6H_3ON^+$) obtained from the thermal degradation in vacuo of dipicolinic acid. For FT-IR, a pyridine ring vibration at 1447−1439 cm^{-1} from the same dipicolinic acid biomarker was found to be highly characteristic of spores.

Rather than the rule induction and GPs replacing ANNs and PLS for the identification of this physiological difference, since all are very complex multivariate mathematical techniques, one can now use a very simple univariate approach since a specific mass ion in the PyMS and specific vibration in the FT-IR have been detected. Indeed, now that a specific characteristic spore biomarker has been elucidated, then it would be possible to fine-tune these two analytical spectroscopies specifically for dipicolinic acid. In the case of FT-IR, it may be possible to devise a simple dispersive spectroscope based on the integration of the absorbance centered at 1443 cm^{-1}, thereby reducing the cost of this analysis, since the necessity for a complex inteferometer-based instrument would be removed.

In conclusion, these results demonstrate that PyMS and FT-IR can be used to detect rapidly whether a *Bacillus* culture is sporulated or not and more importantly that this discrimination can be assigned unequivocally to the spore biomarker dipicolinic acid.

ACKNOWLEDGMENT

R.G. and E.M.T. are indebted to the Wellcome Trust for financial support (Grant 042615/Z/94/Z). R.G. and B.S. thank the Society of Microbiology for a vacational studentship. A.C.M. thanks Zeneca Pharmaceuticals for financial support. D.B.K., R.J.G., and B.K.A. thank the Engineering and Biological Systems Committee of the UK BBSRC and the EPSRC Analytical Science Program for financial support.

Received for review June 16, 1999. Accepted September 24, 1999.

AC990661I