Chemometric Analysis of Diffuse Reflectance-Absorbance Fourier Transform Infrared Spectra Using Rule Induction Methods: Application to the Classification of *Eubacterium* Species

BJØRN K. ALSBERG,* WILLIAM G. WADE, and ROYSTON GOODACRE

Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion, SY23 3DA, United Kingdom (B.K.A., R.G.); and Oral Microbiology Unit, Department of Oral Medicine and Pathology, UMDS, Guy's Hospital, London, SE1 9RT, United Kingdom (W.G.W)

Strains representative of four *Eubacterium* species were analyzed by using diffuse reflectance-absorbance Fourier transform infrared spectroscopy. To identify important wavenumber regions for the classification of these 22 bacterial isolates, we investigated three rule induction methods and various spectral preprocessing regimes. In this study both univariate and multivariate classification and regression trees (CART) methods and the fuzzy multivariate rule-building expert system (FuRES) method were exploited. It was found that the FuRES method was superior in terms of prediction, whereas the rules proposed by the univariate CART method were easier to interpret in terms of which wavenumbers in the IR spectra were important for bacterial class separation. Scaled and detrended FT-IR spectra and first-order numerical differentiation preprocessing steps were necessary to obtain optimal classification models. Finally, a reduction in the classification error for the CART-based methods was observed by analyzing the compressed B-spline coefficients rather than the original spectra representation. The spectral interpretation of these rules is in agreement with analyses using the uncompressed representation.

Index Headings: ***Eubacterium*; Rule induction; Fuzzy rule induction; CART; Discriminant function analysis; Infrared spectroscopy.**

INTRODUCTION

The oral asaccharolytic *Eubacterium* species are a diverse group of organisms that are implicated in periodontitis, endodontic infections, and dentoalveolar abscesses.[1,2] They are slow-growing and difficult to identify by conventional means. The number of documented species continues to increase,[1,3,4] and in a recent study[5] we used pyrolysis mass spectroscopy (PyMS) to confirm the taxonomic position of the three newly described species *E. exiguum, E. infirmum,* and *E. tardum.*

The ideal method for rapid and accurate identification of micro-organisms, particularly in the clinical laboratory, would have minimum sample preparation, would analyze samples directly, and would be rapid, automated, accurate, and (at least relatively) inexpensive.[6] With recent developments in analytical instrumentation, these requirements are being fulfilled by physicochemical spectroscopic methods, often referred to as ''whole-organism fingerprinting.''[7] The most common methods are pyrolysis mass spectrometry,[6] Fourier transform infrared spectroscopy (FT-IR),[8–11] and Raman spectroscopy.[12–14]

FT-IR allows the chemically based discrimination of intact microbial cells, without their destruction, and produces complex biochemical fingerprints that are reproducible and distinct for different bacteria. Naumann and co-workers[9,10] have shown that FT-IR absorbance spectroscopy (in the mid-IR range, usually defined as 4000–400 cm^{-1}) provides a powerful tool with sufficient resolving power to distinguish microbial cells at the strain level. However, the interpretation of the FT-IR spectra has conventionally been by the application of unsupervised pattern recognition methods of correspondence analysis maps and cluster analysis,[11] which can often give rise to subjective interpretation of complicated scatter plots and dendrograms.

More recently, various related but much more powerful chemometric methods have been applied to the ''supervised'' analysis of FT-IR data. For example, Goodacre and colleagues[8] have used diffuse reflectance-absorbance FT-IR spectroscopy and artificial neural networks (ANNs) to successfully identify clinical isolates of *Enterococcus faecalis, E. faecium, Streptococcus bovis, S. mitis, S. pneumoniae,* or *S. pyogenes*; and in a more recent study we have also been able to use ANNs and radial basis function neural networks to identify bacterial isolates associated with urinary tract infection from their FT-IR spectra.[15] However, although chemometric methods based on artificial intelligence have been shown to be effective tools for microbial identification and discrimination from FT-IR data, the information in terms of which wavenumbers in the IR spectrum are important is not *readily* available, and ANNs are often perceived as a ''black box'' approach to modeling spectra.

In a previous study we have used rule induction to solve classification problems by using hyperspectral data from pyrolysis mass spectrometry.[16] This study allowed at least some indication of which masses (*m/z* values; mass-to-charge ratio) were important in the discrimination of three different milks[17] and for detection of the adulteration of extra virgin olive oil with lower grade oils.[18,19]

The aim of this study was to exploit rule induction methods to deconvolute the infrared spectra of strains previously identified by polyphasic phenotypic and genotypic phylogenetic analyses as one of four *Eubacterium* species. To be confident in the ''rules'' or ''trees'' produced by these induction methods, we used an independent test set of each of the four species along with

Received 26 November 1997; accepted 23 February 1998.
* Author to whom correspondence should be sent.

0003-7028/98/5206-0823$2.00/0

TABLE I. Description of bacterial strains used in this study.[a]

Identifier	Species/group	Strain number	Training or test set
Ta	*E. timidium*	ATCC 33093^T	Training
Tb		W557	Training
Tc		W690	Training
Td		W2847	Test
1a	*E. infirmum*	NCTC 12940^T	Training
1b		W687	Training
1c		W1475	Training
1d		W1470	Test
2a	*E. exiguum*	SC142	Training
2b		SC108	Training
2c		W1365	Training
2d		W733	Test
Na	*E. tardum*	SC68	Training
Nb		SC88P	Training
Nc		SC41B	Training
Nd		SC37	Training
Ne		NCTC 12941^T	Test
Ha	Clinical isolates	SBH463	Test
Hb		SBH481	Test
Hc		SBH462	Test
Hd		SBH403	Test
He		SBH477	Test

[a] Group of 22 *Eubacterium* strains analyzed by FT-IR.

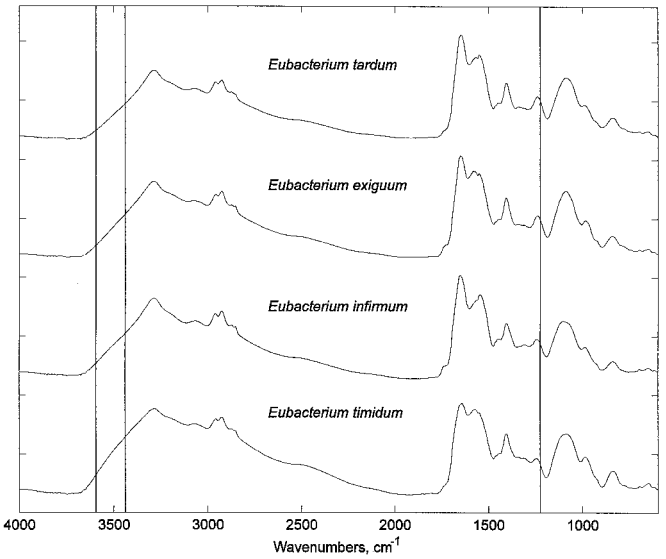

FIG. 1. This figure shows the mean spectrum (of the detrended data) for each of the four *Eubacterium* classes. The three most important variables found by the CART methods are added as vertical bars.

five clinical isolates that had been recently identified as *E. exiguum*.[5]

EXPERIMENTAL

Sample Preparation. Details of the organisms are given in Table I. Strains were cultured on Fastidious Anaerobe agar (Lab M, Bury, U.K.) plus 5% sheep blood and incubated anaerobically in an atmosphere of 80% N_2, 10% CO_2, and 10% H_2 for 72 h. The bacteria were harvested with a nichrome wire loop and suspended in physiological saline (0.9% NaCl) to approximately 20 mg mL^{-1}.

Diffuse Reflectance-Absorbance Fourier Transform Infrared Spectroscopy. Ten microliter aliquots of the above bacterial suspensions were evenly applied onto a sandblasted 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., Banner Lane, Coventry, U.K.) equipped with an MCT (mercury-cadmium-telluride) detector cooled with liquid N_2. The aluminum plate was then loaded onto the motorized stage of a reflectance TLC accessory.[20–24] The background spectrum was recorded from an empty well.

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 to 600 cm^{-1}. Spectra were acquired at a rate of 20 s^{-1}. The spectral resolution used was 4 cm^{-1}. In an effort to improve the signal-to-noise ratio, 256 spectra were coadded and averaged. Each sample was thus represented by a spectrum containing 882 points (set by the digitization interval of the IR instrument), and spectra were displayed in terms of absorbance as calculated from the reflectance-absorbance spectra by using the Opus software. The samples were applied to a 20 × 20 array of wells on a sandblasted aluminum plate. Full details about the preparation and the aluminum plates can be found in the work of Winson and co-workers.[24] Typical FT-IR spectra of the four *Eubacterium* species are shown in Fig. 1.

Preprocessing of FT-IR Data. Before analysis it was necessary to perform a set of preprocessing steps on the raw data sets: (1) Regions in the spectra that are dominated by CO_2 (2403.21–2272.06 cm^{-1} and 682.77–655.77 cm^{-1}) were filtered out to follow the spectral trend of neighboring data points (although the contribution to CO_2 in this data set was hardly detectable in these regions). (2) Next the spectra were normalized so that the smallest absorbance was set to 0 and the highest to +1 for each spectrum. (3) It was also necessary to perform a detrend operation with a linear function increasing from 4000 cm^{-1} to 600 cm^{-1}.[8,25] (4) Finally, it was necessary to remove baseline effects; there are several ways of doing this, and we used the first-order numerical differentiation to each of the spectra using a Savitsky–Golay filter.[26]

From each of the four ''known'' species, one of the samples (in triplicate) was reserved to form a test set (details are given in Table I); in addition all those spectra from the clinical bacterial isolates [(Ha–He; the SBH strain numbers (Table I)] were also used in the ''unknown'' test set. The remaining samples were used as training data for the various rule induction methods.

Data Analysis. *Rule Induction Methods.* Rule induction attempts to partition the space of sample objects into regions of single class memberships.[27] The data set is *recursively split* into smaller subsets where each subset contains objects belonging to as few different classes as possible. The ''purity'' of a subset (i.e., the distribution among the classes of the objects within the set) is often measured by using the concept of *entropy*.[28] For each subset there are fractions or probabilities $P = [p_1, p_2, \ldots, p_J]$, of the objects belonging to the J different classes. The entropy of P

$$H(P) = -\sum_{i=1}^{J} p_i \log(p_i) \qquad (1)$$

has properties in accordance with our intuitive understanding of ''impurity'': $H_{min}(P) = 0$ and $H_{max}(P) = log_2(J)$ when $P_i = 1/J$. Thus, ensuring the highest purity in a subset corresponds to *minimizing* $H(P)$ by selecting an optimal partitioning.

There are two major strategies for finding the best object paritioning, which in general are described as *univariate* and *multivariate* rule induction. In univariate rule induction, the single variable x_i at each recursion step is found that gives rise to the purest subsets (i.e., those that have minimum entropy). In univariate rule induction, a split or a partitioning of the input feature space can be formulated as a question such as ''Is $x_i < c$?'' where c is some value chosen from the ***finite set*** of values variable x_i has among the N calibration objects. All the objects that satisfy the question are grouped into one subset and those that do not into another. It should be emphasized that the rule induction methods described here operate on real valued variables only. It is also possible to use the same rule induction methodology on variables that have a ***finite*** set of values; see, for example, Quinlan's C4.5 method.[29]

In *multivariate rule induction,* a partition of the input feature space is found that depends on a linear (or nonlinear) combination of all the variables instead of just one variable. In this article we will consider only linear combinations of the original variables.

A multivariate rule induction partitioning can thus be formulated as the following type of question: ''Is

$$\sum_{j=1}^{n} w_j x_j \leq c \ ?''.$$

This type of partitioning of the data space is particularly useful if there are any colinearities between the variables. Our aim now is not to find the single best variable, but to find an arbitrary direction in the multivariate space represented by the vector **w** which best separates the data set into subsets that have maximum purity. The object vectors stored as rows in the matrix **X** are projected onto **w**:

$$\mathbf{t} = \mathbf{X}\mathbf{w}/(\mathbf{w}^{\mathrm{T}}\mathbf{w}) - b \qquad (2)$$

where b is a bias (scalar) and **t** is a score vector where the sign of each score determines whether that object is to be classified as class 1 (''yes'') or 2 (''no''). This is similar to the paradigm used in the multivariate rule building expert system (MuRES) approach by Harrington,[30] the oblique rule induction by Murthy et al.,[31] and the linear machine decision trees by Utgoff and colleagues.[32–34] The fuzzy multivariate rule-building expert system (FuRES)[35] approach is a modification of the MuRES algorithm where the central change is in introducing *fuzzy set theory*[36] in the handling of object class memberships. As can be seen from the projections onto **w**, it may be difficult to say on which side of the hyperplane an object really is located. When an object can be on one side (membership value = 1) or the other (membership value = 0) of the hyperplane only, we refer to this as *crisp classification.* To describe intermediate positions relative to the hyperplane fuzzy sets is ideal since they allow the degree of position to be described by a continuous value between 0 and 1. Consequently, an object may exist on both sides of the hyperplane because the plane itself is fuzzy. To obtain a ''fuzziness'' in the object position relative to **w**, one can also use a *logistic function* on the score value t_i.[35,37]

The use of the logistic (sigmoid) function makes FuRES analogous to the feedforward multilayer perceptron architecture commonly associated with artificial neural networks and is thus sometimes referred to as a *minimal neural network* (MNN).[37] Both FuRES and ANNs can produce nonlinear decision surfaces. However, they differ in that FuRES partitions the row space of the input training data matrix and ANNs partition the column space.

For univariate and multivariate classification and regression trees (CART) we used the OC1 program[31] (Department of Computer Science, Johns Hopkins University, Baltimore, MD). The FuRES program was kindly provided by Prof. Peter de B. Harrington. Both programs run under Windows NT 4.0.

Discriminant Function Analysis (DFA). Discriminant function analysis [(also referred to as canonical variates analysis (CVA)] is a multivariate statistical technique that separates objects (samples) into groups or classes by minimizing the within-group variance and maximizing the between-group variance.[38–40]

The general principle of DFA is similar to that of principal components analysis (PCA), but the objective of DFA is to find latent variables that maximize the ratio of the between-group to within-group variance, rather than maximizing the total variance. To find the discriminant functions (DFs) direction, the within-sample matrix of sums of squares and cross products, **W**, and the total sample matrix of sums of squares and cross products, **T**, are first computed. The between-group matrix is computed as $\mathbf{B} = \mathbf{T} - \mathbf{W}$, and the eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ correspond to the CVs onto which our data objects were projected.[38,41]

In the DFA used here, PCA was first used to reduce the dimensionality of the data by using only the A first principal components (PCs). The remaining PCs are usually due to random ''noise'' in the data[42] and can be ignored without reducing the amount of useful information representing the data. The DFA algorithm used here was implemented in MATLAB (The MathWorks, Natick, MA).

B-splines. B-splines[43] is a powerful method for compressing smooth curves and surfaces. It has also been shown to be effective in the compression of spectral profiles.[44–46] B-splines represent curves as a linear combination of a set of bell-shaped basis functions. The shape and position of these basis functions are determined by a set of knots. For a one-dimensional (1D) B-spline, a set of knots corresponds to points along the abscissa. If the number of knots is q and the local polynomial degree is k, then we have $n = q - k - l$ different B-spline basis functions. For a given function $f(x)$ we find a set of coefficients c_j using least-squares fitting so that

$$f(x) = \sum_{i=1}^{n} c_i B_{i,k}(x) + \epsilon \qquad (3)$$

where $B_j(x)$ is the jth B-spline basis function, and ϵ is the reconstruction error. This approach means that if the re-

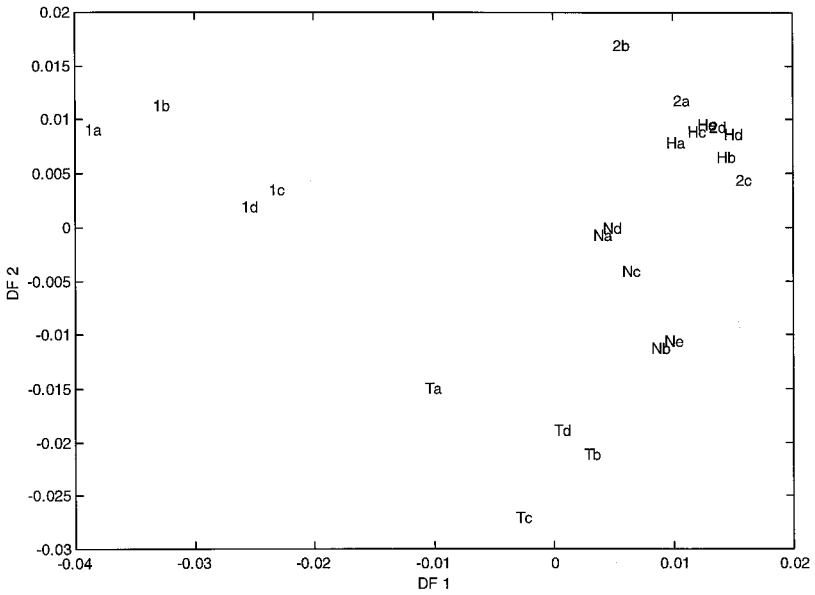

FIG. 2. Plot of the first two latent variables from the discriminant function analysis. Data points are the averages for the replicate samples, and their identifiers are given in Table I.

construction error $\in$ is satisfactory, $f(x)$ is completely described by a knot vector **h** and a coefficient vector **c.** The knot vector is a nondecreasing sequence of position along the abscissa:

$$h_0 \leq h_1 \leq \ldots \leq h_{n+k}$$

where k is the maximum degree of any local polynomial.

The B-spline bases are recursively generated from the information in the knot vector as follows:

$$B_{j,k}(x) = \frac{x - h_j}{h_{j+k} - h_j} B_{j,k-1}(x) + \frac{h_{j+k+1} + x}{h_{j+k+1} - h_{j+1}} B_{j+1,k-1}(x) \quad (4)$$

$$j = 0, \pm 1, \pm 2, \ldots \qquad k = 1, 2, 3, \ldots .$$

For $B_{i,0}(x)$ we have

$$B_{i,0}(x) = \begin{cases} 1 & \text{if } h_i \leq x \leq h_{i+1} \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

To obtain compression of spectral profiles, it is necessary to use the same knot distribution for all spectra in the data set. This condition is to ensure comparability between the bases and coefficients in different spectra. The initial knot distribution is determined in this case by a method based on a maximum entropy assumption (see Ref. 44 for more details).

After B-spline analysis, each spectrum is represented by a coefficient vector **c.** These vectors are later subjected to analysis instead of the original (more redundant) data vectors.

RESULTS AND DISCUSSION

DFA. In order to perform DFA on these data it was necessary to ensure that the variables were independent. When the number of variables is larger than the number of objects (which is the case in almost all applications using whole spectra profiles), the matrix rank is much lower than the total number of variables. One way to ensure a nonredundant variable space is to perform a principal component analysis and use the resulting scores vectors in the DFA analysis—a method commonly used for the analysis of pyrolysis mass spectra,[6,47] and more recently for the analysis of FT-IR data.[8,25] By projecting onto the A first principal components, one obtains an efficient compression of the data set into A co-ordinates, where A is much less than the number of objects. DFA is a supervised method, which means that each feature vector must be associated with a class membership. In this study the *object replicate information* is used as membership value and *not* the bacterial type information associated with each object.

Since the number of *Eubacterium* classes is four, we should have a sufficient discrimination of the objects by using three discriminant function directions, so that each of the four classes is equidistant from each in multidimensional space. The resulting DFA analysis of these 22 *Eubacterium* isolates is shown in Fig. 2 where class representatives of each of the four classes (*E. exiguum, E. infirmum, E. tardum,* and *E. timidum*) are seen. Moreover, as expected from our previous mass spectrometry studies,[5] all the clinical isolates (SBH isolates) group with *E. exiguum.*[2]

CART and FuRES. As is usually the case, the measured infrared spectra are not always suitable for direct analysis. In order to account for any baseline effects and unwanted artifacts due to CO_2 (which in this experiment were minimal), the preprocessing regime of CO_2 removal, followed by scaling and detrending the spectra, was employed (as detailed above). In addition, any residual baseline distortions were removed by computing the first-order numerical differentiation.

The data set was represented to the rule induction algorithms either as (1) detrended data, (2) the Savitsky–Golay first-order differentiated spectra, or (3) the PCA scores of the detrended and differentiated spectra.

PCA was included since it is an excellent way of reducing the high dimensional data. In this case the original 882 variables were compressed to only 25 PC scores,

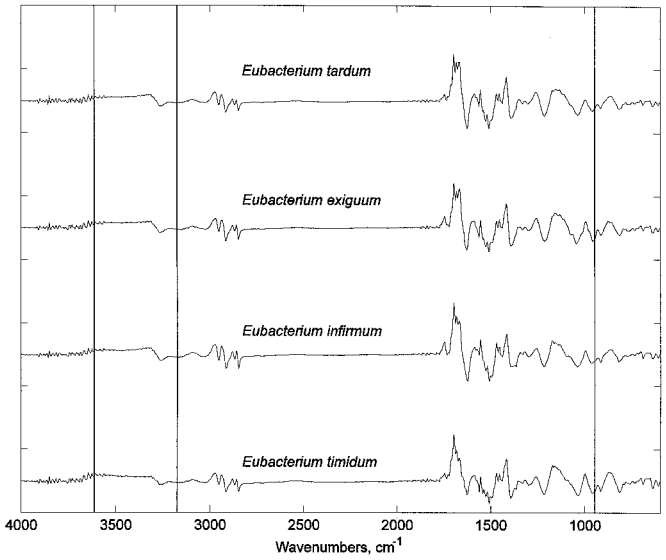

FIG. 3. This figure is analogous to Fig. 1. Here we show the mean differentiated spectra for the four classes together with the three most significant wavenumbers used in the rule induction classification models.

which accounted for more than 99.9% of the total variance in the detrended data set. Twenty-five principal components accounted for 98.4% of the detrended and differentiated data set.

For each of these three representations, four types of rule induction analyses were performed; the results obtained from pruned and unpruned trees are shown in Table II and Table III, respectively. All the CART-based methods perform better on the differentiated spectra than on the 25 PCA scores. The FuRES method applied to either the differentiated spectra or the PCA scores produced perfect predictions on the unseen validation set. The FuRES classification tree for the analysis of the differentiated data set is shown in Fig. 4. The profiles of the three rules (Rules 1,2, and 3) are shown in Fig. 5. We see that Rule 1 separates *E. timidum* from the three other species (Fig. 4). Rule 1 has a profile that is somewhat different from the other rules (Fig. 5). One characteristic feature is that region 1800–4000 cm^{-1}, compared to 600–1800 cm^{-1}, is smaller than it is for the other rules. Both Rule 2 and 3 have a more prominent feature in the region 2800–3000 cm^{-1} than does Rule 1 (Fig. 4). Rule 2 separates out *E. infirmum* from *E. tardum* and *E. exigum* and

TABLE II. Class predictions from pruned trees.[a]

Correct	DT1	DT2	DT3	DT4	D1	D2	D3	D4	S1	S2	S3	S4
1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	2	1	2	2	2	2	2	2	2	2	2
2	1	2	1	2	2	2	2	2	2	2	2	2
2	1	2	1	2	2	2	2	2	2	2	2	2
2	1	2	1	2	1	1	1	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	2	2	2	3
4	4	4	4	4	4	4	4	4	4	4	4	4
4	1	4	1	4	4	4	4	4	3	3	3	4
4	4	4	4	4	4	4	4	4	4	4	4	4
4	3	3	3	4	4	4	4	4	3	3	3	4
3	3	3	3	3	4	4	4	3	3	3	3	3
3	3	3	3	3	4	4	4	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3
3	1	3	1	3	4	4	4	4	4	3	3	3
3	3	3	3	3	4	4	4	4	4	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	1	3	3	3	3	1	1	1	3
3	3	3	3	1	3	3	3	3	1	1	1	3
3	3	3	3	1	3	3	3	3	1	1	1	3
3	3	3	3	1	3	3	3	3	1	1	1	3
3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	2	2	2	3
3	4	4	4	4	3	3	3	3	3	3	3	3
3	4	4	4	4	3	3	3	3	3	3	3	3
3	1	4	1	3	4	4	4	3	3	3	3	3
3	4	4	4	3	4	4	4	3	3	3	3	3
# Error	11	5	11	6	7	7	7	0	8	8	8	0
% Error	30.6	13.9	30.6	16.7	19.4	19.4	19.4	0.0	22.2	22.2	22.2	0.0

[a] Explanation: Each number in the columns represent a predicted class. Code "1" represents *E. timidum,* code "2" represents *E. infirmum,* code "3" represents *E. exiguum,* and code "4" is *E. tardum.* DT = detrended, D = first derivative, S = 25 PCA scores, 1 = univariate CART, 2 = multivariate CART (mix of uni- and multivariate rules) by Breiman, 3 = OC1 rule induction (mix of uni- and multivariate rules), 4 = FuRES. All calibrations are performed with leave-one-out cross-validation before prediction on unknown validation set. Pruned decision trees were used here.

TABLE III. Class predictions from unpruned trees.

Correct	DT1	DT2	DT3	D1	D2	D3	S1	S2	S3
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	2	2	2
4	4	4	4	4	4	4	4	4	4
4	2	2	2	4	4	4	3	3	3
4	4	4	4	4	4	4	4	4	4
4	3	3	3	4	4	4	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	2	2	2	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	1	1	1	1	1	1
3	3	3	3	1	1	1	1	1	1
3	3	3	3	1	1	1	1	1	1
3	3	3	3	1	1	1	1	1	1
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	2	2	2
3	4	4	4	3	3	3	3	3	3
3	4	4	4	3	3	3	3	3	3
3	2	2	2	3	3	3	3	3	3
3	4	4	4	3	3	3	3	3	3
# Error	7	7	7	4	4	4	8	8	8
% Error	19.4	19.4	19.4	11.1	11.1	11.1	22.2	22.2	22.2

[a] Explanation: Each number in the columns represents a predicted class. Code ''1'' represents *E. timidum,* code ''2'' represents *E. infirmum,* code ''3'' represents *E. exiguum,* and code ''4'' is *E. tardum.* DT = detrended, D = first derivative, S = 25 PCA scores, 1 = univariate CART, 2 = multivariate CART (mix of uni- and multivariate rules) by Breiman, 3 = OC1 rule induction (mix of uni- and multivariate rules). FuRES is not included since it does not allow for an unpruned option in the program used. All calibrations are performed with leave-one-out cross-validation before prediction on unknown validation set. Unpruned decision trees were used here. Note that for some of the columns the pruned and unpruned tree are identical.

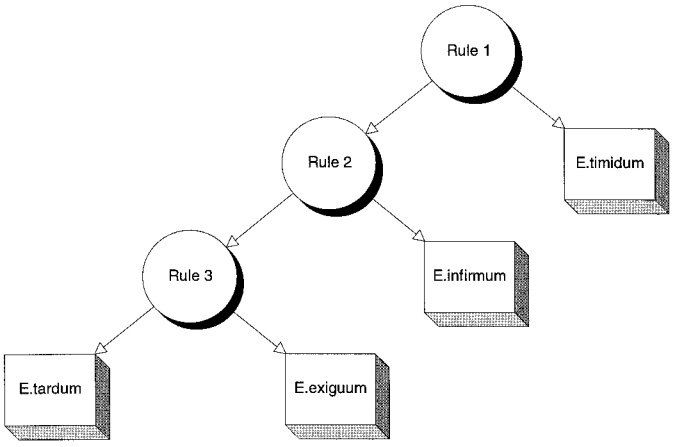

FIG. 4. Illustration of the FuRES classification tree for analysis of the detrended and differentiated spectra. The three multivariate rules referred to here as Rules 1, 2, and 3 are shown in Fig. 5. We see that Rule 1 separates *E. timidum* from the three other species; Rule 2 separates out *E. infirmum* from *E. tardum* and *E. exigum*; Rule 3 distinguishes between *E. tardum* and *E. exigum.*

has a peak around 1200 cm^{-1} that is larger than the two other rules. Rule 3 distinguishes between *E. tardum* and *E. exigum* and has a feature in the region 860–1050 cm^{-1} that is more prominent than in the two other rules.

It is important to realize that the two multivariate CART methods do not always use multivariate rules in their decision trees. A univariate rule is chosen if it is better. We have also frequently observed (unpublished results) that the multivariate rules in Breiman's and OC1 CART methods are very much dominanted by one variable; making them effectively univariate models.

For most CART methods in the OC1 program, the unpruned trees have significantly better prediction error than the pruned trees (see Table III). Since the FuRES available to us does not allow such an option, it is not included in the table. The most drastic improvement in prediction error can be seen in the detrended data set. For univariate CART and the OC1 CART, the prediction error drops by 11.2% with the use of the unpruned trees on the unseen validation set. For the differentiated spectra, the prediction error drops 8.3%. For the PCA scores representation there is no change in the prediction error between pruned and unpruned trees.

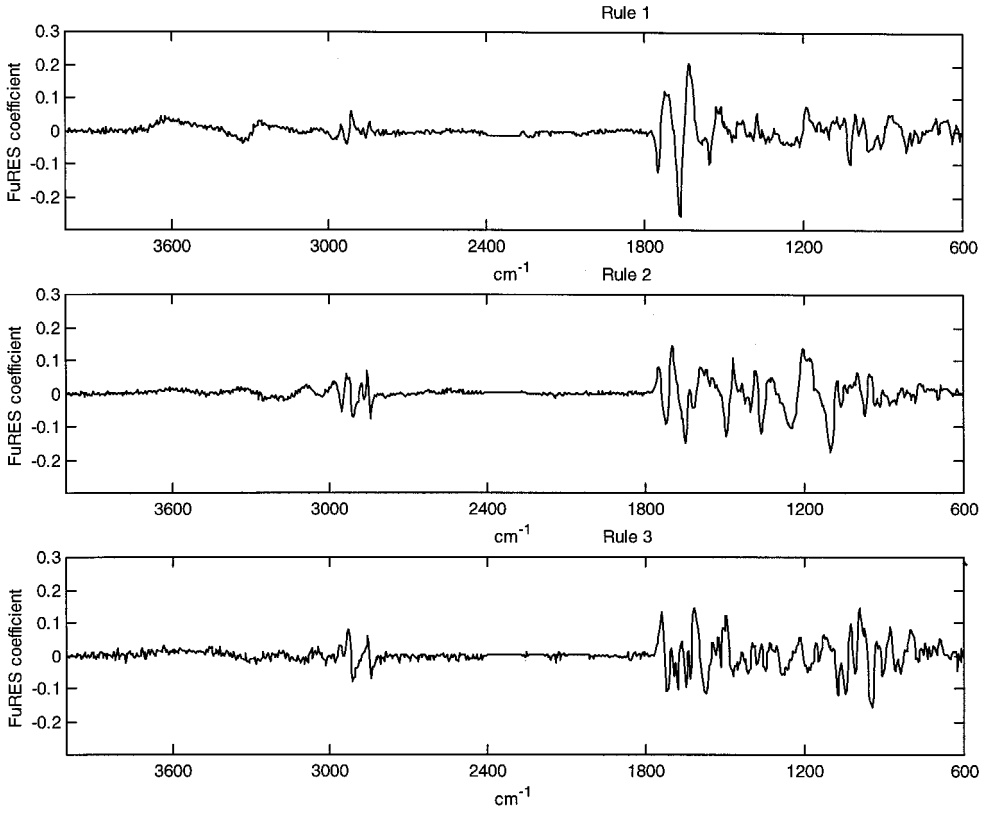

FIG. 5. The three FuRES rules (which are referred to as Rules 1, 2, and 3 in the previous figure) from the analysis of the detrended and differentiated spectra. Please see text for interpretation of these plots.

The pruning algorithm used in the OC1 program is based on a leave-one-out cross-validation approach that tends to create conservative models; that is to say, the algorithm penalizes large trees and hence is biased towards choosing smaller trees with few branches. Such small trees often have less than optimal prediction ability, and in this instance may explain why the unpruned trees give better prediction results than the pruned trees.

The most important variables selected by the CART, Breiman CART, and OC1 methods applied to the detrended data set were wavenumbers 1225, 3440, and 3595 cm^{-1}. Similarly, the methods selected wavenumbers 947, 3614, and 3174 cm^{-1} for the differentiated and detrended data set. In Fig. 1 we have plotted the mean spectra for each of the four classes together with the three most significant wavenumbers found from analysis of the detrended data set. Using a visual inspection of the four mean vectors, we find it difficult to detect features that are selective for some of the classes. In Fig. 3 we present the mean differentiated spectra for the four classes together with the three most important variables depicted as vertical bars. Again, visual inspection does not give any clear indication of any class-dependent features in the differentiated spectra.

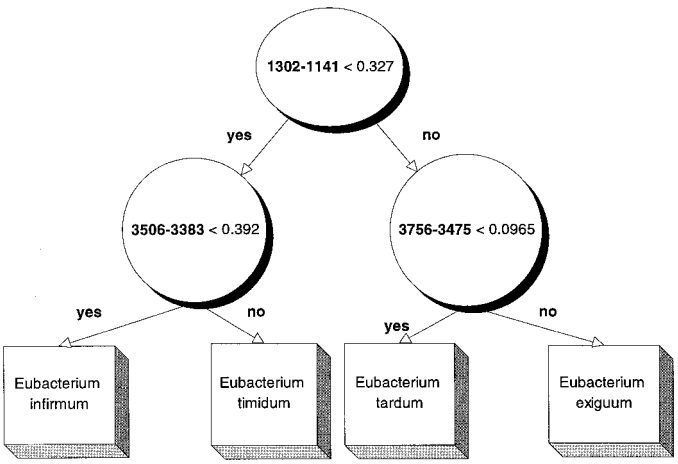

FIG. 6. The decision tree for univariate CART on B-spline coefficients. The most important B-spline functions correspond to coefficients covering variable regions in the original variable domain (882 variables). Bold numbers indicate the wavenumber domains selected by the rule induction methods.

B-splines. Since smooth curves are analyzed, the variables are highly correlated; consequently regions, rather than individual variables, will be important for the classification of the four classes. In order to analyze this aspect further, a B-spline analysis of the data set was performed where compressed B-spline coefficients were used instead of the original 882 variables. Seventy-nine regions were selected by the maximum entropy method used for finding good knot distributions.[44] The maximum local polynomial degree was set to 4. The data set containing 79 rather than 882 variables was subjected to the same rule induction algorithms as mentioned above, and the results were compared (pruned results used here). It was found that for the univariate and the OC1 CART methods (see Fig. 6) a 16.7% prediction error was achieved compared to a 30.6% prediction error with the use of the original 882 variables. No improvement was

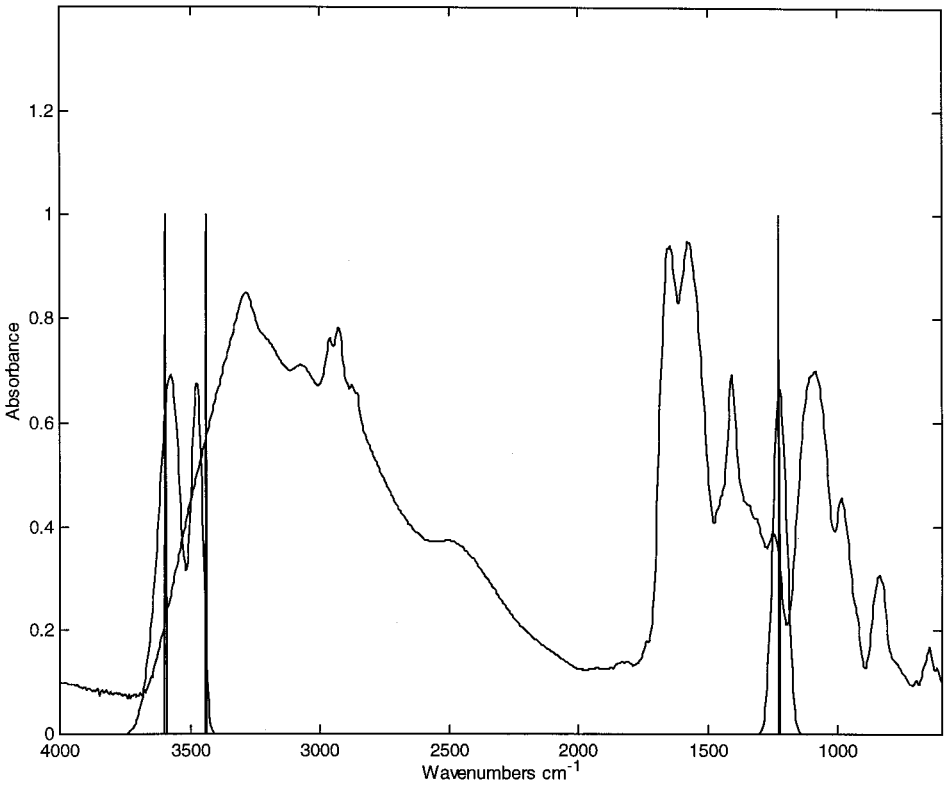

FIG. 7. A comparison of the results from using B-spline compression and detrended data in the original domain (882 variables). The compressed data set contains 79 variables where B-spline coefficients 6 (3756–3475 cm^{-1}), 9 (3506–3383 cm^{-1}), and 65 (1302–1141 cm^{-1}) were found to be important. Each B-spline coefficient is associated with a B-spline function in the original domain, and we have here plotted the three B-spline basis functions together with the three most important variables determined from analysis of the data set in the *original* domain. Note that the B-spline coefficient representation decreases the prediction error for the three CART methods.

achieved for the multivariate CART (Breiman's version). This observation suggests that numerous variables contain the same information and that a ''pooling'' procedure such as B-spline will sometimes be advantageous.[48] The CART methods applied to the B-spline representation selected the coefficients 6, 9, and 65 as the most important. Since each coefficient corresponds to a B-spline function over a finite region (compactly supported), those functions were highlighted and plotted with a representative detrended spectrum in the original data representation (882 variables) (see Fig. 7). Note that coefficient 6 covers spectral region 3756–3475 cm^{-1}, coefficient 9 covers 3506–3383 cm^{-1}, and coefficient 65 covers 1302–1141 cm^{-1}. The three most important variables found by the CART methods applied directly on the original detrended data domain are included as vertical bars. It is observed that these variables are almost located in the middle of all the B-spline function regions.

Often a compression approach in the analysis of very correlated variables produces adequate univariate rather than multivariate CART models. This approach is always desirable because of the superior interpretational advantages provided by the method.

The correlation between the variables in the original data domain is reduced in the B-spline coefficients. In a ''perfect'' compression, we would not expect any correlation between the variables.

Optimal Number of PC Scores. We have previously used PCA as a data compression method prior to ANN analysis of both pyrolysis mass spectra and infrared data and have effected the successful identification and quantifaction of (micro) biological systems.[8,49–51] In the present study the first 25 PC scores were used to capture most of the information available in the differentiated data set. In order to assess the minimum number of PC scores required for this data set to obtain optimal prediction for the various methods we used between 1 and 25 PCA scores vectors sequentially as input data to the four rule induction techniques. The results of this analysis are shown in Table IV. For the FuRES method it was seen that using more than the first four PC scores gives perfect prediction. The OC1 multivariate rule induction method has zero prediction error using five PCs, but the prediction error increases for higher factors (overfitting). Neither the univariate nor the multivariate (Breiman's) CART ever gave perfect predictions, and the prediction error was always 22.2% when more than five PCs scores were used.

The main reason why interpretation of PC scores is difficult is that each score variable is based on a projection onto linear combinations of the original variables. Even though each score variable explains the maximum variance direction, the new factors may not be easy to interpret.

CONCLUSION

Figure 2 shows that, when the full FT-IR spectra were analyzed by DFA, four distinct groups could be seen, one for each of the *Eubacterium* species; moreover the five

TABLE IV. PCA factors and prediction errors from each of the rule induction techniques on the differentiated data.[a]

Scores	CART	Breiman	OC1	FuRES	Explained PCA variance (%)
2	38.9	38.9	38.9	38.9	61.9
3	38.9	38.9	38.9	25.0	74.0
4	30.6	30.6	8.3	2.8	83.0
5	22.2	22.2	0.0	0.0	88.5
6	22.2	22.2	8.3	0.0	91.0
7	22.2	22.2	5.6	0.0	92.6
8	22.2	22.2	5.6	0.0	93.9
9	22.2	22.2	19.4	0.0	94.9
10	22.2	22.2	30.6	0.0	95.4
11	22.2	22.2	8.3	0.0	95.9
12	22.2	22.2	27.8	0.0	96.3
13	22.2	22.2	25.0	0.0	96.7
14	22.2	22.2	25.0	0.0	96.9
15	22.2	22.2	27.8	0.0	97.2
16	22.2	22.2	27.8	0.0	97.3
17	22.2	22.2	22.2	0.0	97.5
18	22.2	22.2	33.3	0.0	97.7
19	22.2	22.2	13.9	0.0	97.8
20	22.2	22.2	36.1	0.0	98.0
21	22.2	22.2	19.4	0.0	98.1
22	22.2	22.2	25.0	0.0	98.2
23	22.2	22.2	22.2	0.0	98.3
24	22.2	22.2	22.2	0.0	98.4
25	22.2	22.2	22.2	0.0	98.4

[a] This table shows how the four different rule induction methods perform using different numbers of scores vectors. The percentage error values are given in the table. Here the number of scores vectors is varied from 2 to 25. Note that five PCA scores would have been sufficient for perfect prediction both for the FuRES and the OC1 methods. Whereas the OC1 method is unstable for more than five scores, the FuRES has zero prediction error for the same region. Both univariate CART and the multivariate CART (Breiman's version) have 22.2% prediction error after five PCA scores vectors.

clinical isolates clustered with *E. exiguum*—a result found in an earlier study.[5]

We have previously exploited ANNs for the identification of these eubacteria from their PyMS;[5] however, in that study the information in terms of which masses in the mass spectrum are important was not *readily* available, and ANNs are often perceived as a ''black box'' approach to modeling spectra.

In this study, using FT-IR, the most important variables selected by the rule induction methods were the wavenumbers 947, 1225, 3174, 3440, 3595, and 3614 cm^{-1} (see Table V). It is very significant that each of the rule induction techniques highlighted a low number of similar wavenumbers that were predominantly important in separating the four *Eubacterium* species because it could now allow the deconvolution in terms of what (bio)chemical information is different in each of the bacteria and thereby could allow specific characteristic biochemical markers to be developed for each of the eubacteria.

A database search using IR Mentor Pro (Bio-Rad Laboratories, Sadtler Division) based on condensed-phase compounds in the mid-IR region (4000–400 cm^{-1}) compiled from Ref. 52 for these six wavenumber regions produced the following suggestions for the functional groups:

- 947 cm^{-1} is a likely vibration from aromatic C–H, carboxylic acid (O–H), ethers (C–O–C), or phosphorus ($P–CH_3$).
- 1225 cm^{-1} could originate from the vibrations of many groups, and, most commonly, alcohols (C–O), alkanes (C–C), amides (CHN and C–N), aromatic C–H, carboxylic acid (O–H), carbonates (O–C–O), esters (C–O), ethers (C–O–C), imides (C–N), phosphorus (P–O–Ph), and P=N.
- 3440 cm^{-1} and 3174 cm^{-1} may be from secondary amide (C–N) or from urea (NH_2) vibrations.
- 3614 cm^{-1} and 3595 cm^{-1} are vibrations from a region dominated by water vapor and hydrogen bond effects. By inspecting the raw spectra for each of the classes, it was possible to detect a correlation between the slope of the start of this peak region and the class membership. We suggest that this is the primary reason why the rule induction methods select these variables, although no reason for this phenomenon is evident to us.

From the above list it is obvious that no single biochemical species vibrates uniquely at each of the wavenumber regions selected. This is to be suspected since it is known that the IR spectra of very complex mixtures contain the vibrations of many materials superimposed on one another. Since the samples under analysis in this study are bacteria and these are known to contain, among many other things, DNA and RNA, proteins, carbohydrates, and lipids, the problem of exact peak assignment is immense.

TABLE V. Summary of rules.[a]

Data set–Method	Class 1: *E. timidum*	Class 2: *E. infirmum*	Class 3: *E. exiguum*	Class 4: *E. tardum*
detrend–CART	If **1225** < 0.265 AND **3440** > 0.3915	If **1225** < 0.265 AND **3440** < 0.3915	If **1225** > 0.265 AND **3595** > 0.0545	If **1225** > 0.265 AND **3595** < 0.0545
detrend–Breiman	If mult(**1225;1696**) < 0.1723 AND **3440** > 0.3915	If mult(**1225;1696**) < 0.1723 AND **3440** < 0.3915	If mult(**1225;1696**) > 0.1723 AND **3595** > 0.0545	If mult(**1225;1696**) > 0.1723 AND **3595** < 0.0545
detrend–OC1	If **1225** < 0.265 AND **3440** > 0.3915	If **1225** < 0.265 AND **3440** < 0.3915	If **1225** > 0.265 AND **3595** > 0.0545	If **1225** > 0.265 AND **3595** < 0.0545
diff–CART	If **947** < −0.0135 AND **3614** > 0.0075	If **947** > −0.0135 AND **3174** < −0.0045	If **947** < −0.0135 AND **3614** < 0.0075	If **947** > −0.0135 AND **3174** > −0.0045
diff–Breiman	If **947** < −0.0135 AND **3614** > 0.0075	If **947** > −0.0135 AND **3174** < −0.0045	If **947** < −0.0135 AND **3614** < 0.0075	If **947** > −0.0135 AND **3174** > −0.0045
diff–OC1	If **947** < −0.0135 AND **3614** > 0.0075	If **947** > −0.0135 AND **3174** < −0.0045	If **947** < −0.0135 AND **3614** < 0.0075	If **947** > −0.0135 AND **3174** > −0.0045

[a] This table summarizes the different rules obtained from the three different CART methods. The bold numbers correspond to the intensities observed at the selected wavenumbers (3614, 3595, 3440, 3174, 1225, and 947 cm^{-1}). diff indicates the first-order differentiated spectra. CART indicates the univariate method, Breiman is the multivariate CART suggested by Breiman, and OC1 is the multivariate CART methods included in the OC1 package. Mult(x,y) indicates a multivariate rule with two very dominating variables x and y.

Although, in this instance, the deconvolution of these spectra, in terms of which were the characteristic biochemical species, was not possible, these methods have provided significant data reduction. The original spectra contain 882 variables (wavenumbers) that can be reduced by univariate CART to three variables. While multivariate rule induction is more difficult to interpret, particularly for FuRES, Breiman's multivariate rules were dominated by only four variables.

The reason why univariate methods did worse than the multivariate methods is probably because of the strong correlations between the variables (wavenumbers). This supposition was confirmed by the results from analyses using the B-spline representation. It was observed that the univariate rule induction prediction error was significantly reduced. The reason for this result is that each compressed variable has very little correlation and more ''explanatory power'' than the original variables.

In conclusion, this is the first study that has shown that rule induction techniques can be applied successfully to the accurate identification of bacteria by analysis of their FT-IR spectra. Unlike DFA and ANNs, these regression tree methods have the additional benefit of enabling the deconvolution of the IR spectra in terms of which wavenumbers were characteristic for each of the bacterial species studied.

ACKNOWLEDGMENTS

The clinical isolates, SBH 403–481, were kindly donated by Dr. D. Murdoch. R. G. is indebted to the Wellcome Trust for financial support (Grant Number 042615/Z/94/Z). The authors thank Prof. P. de B. Harrington for kindly providing the FuRES program. B.K.A. thanks the Chemicals and Pharmaceuticals Directorate of the UK BBSRC, Glaxo Wellcome, and Bruker/Spectrospin for financial support.

1. W. G. Wade, M. A. O. Lewis, S. L. Cheeseman, E. G. Absi, and P. A. Bishop, J. Med. Microbiol. **40,** 115 (1993).
2. W. G. Wade, Microbial Ecol. Health Disease **9,** 367 (1996).
3. S. L. Cheeseman, S. J. Hiom, A. J. Weightman, and W. G. Wade, Int. J. Syst. Bact. **46,** 957 (1996).
4. S. Poco, F. Nakazawa, T. Ikeda, M. Sato, T. Sato, and E. Hoshino, Int. J. Syst. Bact. **46,** 1120 (1996).
5. R. Goodacre, S. J. Hiom, S. L. Cheeseman, D. Murdoch, A. J. Weightman, and W. G. Wade, Curr. Microbiol. **32,** 77 (1996).
6. R. Goodacre and D. B. Kell, Curr. Opin. Biotechnol. **7,** 20 (1996).
7. J. T. Magee, in *Handbook of New Bacterial Systematics,* M. Goodfellow and A. G. O'Donnell, Eds. (Academic Press, London, 1993), p. 383.
8. R. Goodacre, É. M. Timmins, P. J. Rooney, J. J. Rowland, and D. B. Kell, FEMS Microbiol. Lett. **140,** 233 (1996).
9. D. Helm, H. Labischinski, G. Schallehn, and D. Naumann, J. Gen. Microbiol. **137,** 69 (1991).
10. D. Naumann, V. Fijjala, H. Lavischinski, and P. Giesbrecht, Mol. Struct. **174,** 165 (1988).
11. D. Naumann, D. Helm, H. Labischinski, and P. Giesbrecht, in *Modern Techniques for Rapid Microbiological Analysis,* W. H. Nelson, Ed. (VCH Publishers, New York, 1991), p. 43.
12. G. J. Puppels and J. Greve, Adv. Spectrosc. **20A,** 231 (1993).
13. W. H. Nelson and J. F. Sperry, in *Modern Techniques for Rapid Microbiological Analysis,* W. H. Nelson, Ed. (VCH Publishers, New York, 1991), p. 97.
14. W. H. Nelson, R. Manoharan, and J. F. Sperry, Appl. Spectrosc. Rev. **27,** 67 (1992).
15. R. Goodacre, É. M. Timmins, R. Burton, N. Kaderbhai, A. Woodward, D. B. Kell, and P. J. Rooney, Microbiology, submitted (1998).
16. B. K. Alsberg, R. Goodacre, J. J. Rowland, and D. B. Kell, Anal. Chim. Acta **348,** 389 (1997).
17. R. Goodacre, Appl. Spectrosc. **51,** 1144 (1996).
18. R. Goodacre, D. B. Kell, and G. Bianchi, Nature **359,** 594 (1992).
19. R. Goodacre, D. B. Kell, and G. Bianchi, J. Sci. Food Agric. **63,** 297 (1993).
20. G. Glauninger, K. A. Kovar, and V. Hoffmann, Fresenius' J. Anal. Chem. **338,** 710 (1990).
21. M. B. Mitchell, Adv. Chem. Ser. **236,** 351 (1993).
22. S. P. Bouffard, J. E. Katon, A. J. Sommer, and N. D. Danielson, Anal. Chem. **66,** 1937 (1994).
23. R. Goodacre, É. M. Timmins, P. J. Rooney, J. J. Rowland, and D. B. Kell, FEMS Microbiol. Lett. **140,** 233 (1996).
24. M. K. Winson, R. Goodacre, A. M. Woodward, É. M. Timmins, A. Jones, B. K. Alsberg, J. J. Rowland, and D. B. Kell, Anal. Chim. Acta **348,** 273 (1997).
25. É. M. Timmins, S. A. Howell, B. K. Alsberg, W. C. Noble, and R. Goodacre, J. Clin. Microbiol. **36,** 367 (1998).
26. A. Savitzky and M. J. E. Golay, Anal. Chem. **36,** 1627 (1964).
27. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees* (Wadsworth, Pacific Grove, California, 1984).
28. C. E. Shannon, Bell System Tech. J. **379,** 379 (1948).
29. J. R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann Publishers, San Mateo, California, 1993).
30. P. D. Harrington and K. J. Voorhees, Anal. Chem. **62,** 729 (1990).
31. S. K. Murthy, S. Kasif, and S. Salzberg, J. Artificial Intell. Res. **2,** 1 (1994).
32. P. E. Utgoff and C. E. Brodley, ''Linear Machine Decision Trees'', Report No. Tech. Rep. 10 (1991).
33. B. A. Draper, C. E. Brodley, and P. E. Utgoff, IEEE Trans. Pattern Anal. Machine Intell. **16,** 888 (1994).
34. C. E. Brodley and P. E. Utgoff, Machine Learning **19,** 45 (1995).
35. P. d. B. Harrington, J. Chemom. **5,** 467 (1991).
36. L. A. Zadeh, *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers,* Vol. 6 (World Scientific, River Edge, New Jersey, 1996).
37. P. d. B. Harrington, Chemom. Intell. Lab. Syst. **18,** 157 (1993).
38. B. F. J. Manly, *Multivariate Statistical Methods: A Primer* (Chapman and Hall, London, 1994).
39. H. J. H. MacFie, C. S. Gutteridge, and J. R. Norris, J. Gen. Microbiol. **104,** 67 (1978).
40. W. Windig, H. L. C. Meuzelaar, B. A. Haws, W. F. Campbell, and K. H. Asay, J. Anal. Appl. Pyrolysis **5,** 183 (1983).
41. R. Goodacre and R. C. W. Berkeley, FEMS Microbiol. Lett. **71,** 133 (1990).
42. I. T. Jolliffe, *Principal Component Analysis* (Springer-Verlag, New York, 1986).
43. G. Farin, *Curves and Surfaces for Computer Aided Geometric Design: A Practical Guide* (Academic Press, San Diego, California, 1990), 2nd ed.
44. B. K. Alsberg and O. M. Kvalheim, J. Chemom. **7,** 61 (1993).
45. B. K. Alsberg and O. M. Kvalheim, Chemom. Intell. Lab. Syst. **23,** 29 (1994).
46. B. K. Alsberg, E. Nodland, and O. M. Kvalheim, J. Chemom. **8,** 127 (1994).
47. C. S. Gutteridge, L. Vallis, and H. J. H. MacFie, in *Computer-assisted Bacterial Systematics,* M. Goodfellow, D. Jones, and F. Priest, Eds. (Academic Press, London, 1985), p. 369.
48. B. K. Alsberg, J. Chemom. **7,** 177 (1993).
49. R. Goodacre, D. Hammond, and D. B. Kell, J. Anal. Appl. Pyrolysis **40/41,** 135 (1997).
50. R. Goodacre, P. J. Rooney, and D. B. Kell, J. Antimicrob. Chemother. **41,** 27 (1998).
51. É. M. Timmins and R. Goodacre, J. Appl. Microbiol. **83,** 208 (1997).
52. N. B. Colthup, L. H. Daly, and S. E. Wiberly, *Introduction to Infrared and Raman Spectroscopy* (Academic Press, New York, 1990).