PERGAMON

Phytochemistry 52 (1999) 407–417

PHYTOCHEMISTRY

Fourier Transform Infrared microspectroscopy and chemometrics as a tool for the discrimination of cyanobacterial strains

Mustafa Kansiz[a], Philip Heraud[a], Bayden Wood[a], Frank Burden[a], John Beardall[b], Don McNaughton[a,]*

[a]*Department of Chemistry, Monash University, Wellington Road, Clayton, Victoria, 3168, Australia.*
[b]*Department of Biological Sciences, Monash University, Wellington Road, Clayton, Victoria, 3168, Australia.*

Received 7 December 1998; received in revised form 26 March 1999; accepted 26 March 1999

Abstract

Fourier Transform Infrared (FTIR) microspectroscopy, in combination with chemometrics, was investigated as a novel method to discriminate between cyanobacterial strains. In total, 810 absorbance spectra were recorded from one eukaryotic and five cyanobacterial taxa spanning three genera and including two strains of one species, *Microcystis aeruginosa.* Principal Component Analysis (PCA) based classification techniques such as Soft Independent Modelling of Class Analogy (SIMCA) and K-Nearest Neighbours (KNN) were investigated. Different spectral regions using derivative spectra were investigated to find the best combinations for classification. The highest rate of correct classifications (99–100%) was achieved using first derivative spectra with a spectral region of 1800–950 cm^{-1} for both the SIMCA and KNN. A dendrogram constructed using averaged spectra of the six taxa studied showed that the two strains of *Microcystis aeruginosa* exhibited the highest degree of similarity, while the eukaryotic taxon was the most dissimilar from the prokaryotic taxa. © 1999 Published by Elsevier Science Ltd. All rights reserved.

Keywords: Cyanobacteria; Discrimination; Infrared spectroscopy; Chemometrics

1. Introduction

Cyanobacteria (blue–green algae) comprise a large and diverse group of organisms that are widely distributed in freshwater, marine and terrestrial habitats (Brock, Madigan, Martinko, & Parker, 1994). During the past three decades, cyanobacteria have become generally recognised as a major group in the prokaryotic kingdom (Carr & Whitton, 1982), with importance as a possible food and energy source and as simple organisms for the study of fundamental cellular processes (Fay, 1983). These studies are, of necessity, dependent on the ability to differentiate between cyanobacterial strains.

Cyanobacterial identification and classification has until now relied upon: (i) morphological characterisation; (ii) the identification of various chemical markers including lipids, polyamines and carotenoids; and (iii) a number of macromolecular methods including protein electrophoresis patterns, phycobiliprotein patterns, immunological studies, DNA studies and 16S rRNA gene sequence analysis. These techniques are summarised by Wilmotte & Golubic (1991).

FTIR spectroscopy has been extensively applied to the identification and characterisation of bacteria (Naumann, Helm, & Labischinski, 1991) such as *Lactobacillus* (Curk, Peladan, & Hubert, 1994), *Staphylococcus* (Helm, Labischinski, & Naumann, 1991; Helm, Labischinski, Schallen, & Naumann, 1991), *Escherichia* (Helm, Labischinski, Schallen et al., 1991), *Streptococcus* (Goodacre, Timmins, Rooney,

* Corresponding author. Tel.: +61-3990-54525; fax: +61-3990-54597.

E-mail address: d.mcnaughton@sci.monash.edu.au (D. McNaughton)

0031-9422/99/$ - see front matter © 1999 Published by Elsevier Science Ltd. All rights reserved.
PII: S0031-9422(99)00212-5

Rowland & Kell, 1996; Helm, Labischinski & Naumann, 1991; Helm, Labischinski, Schallen et al., 1991) and *Enterococcus* (Goodacre et al., 1996). These studies have shown that the technique has the ability to provide excellent discrimination between different bacterial taxa down to the strain level. Some recent studies have also attempted to differentiate and classify algae by resonance Raman spectroscopy (Wu et al., 1998) and fluorescence spectroscopy (Henrion, Henrion, Bohme, & Behrendt, 1997). FTIR spectroscopy is a routine chemical tool for the study of

Table 1
Summary of the spectra collected

Taxa	Number of spectra collected
Synechococcus sp.	108
Microcystis aeruginosa LB-A1	162
Chlorella emersonii	162
Anabaena variabilis Mon	162
Microcystis aeruginosa DTW	108
Oscillatoria sp.	108
Total	810

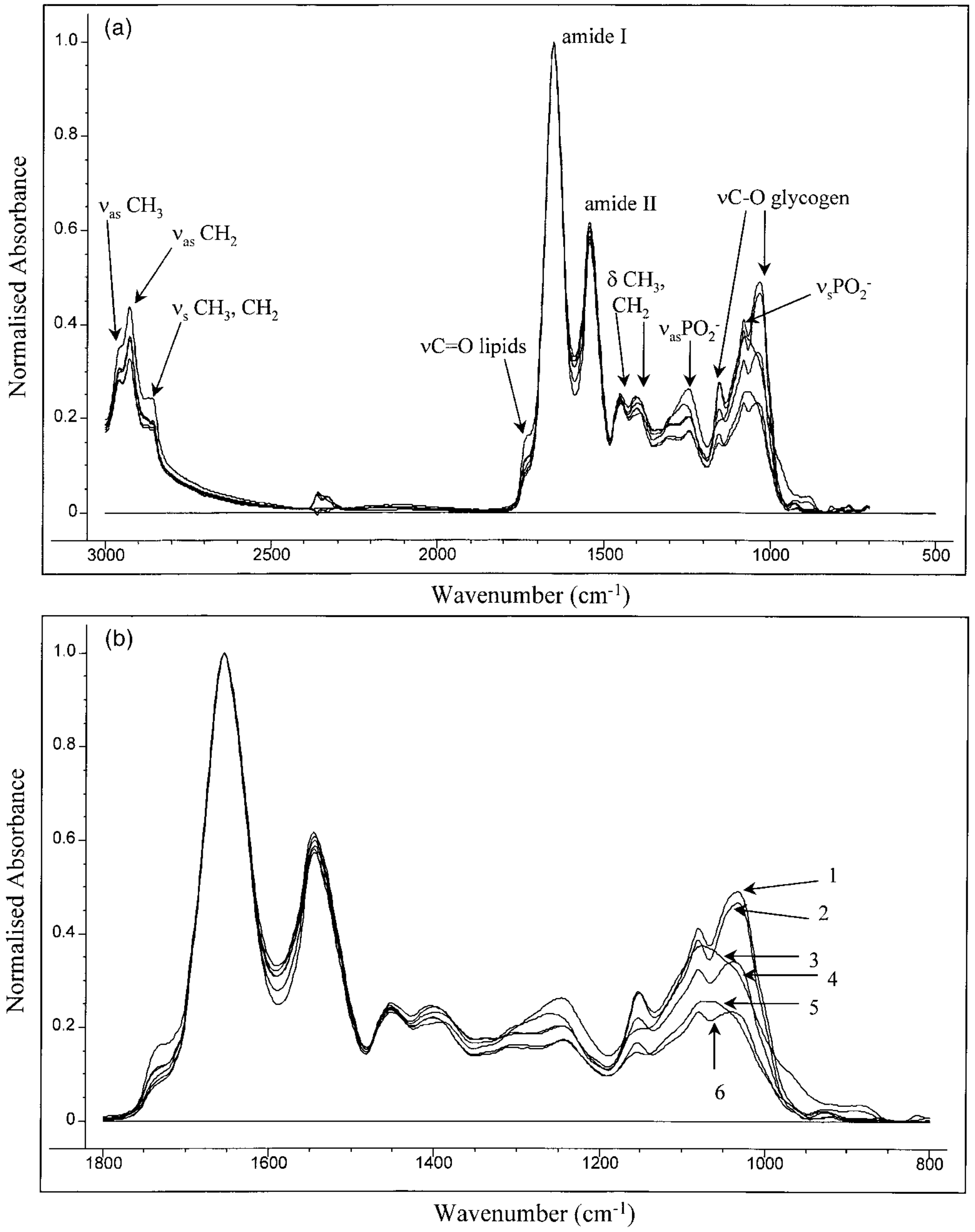

Fig. 1. (a) Averaged spectra from all six taxa. Spectra have been baseline corrected and normalised to the amide I band. Each spectrum is an average of between 108 to 162 spectra. (b) An enlargement of (a). 1 = *Microcystis aeruginosa* LB-A1, 2 = *Microcystis aeruginosa* DTW, 3 = *Chlorella emersonii*, 4 = *Synechococcus* sp., 5 = *Oscillatoria* sp., 6 = *Anabaena variabilis*.

Table 2
List of band assignments

Wavenumber (cm^{-1})	Assignment[a]	Comments
∼ 2960	ν_{as} C–H of methyl groups (Nelson, 1991)	–
∼ 2929	ν_{as} C–H of methylene groups (Nelson, 1991)	–
∼ 2875	ν_s C–H of methyl groups (Nelson, 1991)	–
∼ 2850	ν_s C–H of methylene groups (Nelson, 1991)	–
∼ 1740	ν C=O of ester functional groups primarily from lipids and fatty acids (Hedrick et al., 1991; Zeroual et al., 1995; Williams and Fleming, 1996)	–
∼ 1650	ν C=O of amides associated with proteins (Nelson, 1991; Williams and Fleming, 1996)	usually called the amide I band, may also contain contributions from C=C stretches of olefinic and aromatic compounds
∼ 1540	δ N–H of amides associated with proteins (Nelson, 1991; Williams and Fleming, 1996)	usually called the amide II band, may also contain contributions from C=N stretches
∼ 1455	δ_{as} CH_3 and δ_{as} CH_2 of proteins (Zeroual et al., 1994)	the positions of these assignments can vary in the literature
∼ 1398	δ_s CH_3 and δ_s CH_2 of proteins and ν_sC–O of COO^- groups (Nelson, 1991; Zeroual et al., 1994)	the positions of these assignments can vary in the literature
∼ 1242	ν_{as} P=O of the phosphodiester backbone of nucleic acid (DNA and RNA) (Nelson, 1991; Wong et al., 1991)	may also be due to the presence of phosphorylated proteins and polyphosphate storage products
∼ 1080	ν_s P=O of the phosphodiester backbone of nucleic acid (DNA and RNA) (Nelson, 1991; Wong et al., 1991)	may also be due to the presence of phosphorylated proteins and polyphosphate storage products
1200–900	ν C–O–C of polysaccharides (Wong et al., 1991; Zeroual et al., 1994)	the predominant polysaccharide in cyanobacteria is glycogen

[a] ν_{as} = asymmetric stretch, ν_s = symmetric stretch, δ_{as} = asymmetric deformation (bend), δ_s = symmetric deformation (bend).

molecular structure, but when applied to a large collection of intact microbial cells, the resulting spectra reflect the total biochemical composition of the cells (Naumann et al., 1991) and can thus be considered as a total and simultaneous chemical analysis. Typically, the bands are due to major cellular constituents such as proteins, lipids, polysaccharides and nucleic acids (Curk et al., 1994; Goodacre et al., 1996; Helm, Labischinski & Naumann, 1991; Helm, Labischinski, Schallen et al., 1991; Naumann et al., 1991). Consequently the IR spectra of cells produce distinct patterns that are highly reproducible and unique for different cells and it is on this basis that the differentiation between different taxa is performed. FTIR spectroscopy has distinct advantages over other conventional methods of biochemical analysis in that it is rapid, reliable and requires a relatively small sample size and simple sample preparation procedure. It is also possible to couple the spectrometer to an FTIR microscope which enables the precise selection and isolation of a desired sample area.

Chemometrics methods such as principal component analysis (PCA) and soft independent modelling by class analogy (SIMCA) have been extensively applied to the analysis of infrared spectra in a variety of different areas within the fields of medicine, biology and forensic science. These applications include the differentiation of normal and malignant cervical ‘pap’ smears (Wood, Quinn, Burden, & McNaughton, 1996), diagnosis of Alzheimer’s Disease from autopsy tissue (Choo et al., 1995), determination of the types of fruit used in jams (Derfernez, Kemsley, & Wilson, 1995; Defernez & Wilson, 1995), corn starch classification (Dupuy, Wojciechowski, Ta, Huvenne & Legrand, 1997) and fabric sample discrimination and matching (Gilbert & Kokot, 1995). PCA modelling is used to replace a complex multi-dimensional data set, such as infrared spectra, by a simplified version involving fewer dimensions which facilitates an easier analysis of the variance in the data set.

This study demonstrates that FTIR spectroscopy, in conjunction with chemometrics methods, can consistently discriminate between cyanobacterial strains.

2. Results and discussion

In total, 810 spectra (between 108 to 162 spectra from each of the six taxa) were collected from late-exponential to early-stationary phase cultures (Table 1).

The recorded spectra display the typical bands associated with biological samples. All spectra have been normalised to the strongest band: the amide I band. Band assignments are based on previous studies on whole cells, organelles and macromolecules (Hedrick, Nivens, Stafford, & White, 1991; Nelson, 1991; Williams & Fleming, 1996; Wong, Wong, Caputo, Godwin, & Rigas, 1991; Zeroual, Choisy, & Doglie,

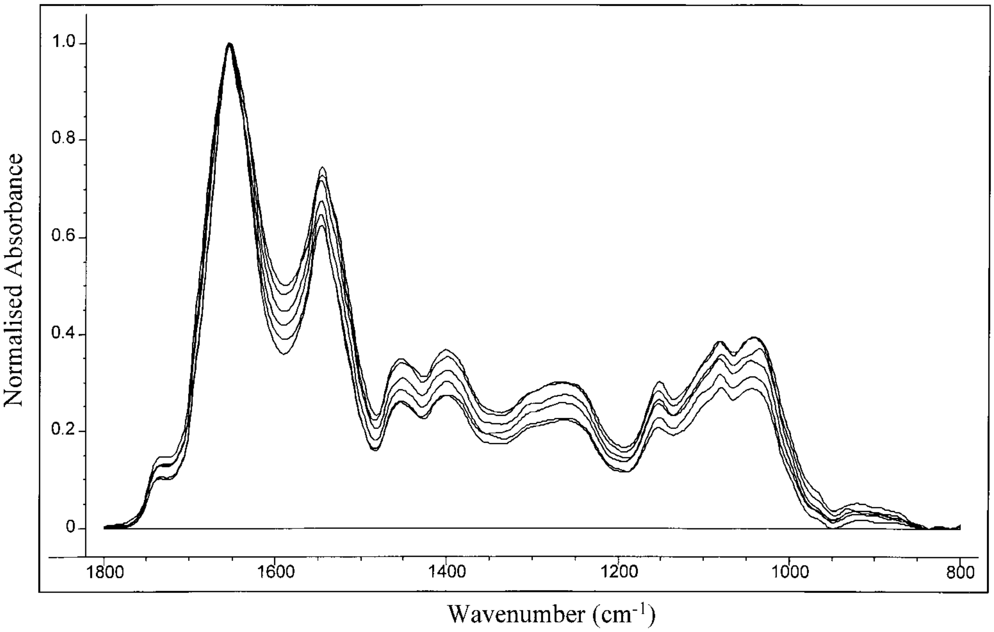

Fig. 2. Six replicate spectra from one deposit of *Synechococcus* sp. at the stationary phase.

1994; Zeroual, Manfait, & Choisy, 1995) and are detailed in Table 2. Fig. 1(a) depicts the averaged spectra (3000–700 cm^{-1}) of the six taxa studied, together with some band assignments. Fig. 1(b) is an enlargement of the region between 1800–800 cm^{-1}.

Examination of the loadings plots and raw spectra reveal that there are a number of spectral regions which can account for the chemical differences between the taxa. The most significant differences occur in the range 1200–950 cm^{-1}, which contains absorptions from the C–O–C stretching vibrations of polysaccharides, together with the symmetric and asymmetric stretching vibrations of PO_2^- functional groups at ~1080 cm^{-1} and ~1242 cm^{-1} respectively. *Chlorella emersonii* is characterised by a distinct absorption at ~1740 cm^{-1} arising from the stretching vibration of the ester C=O groups in lipids and fatty acids. This correlates with increased absorption in the C–H stretching region between 3000–2800 cm^{-1} due to the long carbon chains containing relatively large numbers of CH_2 groups per fatty acid/lipid molecule.

2.1. Spectral reproducibility

The spectral reproducibility was assessed within a single deposit, across differing deposit concentrations/thicknesses and between the triplicate cultures. The degree of variation within each taxon throughout these levels of reproducibility is important, because correct classification could be obscured if the variance within the spectra of one taxon is greater than the variance between the taxa.

Fig. 2 depicts six replicate spectra from within a single deposit of *Synechococcus* sp. taken from an early stationary phase culture. The spectra can be considered to be highly reproducible as indicated by the consistency between the relative peak intensities. However, the integrated domain (1800–800 cm^{-1}) for each spectrum does vary.

The factors contributing to such variability include incomplete baseline correction and sample-associated factors, the most important of the latter being the variation in sample thickness and the non-uniformity of the deposits leading to variations in the scattering, diffraction and refraction which occurs as the infrared beam passes through the sample.

2.2. Improving spectral reproducibility — spectral derivatives

The variation between replicate spectra as described above, can be minimised by the use of spectral derivatives. Because the spectra in Fig. 2 have essentially the same shape and contours, a derivatisation is able to highlight these and remove the differences in raw absorption. The use of such spectral derivatives is widely reported, with most classifications based on FTIR spectroscopy using spectral derivatives (Curk et al., 1994; Goodacre et al., 1996; Helm, Labischinski & Naumann, 1991; Naumann et al., 1991; Helm, Labischinski, Schallen et al., 1991). A Savitzky–Golay algorithm (Savitzky & Golay, 1964), one of the algorithms most commonly applied, was used to calculate first and second order derivatives. Fig. 3 presents the

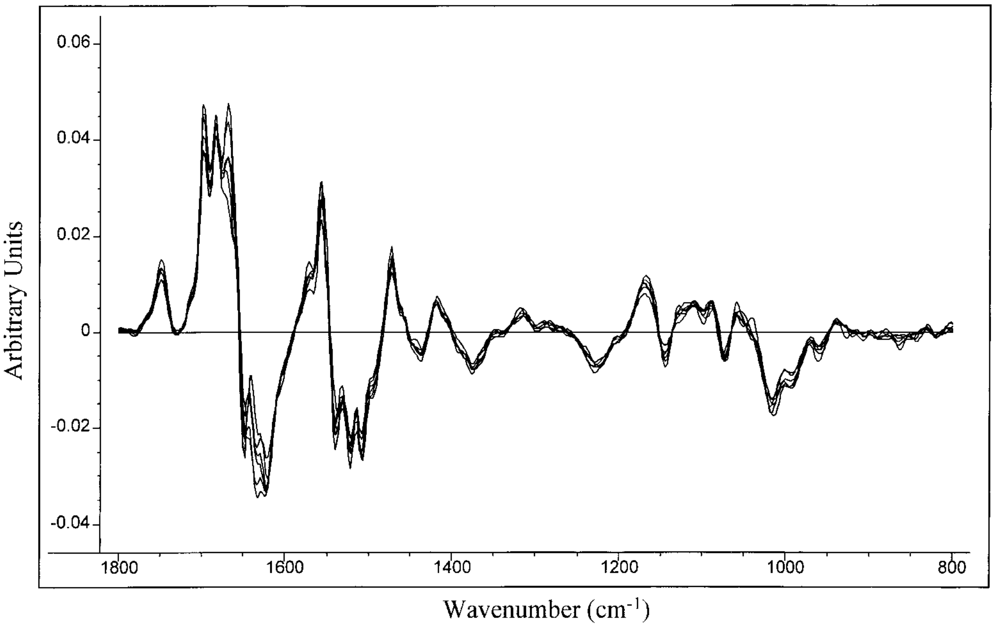

Fig. 3. Six replicate spectra from one deposit of *Synechococcus* sp. at the stationary phase. All spectra have been derivatised using a 3-point first derivative.

same six replicate spectra used in Fig. 2, after a 3-point first order derivatisation.

Upon comparison, it is clear that the variability between replicate spectra due to baseline problems is greatly reduced by the use of spectral derivatives.

2.3. Spectral changes over the growth phase

The spectra of the six taxa showed changes over the growth cycle, particularly in the polysaccharide region 1200–900 cm^{-1}. This is best exemplified by *Microcystis aeruginosa* DTW. Averaged spectra over the growth cycle from early exponential to early stationary phase are shown in Fig. 4. Spectra were taken at 5, 14, 29 and 43 days, which roughly correspond to the early, mid, late exponential and early stationary phases of growth respectively (insert in Fig. 4).

The changes in the polysaccharide region are probably attributable to glycogen, which serves as a carbon and energy reserve which is known to accumulate up to 10–20% of dry weight in exponential growth phase cells (Carr & Whitton, 1982). It can be seen in Fig. 4 that the absorbance in the 1200–900 cm^{-1} region (attributable to glycogen) increases during the exponential phase of growth, with a slight decrease as the cells enter the stationary phase. Similar spectral changes were observed in the other taxa. Although the 1200–900 cm^{-1} region was found to vary over the growth cycle, these differences were observed to diminish towards the end of the exponential phase as shown in Fig. 4. Therefore to minimise the intra-class variation, only late exponential to stationary phase spectra were used for the chemometrics analysis. All similar FTIR studies performed on bacteria (Curk et al., 1994; Goodacre et al., 1996; Helm, Labischinski & Naumann, 1991; Helm, Labischinski, Schallen et al., 1991; Naumann et al., 1991) also used cells collected from late exponential to stationary phase.

2.4. Chemometric analysis and classification

2.4.1. PCA

A mean centered[1] PCA was performed on six underivatised spectra, where each spectrum was the average of the total number of collected spectra for that taxa. The PCA loadings plot is shown in Fig. 5.

When interpreting such a loadings plot, which highlights the variables (wavenumbers) that contribute most to the total variance, it is important to note that the actual contribution to the total variance by a variable is proportional to the total variance explained by that principal component (PC). For example, variables between 1200 and 950 cm^{-1} contribute significantly more to the total variance than do other variables, demonstrated by a large loading on PC 1 for this region, which itself accounts for 68% of the total variance. The variables investigated were those from 3000–2800 cm^{-1} and 1800–950 cm^{-1}. These variables were divided into four spectral windows W1–4 (see

[1] Mean centering involves the subtraction of the average from each variable, ensuring that all results will be interpretable in terms of the variation around the mean.

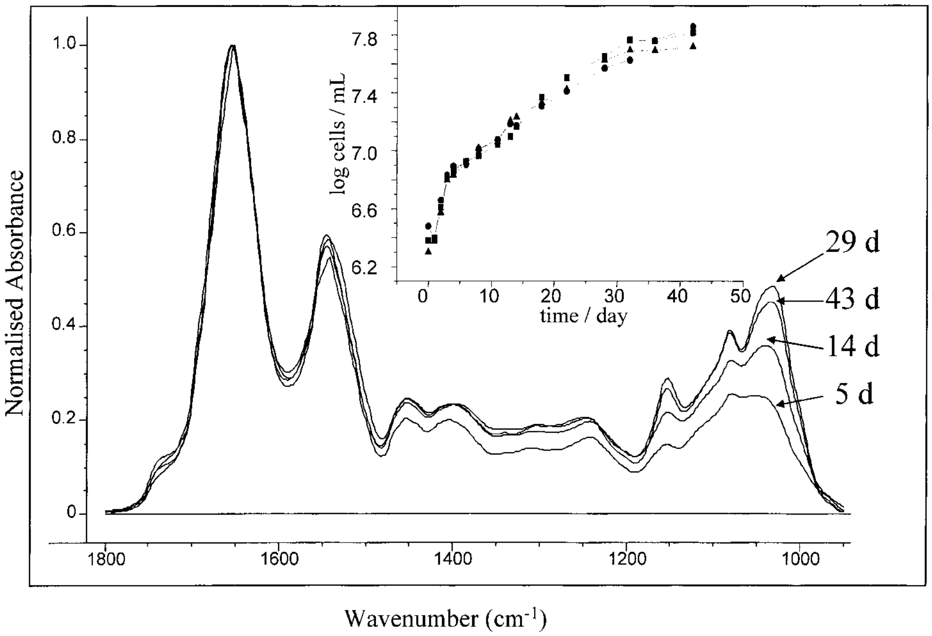

Fig. 4. Averaged spectra of *Microcystis aeruginosa* DTW over the growth phase. Insert: Growth curve of *Microcystis aeruginosa* DTW.

Fig. 5) for investigation of their individual and/or combined classification powers.

A mean centered PCA was also performed on the first derivatised (3-point) spectra of the 810 collected, in all four regions, i.e. W1–4. The PC 1 versus PC 2 scores plot from this PCA is shown in Fig. 6. A scores plot is a projection of the original data onto the new principal components, which allows the most revealing perspective for the visualisation of spatial relationship between samples, therefore allowing the easier detection of sample patterns, groupings or outliers. The scores plot in Fig. 6 displays some clustering of data, however the clusters are not distinct enough (except for *Chlorella emersonii*) to allow classification using PCA alone, indicating that a more sophisticated classification technique is required.

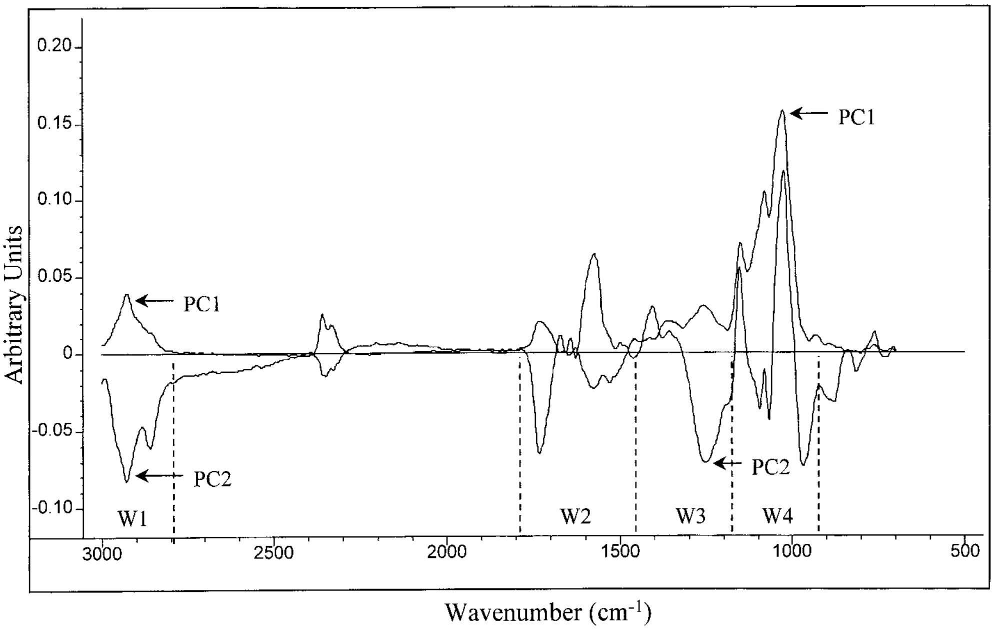

Fig. 5. Loadings plot showing PC 1 and PC 2, accounting for 68 and 23% of the total variance respectively. W1 = Window 1 (3000–2800 cm^{-1}), W2 = 1800–1480 cm^{-1}, W3 = 1480–1200 cm^{-1}, W4 = 1200–950 cm^{-1}.

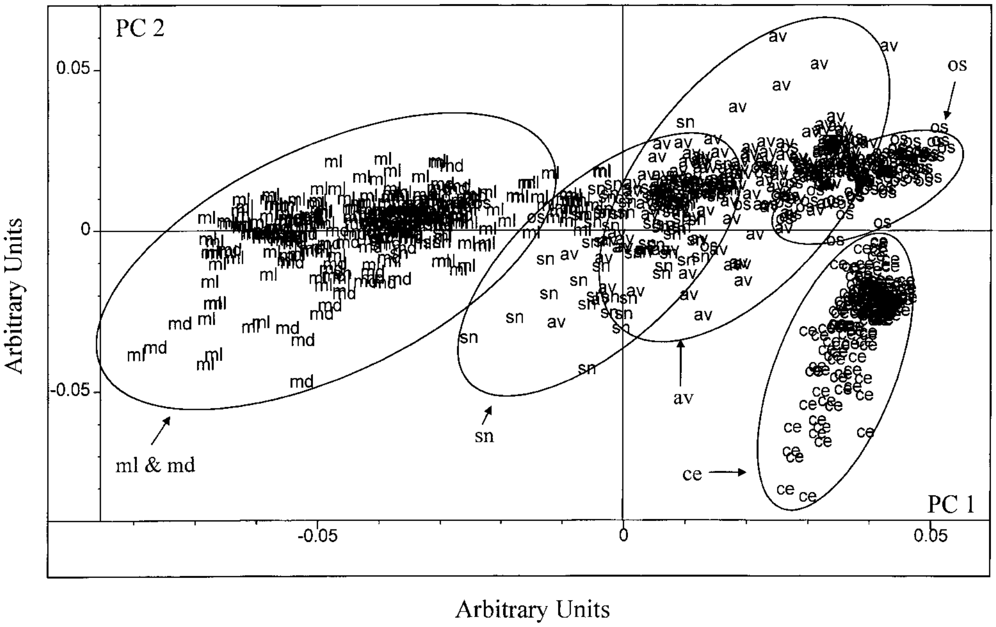

Fig. 6. PC 1 vs. PC 2 scores plot of the 810 spectra collected using the first derivative (3-point) of regions W1–4). PC 1 and PC 2 account for 39 and 12% of the variance respectively. ml = *Microcystis aeruginosa* LB-A1, md = *Microcystis aeruginosa* DTW, ce = *Chlorella emersonii*, sn = *Synechococcus* sp., os = *Oscillatoria* sp., av = *Anabaena variabilis* Mon.

2.4.2. SIMCA and KNN

The data sets for each of the classes (taxa) to be used in the SIMCA analysis were divided randomly into test and training sets containing two thirds and one third, respectively of the total data. The number of PCs required to satisfactorily describe the model was determined by finding the number of PCs for which the explained percentage variance changes by about 1%. The optimal number of PCs varied between 10 and 20 for each class modelled. New unknown samples were classified using these optimal number of PCs. The effects of using derivatives and different spectral regions were investigated.

The KNN was performed using 20 PCs with $K = 1$ through to $K = 9$, with best results coming from $K = 1$. The results of the SIMCA and KNN classifications (where $K = 1$) are presented in Table 3.

The best classification rate was achieved by applying a first (3-point) derivative on a combination of regions, referred to as W234, with a correct classification rate of 99–100%. Another combined region (W34) also had a very high rate of correct classification of 98–99%. The use of the combined regions W234 is favoured, not only because it produces the best rate of correct classification, but also because it includes more of the spectrum and hence utilises more of the chemical information in assigning the samples.

The application of a first (3-point) derivative produced better classification results consistently (except for un-derivatised spectra in W1). This can be attributed to the minimisation of the intra-class variation, resulting in tighter clustering of data points and a reduction in the overlap with neighbouring clusters in the multi-dimensional space and hence improving its classification performance. The use of even higher order derivatives, however, has produced poorer results. This can be attributed to the smoothing effect which, although reducing intra-class variation, also reduces the inter-class variation. The smoothing effect is reversed in regions with steep gradients, such as those in the W2 region, which contain the intense amide I and amide II bands. The use of higher order derivatives in such regions amplifies the signals and any small aberrations, to such an extent that they dominate the variance.

The KNN classification method generally performed with comparable reliability to the SIMCA method except in combinations which included the W2 region (particularly in conjunction with a second derivative). For example, the combination of W23 with a second derivative produced a correct classification rate of 84% for SIMCA, yet for the KNN it produced only 54%. This suggests that the SIMCA method models the data more thoroughly in these regions, because it creates a separate PCA model for each class, whereas a KNN is performed on the entire data set in the PCA reduced multi-dimensional space. By creating a separate model for each class, more information about its structure is expressed.

The averaged spectra (135 spectra), from the six replicate spectra of the individual deposits, were also analysed with a SIMCA classification, using the

Table 3
Results of SIMCA and KNN classifications. Figures in parenthesis indicate the results from a second training/test set

Spectral window	Percentage correct classification[a]					
	un-derivitised		first derivitised 3-point		second derivitised 5-point	
	KNN	SIMCA	KNN	SIMCA	KNN	SIMCA
W1 (3000–2800 cm^{-1})	63	74	72	61	70	58
W2 (1800–1480 cm^{-1})	82	84	79	86	55	68
W3 (1480–1200 cm^{-1})	84	90	93	91	75	82
W4 (1200–950 cm^{-1})	92	81	93	84	80	84
W23 (1800–1200 cm^{-1})	89	93	85	97 (95)	54	84
W34 (1480–950 cm^{-1})	94	91	98	98 (99)	88	93
W234 (1800–950 cm^{-1})	96	96 (99)	97	99 (100)	80	93 (99)

[a] Average percentage correct classification for all six taxa.

models built from the training set. The use of these averaged spectra yielded a 100% correct classification for both strains and genera when using a first (3-point) derivative in the W23, W34 and W234 regions. This result is particularly significant because, for the classification of real unknown samples, one would use the averages of the replicate spectra from each deposit, to ensure the collection of a representative spectrum.

2.4.3. Dendrograms

The dendrogram in Fig. 7 was constructed from the averaged spectra from each of the six taxa, using the best combination of derivative and region i.e. a first (3-point) derivative and region W234.

The dendrogram shows that the two strains, *Microcystis aeruginosa* LB-A1 and *Microcystis aeruginosa* DTW have been clustered the closest. The only two filamentous taxa, *Anabaena variabilis* and *Oscillatoria* sp. have also clustered closely. In addition, the only eukaryotic taxon, *Chlorella emersonii* appears to be the most distant in the dendrogram, again as one would expect, since it is eukaryotic and thus has a completely different cell structure to that of the prokaryotic cyanobacteria.

This study has demonstrated that FTIR microspectroscopy combined with chemometrics has an excellent ability to discriminate between different cyanobacterial taxa, even down to the strain level, with correct classifications of up to 100%. FTIR microspectroscopy combined with chemometrics, thus has considerable potential as an alternative discriminative method for cyanobacteria.

Apart from being a discriminative method, the technique may provide information on the phylogeny of cyanobacteria, as has been demonstrated by the many bacterial studies (Helm, Labischinski & Naumann, 1991; Helm, Labischinski, Schallen et al., 1991;

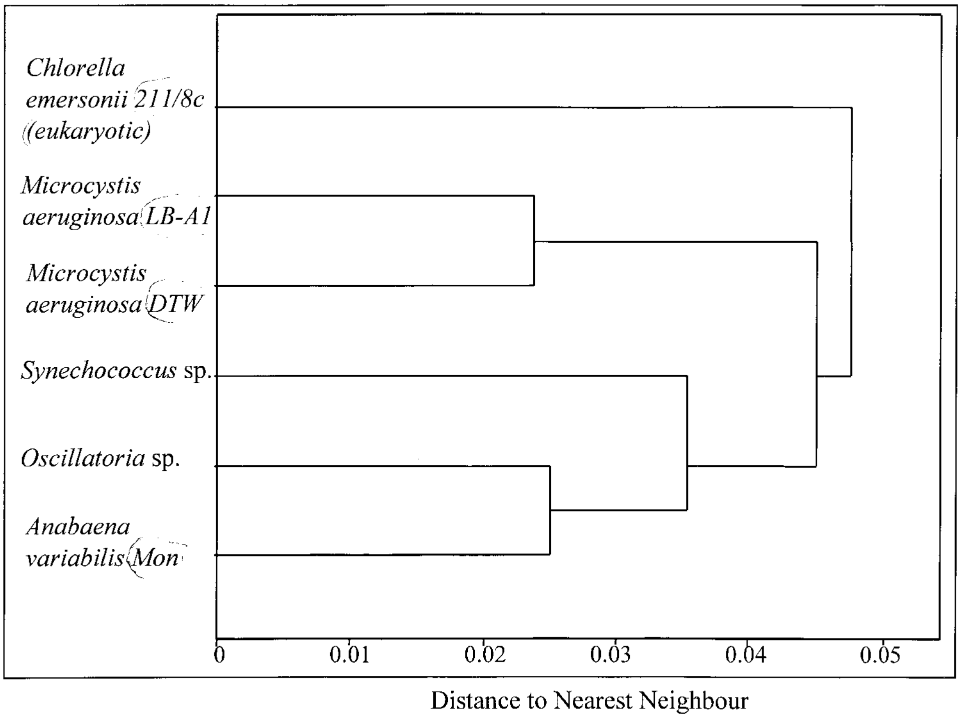

Fig. 7. Dendrogram constructed by a method of hierarchical clustering using the euclidian distance between points. The dendrogram was constructed using a first (3-point) derivative, in the W234 (1800–950 cm^{-1}) region.

Naumann et al., 1991; Curk et al., 1994; Goodacre et al., 1996). We now intend to perform a parallel study using a much larger number of cyanobacterial strains, which compares the phylogenetic relationships, determined by molecular genetic methods, with a measure of similarity derived from FTIR spectroscopic data.

3. Experimental

3.1. Organisms

Five different cyanobacterial cultures were examined. The two strains of *Microcystis aeruginosa* (LB-A1 and DTW) were provided by Ben Long from the School of Botany, La Trobe University. *Synechococcus* sp., *Oscillatoria* sp. and *Anabaena variabilis* were obtained from the Monash University Algal Culture Collection maintained by one of the authors (John Beardall). One eukaryote microalgal species was tested, *Chlorella emersonii* (strain 211/8c) which was obtained from the Cambridge Culture Collection. Cultures of all organisms used are maintained at Monash University. The cultures used were not axenic, although the levels of contaminants were low (less than 1% by microscopic examination) and would not be expected to have a significant effect on the spectra.

3.2. Media and growth conditions

All cultures were grown in 1 L volumes in 2 L conical flasks with BG-11 growth medium (Rippka, Deruelles, Waterbury, Herdman, & Stainer, 1979) at 25° in a 15 h-light/9 h-dark cycle with a photon flux density of 20 μmol photons m^{-2} s^{-1} provided by Sylvania Gro-Lux WS fluorescent tubes.

3.3. Cell deposition

Cells were deposited on to a KRS-5 (thallium-bromide–iodide) infrared transparent substrate. Preliminary trials were performed to determine the density of cell suspensions necessary to produce spectra with a good signal to noise ratio without band saturation. The required densities are shown in Table 4. The cells were also collected at various times during the exponential growth to early stationary phase.

Table 4
Approximate cell concentrations for the collection of high quality spectra. For the filamentous taxa, (*Anabaena variabilis* Mon and *Oscillatoria* sp.), optical density measurements were used instead of cell counts. Optical densities corresponded to those of the unicellular cyanobacterial suspensions

Taxa	Cell concentration (cells ml^{-1})
Synechococcus	130–600 × 10^6
Microcystis aeruginosa LB-A1	100–300 × 10^6
Microcystis aeruginosa DTW	100–400 × 10^6
Chlorella emersonii	10–40 × 10^6

The volumes required to obtain the above cell concentrations were removed from the flasks without stirring (except for the taxa that settle out under gravity) and transferred to appropriately sized centrifuge tubes. Lugol's[2] iodine solution was added, at a concentration of 10 μl per 10 ml of solution, to fix the cells, thereby preventing possible biochemical changes that might have otherwise occurred during the subsequent drying step. The tubes were centrifuged at 3000 rpm for 5 min, the supernatants removed and the resultant pellet resuspended in distilled water. This process was repeated twice to thoroughly wash the cells of growth medium which would otherwise interfere with the spectra. After the final wash, 150 μl of distilled water was left behind in the centrifuge tube and the cellular pellet resuspended in this volume. A 100 μl aliquot was pipetted onto a modified (teflon top-plate) multi-cavity KRS-5 IR cell (Wood et al., 1996). The samples were then desiccated under vacuum for several hrs. After removal of the teflon top-plate, the dried cellular deposit on the KRS-5 plate was placed on the FTIR microscope stage for spectral acquisition.

3.4. Spectroscopic methods

Spectra were collected on a Bruker IFS-55 FTIR spectrometer coupled to a Bruker infrared microscope fitted with a liquid N_2 cooled MCT (mercury–cadmium–tellurium) detector. The Bruker system was controlled by an IBM-compatible PC running OPUS version 2.2 software. The absorbance spectra were collected between 3650 and 700 cm^{-1} at a spectral resolution of 8 cm^{-1} with 10 scans co-added and averaged. A Blackman–Harris 4-term apodisation function was used, with a zero-filling factor of 2. The pin-hole aperture on the microscope was set at a diameter of 1.2 mm, with the diameter of the actual region scanned being 80 μm.

To minimise differences between spectra due to baseline shifts, the spectra were baseline corrected using the 'Rubber Band' algorithm within the *OPUS 2.2* software using 200 baseline points and excluding the CO_2 bands. The spectra were then exported as JCAMP-DX files and imported into the multivariate statistics program *Unscrambler 6.11*, running on an IBM compatible PC. The spectra were normalised within *Unscrambler 6.11* using the 'maximum' normali-

[2] Preparation of Lugol's solution: 6 parts KI is added to 100 parts water and completely dissolved before the addition of 4 parts I_2.

sation function, which normalises to the most intense band in the spectrum. The PCA and the SIMCA classifications were performed using *Unscrambler 6.11*, while the KNN classification and construction of dendrograms were performed using *MATLAB* for which the required files were imported from *Unscrambler 6.11* in an ASCII format.

3.5. Chemometrics

Chemometrics can generally be described as the application of mathematical and statistical methods to; (i) improve chemical measurement processes and (ii) extract more useful chemical information from chemical and physical measurement data (Defernez & Kemsley, 1997; Workman, Mobley, Kowalski, & Bro, 1996). The latter is particularly relevant to infrared spectroscopy which can produce large data sets of multi-dimensional complexity.

The two classification methods employed, SIMCA and KNN are based on PCA, which is one of the most commonly used multivariate statistical techniques for the extraction and interpretation of the systematic variance in a data set. The underlying idea in PCA modelling is to replace a complex multi-dimensional data set by a simplified version involving fewer dimensions, but still fitting the original data closely enough to be considered a good approximation (Adams, 1995; Brereton, 1993; Unscrambler), thus allowing the best possible view of the variance in a multivariate data set.

SIMCA (Frank & Lanteri, 1989; Unscrambler) involves the construction of separate PCA models for each class to describe and model the variation. Once each class has been modelled, provided that the classes are distinct enough, new samples can be fitted to each model. This involves the calculation of new values for all variables which are computed using the scores and loadings of each model and comparison with the actual values. The residuals are then combined into a measure of the object-to-model distance. The scores are also used to build up a measure of the sample's distance to the model centre, called leverage. Finally both object-to-model distance and leverage are taken into account to decide to which class(es) the sample belongs.

KNN (Adams, 1995; Brereton, 1993) is a similarity-based classification method which attempts to categorise unknown samples exclusively on their multivariate proximity to other samples of pre-assigned categories. It consists of drawing a circle (sphere in three dimensions and 'hypersphere' in more than three dimensions) around a point to be classified. The circle is drawn to encompass *K* nearest neighbours, where *K* is a user-set parameter. The majority vote of classes of neighbours in the circle is then assigned to the unclassified point.

Dendrograms are a graphical representation of a clustering process. One of the most commonly used clustering techniques is hierarchical clustering, which involves grouping the two most similar objects within a multidimensional space to form a cluster. The next most similar object is then grouped to form a larger cluster. This process is continued until all the objects have been grouped together by some measure of similarity, usually the euclidian distance.

Acknowledgements

This research was supported by the Australian Research Council (ARC). The authors would like to thank Ben Long from La Trobe University for providing some of the cultures and Steve Brett for his help and advice.

References

Adams, M. J. (1995). *Chemometrics in analytical spectroscopy*. Cambridge: The Royal Society of Chemists.

Brereton, R. G. (1993). *Chemometrics — application of mathematical and statistics to laboratory systems*. New York: Ellis Horwood Ltd.

Brock, T. D., Madigan, M. T., Martinko, J. M., & Parker, J. (1994). *Biology of microorganisms* (7th ed.). New Jersey: Prentice Hall.

Carr, N. G., & Whitton, B. A. (1982). *The biology of cyanobacteria*. Oxford: Blackwell Scientific Publications.

Choo, L. P., Mansfield, J. R., Pizzi, N., Somorjal R, L., Jackson, M., Halliday, W. C., & Mantsch, H. H. (1995). *Biospectroscopy*, *1*, 141.

Curk, M. C., Peladan, F., & Hubert, J. C. (1994). *Fems Microbiology Letters*, *123*, 241.

Defernez, M., & Kemsley, E. K. (1997). *Trends in Analytical Chemistry*, *16*(4), 216.

Derfernez, M., Kemsley, E. K., & Wilson, R. H. (1995). *Journal of Agricultural Food Chemistry*, *43*, 109.

Defernez, M., & Wilson, R. H. (1995). *Journal of the Science Food Agriculture*, *67*, 461.

Dupuy, N., Wojciechowski, C., Ta, C. D., Huvenne, J. P., & Legrand, P. (1997). *Journal of Molecular Structure* , *410–411*, 551.

Fay, P. (1983). *The blue–greens (cyanophyta–cyanobacteria)*. In *Studies in Biology/Institute of Biology*, *No. 160*. London: Edward Arnold Ltd.

Frank, I. E., & Lanteri, S. (1989). *Chemometrics and Intelligent Laboratory Systems*, *5*, 247.

Gilbert, C., & Kokot, S. (1995). *Vibrational Spectroscopy*, *9*, 161.

Goodacre, R., Timmins, E. M., Rooney, P. J., Rowland, J. J., & Kell, D. B. (1996). *Fems Microbiology Letters*, *140*, 233.

Hedrick, D. B., Nivens, D. E., Stafford, C., & White, D. C. (1991). *Journal of Microbiological Methods*, *13*, 67.

Helm, D., Labischinski, H., & Naumann, D. (1991). *Journal of Microbiological Methods*, *14*, 127.

Helm, D., Labischinski, H., Schallen, G., & Naumann, D. (1991). *Journal of General Microbiology*, *137*, 69.

Henrion, R., Henrion, G., Bohme, M., & Behrendt, H. (1997). *Fresenius Journal of Analytical Chemistry*, *357*, 522.

Naumann, D., Helm, D., & Labischinski, H. (1991). *Nature*, *351*(6321), 81.

Nelson, W. H. (1991). *Modern techniques for rapid microbiological analysis*. New York: VCH Publishers.

Rippka, R., Deruelles, J., Waterbury, J. B., Herdman, M., & Stainer, R. Y. (1979). *Journal of General Microbiology*, *111*, 1.

Savitzky, A., & Golay, M. J. E. (1964). *Analytical Chemistry*, *36*(8), 1627.

The Unscrambler® 6.11, User's guide. Camo ASA.

Williams, D. H., & Fleming, I. (1996). *Spectroscopic methods in organic chemistry* (5th ed.). London: McGraw-Hill International Ltd.

Wilmotte, A., & Golubic, S. (1991). *Algological Studies*, *64*, 1.

Wong, P. T. T., Wong, R. H., Caputo, T. A., Godwin, T. A., & Rigas, B. (1991). *Proceedings of the National Academy of Sciences*, *88*, 10988.

Wood, B. R., Quinn, M. A., Burden F, R., & McNaughton, D. (1996). *Biospectroscopy*, *2*(3), 155.

Workman Jr, J. J., Mobley, P. R., Kowalski, B. R., & Bro, R. (1996). *Applied Spectroscopy Reviews*, *31*(1–2), 73.

Wu, Q., Nelson, W. H., Hargraves, P., Zhang, J., Brown, C. W., & Seelenbinder, J. A. (1998). *Analytical Chemistry*, *70*, 1782.

Zeroual, W., Choisy, C., Doglie, S. M. et al. (1994). *Biochimica et Biophysica Acta*, *1222*, 171.

Zeroual, W., Manfait, M., & Choisy, C. (1995). *Pathologie et Biologie*, *43*(4), 300.