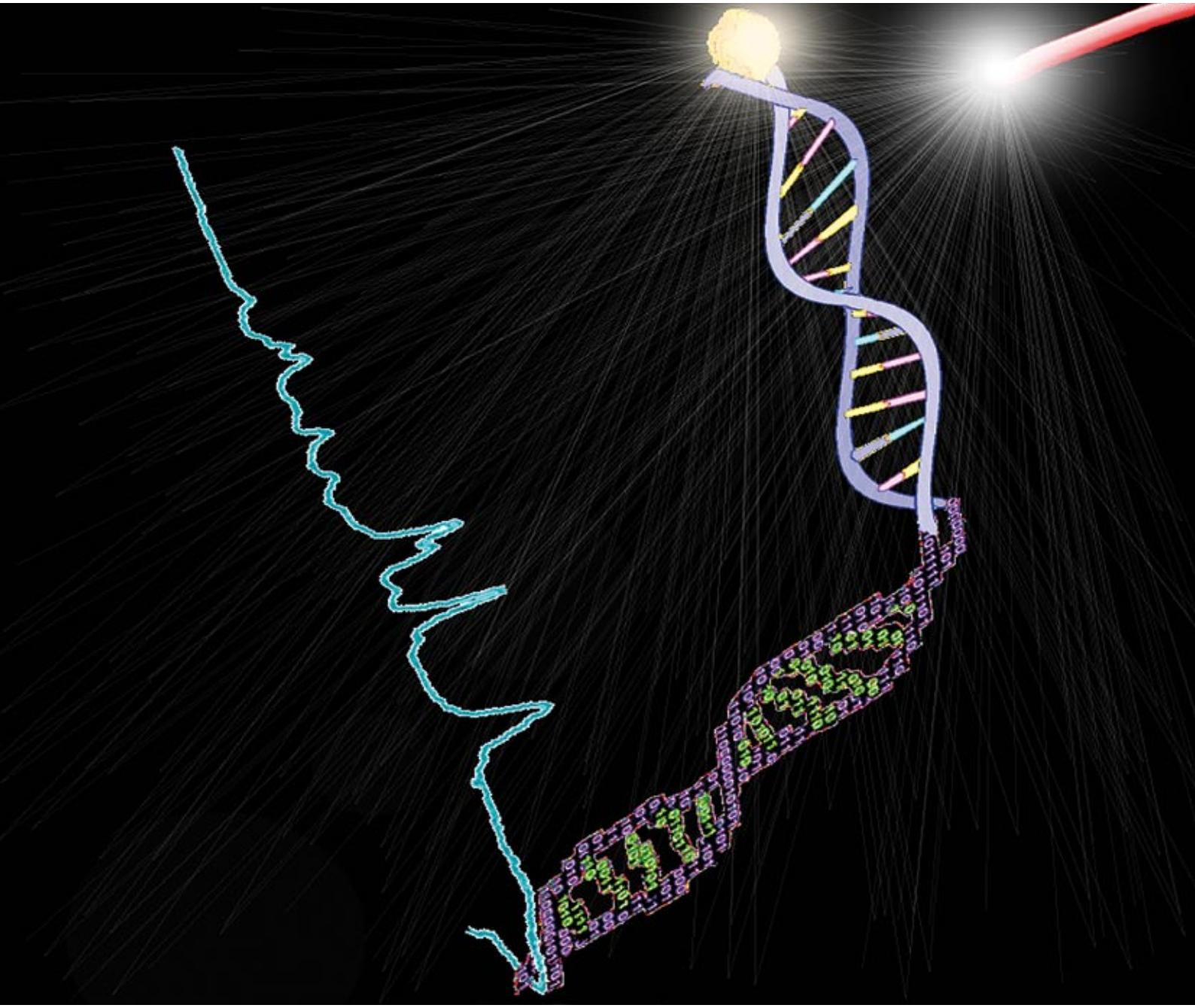
The Analyst

Interdisciplinary detection science

www.rsc.org/analyst

Volume 133 | Number 11 | November 2008 | Pages 1457–1628



ISSN 0003-2654

RSC Publishing

0003-2654(2008)133:11;1-0

Multiplexed detection of six labelled oligonucleotides using surface enhanced resonance Raman scattering (SERRS)†

Karen Faulds,*[a] Roger Jarvis,[b] W. Ewen Smith,[a] Duncan Graham[a] and Royston Goodacre*[b]

Received 10th January 2008, Accepted 27th June 2008
First published as an Advance Article on the web 28th August 2008
DOI: 10.1039/b800506k

The labelling of target biomolecules followed by detection using some form of optical spectroscopy has become common practice to aid in their detection. This approach has allowed the field of bioanalysis to dramatically expand; however, most methods suffer from the lack of the ability to discriminate between the components of a complex mixture. Currently, fluorescence spectroscopy is the method of choice but its ability to multiplex is greatly hampered by the broad overlapping spectra which are obtained. Surface enhanced resonance Raman scattering (SERRS) holds many advantages over fluorescence both in sensitivity and, more importantly here, in its ability to identify components in a mixture without separation due to the sharp fingerprint spectra obtained. Here the first multiplexed simultaneous detection of six different DNA sequences, corresponding to different strains of the *Escherichia coli* bacterium, each labelled with a different commercially available dye label (ROX, HEX, FAM, TET, Cy3, or TAMRA) is reported. This was achieved with the aid of multivariate analysis, also known as chemometrics, which can involve the application of a wide range of statistical and data analysis methods. In this study, both exploratory discriminant analysis and supervised learning, by partial least squares (PLS) regression, were used and the ability to discriminate whether a particular labelled oligonucleotide was present or absent in a mixture was achieved using PLS with very high sensitivity (0.98–1), specificity (0.98–1), accuracy (range 0.99–1), and precision (0.98–1).

Introduction

The detection of specific DNA sequences is vitally important in numerous areas of molecular biology and is key to many modern methods of disease state analysis. The most common approaches for the detection of a DNA sequence involve the detection of a specific DNA sequence within a mixture of other sequences. The majority of methods use dye labels attached to the desired sequence to allow detection after some biological manipulation which isolates the sequence of interest, or, in some cases, several sequences of interest. Polymerase chain reaction (PCR)-based methods dominate the field due to their ability to incorporate labels as part of the PCR, to increase the amount of target sequence to levels that can be detected and to reduce the complexity of hybridization when using labelled probes. The target-labelled sequence is then detected using a number of different techniques based mainly on spectroscopic analysis. Currently, fluorescent labels are most widely used in biological characterization. These labels can then be detected using fluorescence spectroscopy which offers a high degree of sensitivity with single molecule detection reported.[1] An ideal biological label for high-throughput multiplexed analysis will not interfere with the specificity of the detection system, will be robust and reproducible in performance, will provide sensitivity approaching that of a single molecule, give a linear response over a large concentration range, and be distinct from the natural response of the sample and from the output from other probes within the system. In addition, it will ideally be stable in storage, in contact with complex samples and under interrogation. These criteria form a good basis for assessing the fitness for purpose of different labelling strategies.

There are several drawbacks to using fluorescence as a detection technique. The main problem is the nature of the fluorescence emission spectrum, which is broad and gives limited characteristic information about the target analyte. This makes the detection of multiple analytes in a mixture difficult due to the broad spectral overlap that occurs from more than one fluorophore. In practice, in using a single excitation light source only four labels are generally detected at once, three if an internal standard is used unless some sort of physical separation method, such as flow cytometry, is employed. Thus, to increase the amount of data per experiment and to reduce the number of measurements required for DNA analysis, it is desirable to increase the number of DNA sequences that can be detected simultaneously without separation.

The vibrational technique of surface enhanced resonance Raman scattering (SERRS)[2–4] is ideal for the detection of multiple analytes due to the sharp fingerprint spectra which are obtained. This allows the spectral peaks to be used to

[a]*Centre for Molecular Nanometrology, WestCHEM, Department of Pure and Applied Chemistry, University of Strathclyde, 295 Cathedral Street, Glasgow, UK G1 1XL. E-mail: Karen.Faulds@strath.ac.uk*
[b]*School of Chemistry and Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester, UK M1 7ND. E-mail: Roy.Goodacre@manchester.ac.uk*

† Electronic supplementary information (ESI) available: PLS1 bootstrap results for dye labels Cy3, FAM, HEX, ROX, TAMRA and TET. See DOI: 10.1039/b800506k

differentiate multiple analytes in a mixture when the labels are carefully chosen to allow spectral separation of the components. For SERRS to occur, a molecule with a chromophore coincident or close to the laser excitation frequency is required and it must also have the ability to adsorb onto a metal surface. If the molecule does not naturally have these properties then they can be achieved by attaching a SERRS-active label to the analyte of interest. These labels can be either specifically designed for SERRS[5,6] or commercially available fluorescent labels,[7–9] and both these types of labels have been used successfully for the detection of DNA sequences.[10–12] The use of a metal surface quenches any fluorescence emitted by the label, allowing common fluorophores to be used;[13] however, since the nanoparticles used in SERRS are generally negatively charged, this means that adsorption of negatively charged labels is difficult. Modification of the DNA sequences with propargylamino-modified nucleobases allows DNA with a negatively charged label to be easily adsorbed onto a metal surface.[9] This has allowed the quantitative detection of 12 commercially available dye labels attached to oligonucleotide sequences.[8,14] SERRS has also previously been shown to be generally three orders of magnitude more sensitive than fluorescence for the detection of dye-labelled oligonucleotides.[15]

The multiplexing potential of SERRS has already been reported, where this approach was developed for the multiplex genotyping of the mutational status of the cystic fibrosis gene using two different labels, HEX and R6G.[16] Another format that has been used is that of lab-on-a-chip. In this example, microfluidics chips were generated from PDMS, and DNA sequences labelled with Cy3, FAM and TET were introduced into the chip, with the SERRS signals being measured at a point further down the channel.[17] A microfluidics approach allowed the simultaneous detection of three different DNA sequences corresponding to different strains of the *Escherichia coli* bacterium. When a SERRS multiplex is carried out using three carefully chosen labels it is possible to easily distinguish between the analytes by eye. However, when four or more labels are used, or when the concentrations of the labels are varied, identification of the substituents becomes much more complicated, and it is generally not possible by simple visual inspection since there is considerable overlap in the Raman bands. By using more than one excitation wavelength the amount of labels which can be distinguished visually can be increased, and using this approach a 5-plex DNA analysis has been performed.[18] However, using a single excitation source it quickly becomes necessary to use chemometric techniques to identify which labels are present when the number of labels is increased.

In this approach, rather than looking for specific Raman bands which may be discriminatory, a multivariate analysis (MVA) approach is adopted, which has been used widely in a multitude of vibrational spectroscopy studies.[19–21] In MVA, the whole of the SERRS spectrum is considered and each Raman scatter is thought to constitute a different dimension, such that if there are *n* variables (Raman scatters) each object (thing measured) may be said to reside at a unique position in an abstract entity referred to as *n*-dimensional hyperspace.[22] In a typical SERRS spectrum there will be hundreds of variables and hence hundreds of dimensions; this hyperspace is therefore difficult to visualize or use for predictive modelling. The underlying theme of MVA is thus *simplification* or dimensionality reduction. This dimensionality reduction occurs broadly in one of two ways: either using an unsupervised algorithm to summarize the natural variance in the data, or by using supervised learning *via* partial least squares (PLS) regression or artificial neural networks which is targeted based upon the experimenters' *a priori* knowledge of the samples being studied.[23–26] In the first case, variance due to experimental error may mask the interesting differences relating to the hypothesis being studied. If this is the case, then supervised methods, which need to be used with suitable model validation steps, can offer a targeted approach to studying spectral variance which correlates with the patterns expected from the data.

Therefore, the aim of the present study was to develop SERRS with suitably robust discriminant PLS, validated using cross-validation by iterative random resampling, for the detection of up to six labelled oligonucleotides in a mixture.

Experimental

Labelled oligonucleotides

The labelled oligonucleotides were purchased from Eurogentec (Hampshire, UK) and were HPLC purified. The oligonucleotide sequences used corresponded to probes for specific sequences corresponding to *E. coli* O157:H7′ and are detailed in Table 1.

Silver nanoparticle preparation

A colloidal suspension of citrate-reduced silver nanoparticles was prepared using a modified[27] Lee and Meisel procedure.[28]

Instrumentation

A Renishaw Model 100 probe system with a 514.5 nm argon ion laser, utilizing a long working distance ×20 objective and giving 3.5 mW of laser power at the sample, was used. The beam was focused into a 1 cm plastic cuvette containing the sample.

Table 1 Dye labels and the sequence of each *E. coli* oligonucleotide strand and the absorbance maxima of the dye label are given, where T* is 5-propargylamine-2′-deoxyuridine

5′-Dye label	Oligonucleotide sequence	λ_{max} of dye label/nm
HEX	Int1: 5′ T*CT*CT*CT*CT*CT*CGGGCGCTCATCATAGTCTTTCTTA 3′	535
TAMRA	VT1: 5′ ATAAATCGCCATTCGTTGACTAC 3′	565
ROX	VT2: 5′ GCGTCATCGTATACACAGGAGCAG 3′	585
Cy3	VT1: 5′ T*CT*CT*CT*CT*CT*CATAAATCGCCATTCGTTGACTAC 3′	552
FAM	GP1: 5′ T*CT*CT*CT*CT*CT*CCCCCACTGCTGCCTCCCGTAG 3′	494
TET	EB1: 5′ T*CT*CT*CT*CT*CT*CGAAGGTCCCCCTCTTTGGTCTTGC 3′	521

Sample preparation

Multiplexing was carried out using six dye-labelled oligonucleotides. The labels used were TAMRA, ROX, HEX, TET, FAM and Cy3. The multiplex samples were all prepared using initial stock solutions of labelled oligonucleotides that were prepared to be at a concentration of 10^{-7} mol dm^{-3}. The multiplex samples were then prepared by making solutions containing every possible combination of the six labelled oligonucleotides, resulting in 64 samples. In the multiplex sample mixtures, water was used to replace missing oligonucleotides in the matrix samples, thus allowing the overall concentration of the labelled oligonucleotides in each sample to remain the same. The final concentration of each oligonucleotide in the multiplex sample was 1.92×10^{-9} mol dm^{-3}. All samples were prepared for SERRS analysis using the following amounts of reagents: 60 μl of dye-labelled oligonucleotide, 10 μl of spermine tetrahydrochloride (0.1 mol dm^{-3}, Sigma-Aldrich), 190 μl of distilled water and 250 μl of citrate-reduced silver nanoparticles. The samples were analyzed within 1 min of the addition of the colloid and spermine and five replicates of each multiplex concentration were prepared and analyzed in a random fashion. The spectra obtained were the result of a 1 s accumulation time with the spectrometer grating centered at 1400 cm^{-1}. Each sample was represented by a SERRS spectrum containing 574 points and spectra were displayed in terms of the Raman scattered photon count (see Fig. 1 for examples).

Data analysis

For spectral analysis, both exploratory discriminant analysis and supervised learning, using cross-validated partial least squares (PLS) regression by random resampling, were employed as detailed below. As the concentration of the six labelled oligonucleotides was equivalent the data were analysed with no scaling and no smoothing.

Discriminant analysis involved the use of principal components-discriminant function analysis (PC-DFA) as detailed elsewhere.[26] Briefly, principal components analysis (PCA[30]) was used to reduce the dimensionality of the multivariate SERRS data whilst preserving most of the variance.[27] Following this process, discriminant function analysis [DFA; also known as canonical variates analysis (CVA)] was programmed to discriminate between groups on the basis of the

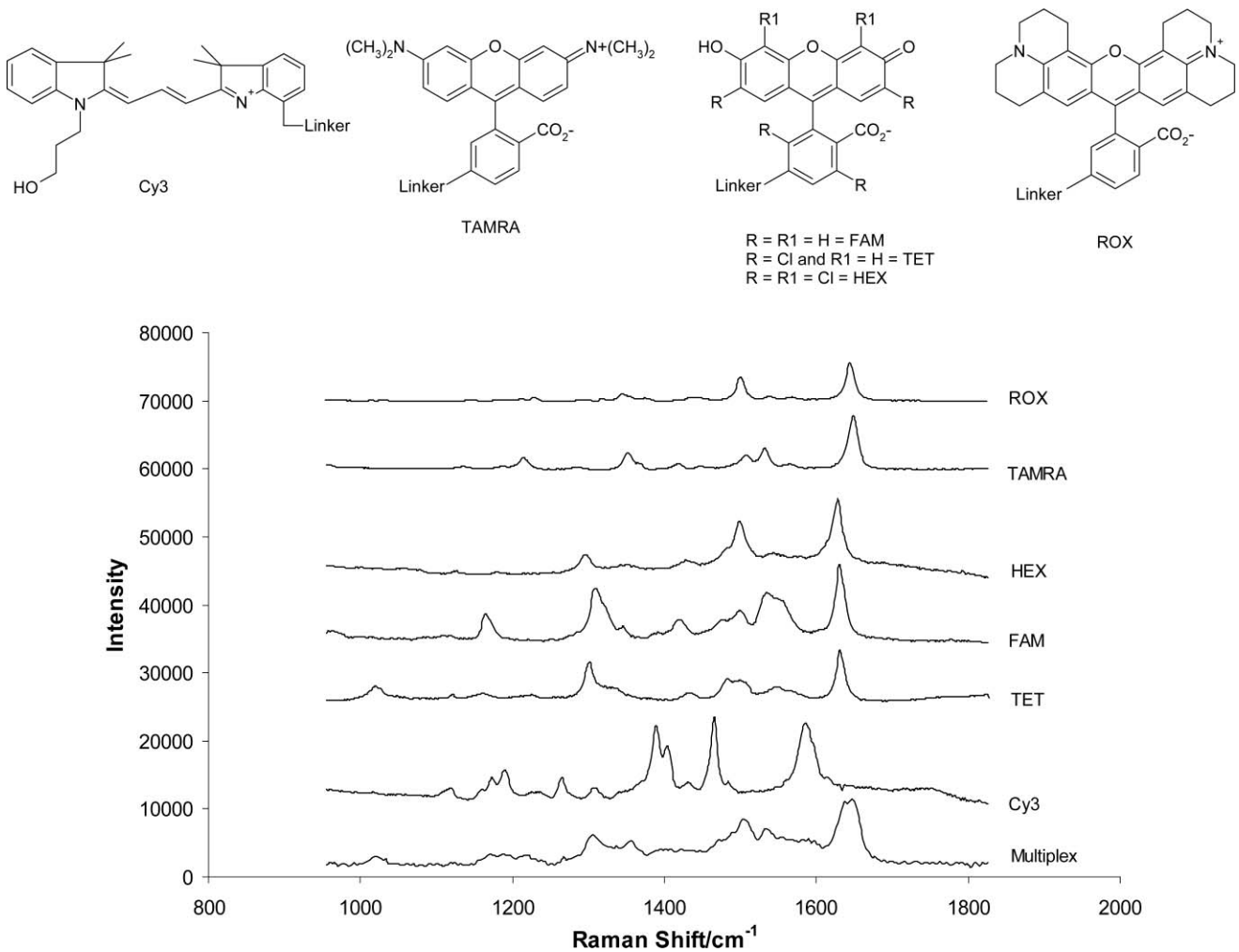


Fig. 1 Structures and SERRS spectra of the six dye-labelled oligonucleotides used in the multiplexing study (Cy3, TAMRA, ROX, FAM, TET and HEX); also shown is a multiplexed mixture containing all six dyes. For visualization purposes the spectra have been baseline corrected and offset on the *Y*-axis. The intensity scale has been left to show that each dye has a Stokes Raman shift maximum intensity of *ca.* 10 000 photons.

first 20 principal components (which accounted for 99.99% of the explained variance) and the *a priori* knowledge of which spectra were machine replicates, and thus this process does not bias the analysis in any way.[21,29]

When the desired responses (targets; *i.e.* dye label) associated with each of the inputs (SERRS spectra) are known then a supervised learning approach can be used. The goal of supervised learning is to find a model that will correctly associate the inputs with the targets. The outputs of the supervised models were encoded as '1' for when the dye was present and '0' when it was absent. PLS regression was used following the pseudocode given in Martens and Næs's *Multivariate Calibration*.[31] PLS1 was employed instead of PLS2 (which allows for multiple *Y*-constituents or outputs), since this algorithm results in a lower residual error for model predictions typically. In this case six PLS1 models were constructed as there were six constituents, one for each dye, these being Cy3, FAM, HEX, ROX, TAMRA, and finally TET. In addition, a random resampling cross-validation approach[32] was used to assess the predictive ability of PLS1 with respect to the calibration of each dye compound, to ensure that the models would be able to generalize accurately. The cross-validation exercise involves iterative random resampling of the data to generate a series of training and cross-validation samples that are assessed against the chosen model (in this case PLS1). Therefore, each sample can be used in either the training or test sets multiple times, and a statistical distribution of model performance based upon the aggregation of these sub-models can be achieved. This can allow for the approximation of confidence intervals and calculation of other statistics, such as model skewness, and also gives a more robust indication of model accuracy. We performed 200 PLS-resampling iterations as detailed below across six models, predicting the presence of one of the six dyes present in each mixture. Using this approach we were able to demonstrate the classification performance of these models, using a variety of methods described below.

All methods were implemented in MATLAB (The Math Works, Natick, MA, USA), and are compatible with the R2007a release of the software, at the time of writing.

Results and discussion

A range of labelled oligonucleotides were assessed for their SERRS signals, and ultimately six were selected for this study. The six labelled oligonucleotides chosen for multiplexing were picked to have approximately similar SERRS intensities per label and all involved electrostatic attractions as the means of adsorption onto the Ag nanoparticles. The multiplex samples were prepared in such a way that samples were produced that contained every possible combination of the six labelled oligonucleotides, which equates to 2^6 combinations resulting in 64 samples. This allowed the prediction of whether a particular label was present or not in the multiplex mixture.

As stated above, SERRS requires a molecule with a chromophore which has an absorption maximum close to the excitation wavelength of the incident laser light. Also, the molecule must come into contact with, or be very close to, the metal surface used for enhancement. This study uses citrate-reduced silver nanoparticles which have a net negative charge in aqueous solution due to a layer of citrate that exists on the surface of the silver particles.[27] Since DNA is overall negatively charged, due to the phosphate groups present in the DNA backbone, it is unable to adsorb efficiently onto the surface of the silver colloid. However, the polyamine spermine hydrochloride has been shown to interact with the DNA backbone and neutralize the charges.[33] This has been used previously in aiding oligonucleotides to adsorb onto silver nanoparticles and act as an aggregating agent when in excess.[6]

Since DNA does not contain a chromophore it is necessary to add a dye label. The labels used here were chosen as they are all commercially available and routinely used in the fluorescence detection of oligonucleotides. Of the six dye labels chosen, two have a net positive charge in aqueous solution (ROX and TAMRA; see Fig. 1) and no further modification was required to allow oligonucleotides modified with these labels to attach to the negatively charged silver surface. However, three of the labels have a net negative charge in aqueous solution (HEX, FAM and TET; see Fig. 1) and therefore further modification of the oligonucleotide was required for effective surface adsorption to occur. Propargylamino-modified deoxyuridine residues were used with negatively charged dye labels as previously reported[7] and involve the addition of six modified nucleobases at the 5′-terminus. In aqueous solution the primary amino groups are protonated giving a positive charge, allowing the DNA to adsorb onto the surface of the negatively charged nanoparticle. Therefore, the action of spermine combined with either a positively charged dye or a negatively charged dye and the propargylamino-modified bases allows good adsorption of DNA and hence successful SERRS can be obtained. The Cy3-labelled oligonucleotide was also labelled with propargylamino-modified bases to aid in its adsorption onto the negatively charged surface. The spectra obtained for each of the six dye-labelled oligonucleotides are given in Fig. 1. The full conditions for the optimisation of the SERRS conditions to allow quantitative detection of these labelled oligonucleotides have been published elsewhere and will not be discussed here.[8]

As can be seen from Fig. 1, each of the individual dye-labelled oligonucleotides generally have different SERRS spectra with the exception of ROX and TAMRA which are quite similar. When all six dye-labelled oligonucleotides are mixed and analysed by SERRS the result (Fig. 1) is a compound spectrum of all six chromophores. This spectrum is multivariate in nature and is difficult to interpret objectively by eye; the same is true for all of the other spectra of mixed oligonucleotides (data not shown), and thus for interpretation of which of the six dye-labelled oligonucleotides are present or absent in multiplex samples these SERRS spectra will necessitate analysis by multivariate statistical techniques.

In order to assess if there were any obvious features in the SERRS spectra that could be used to identify the six dyes we used principal components-discriminant function analysis (PC-DFA) to observe the relationships between the dye-labelled oligonucleotides. The SERRS spectra that were collected were coded so as to give 64 classes: one for each of the possible 2^6 dye-labelled combinations. The resulting ordination plots for the first discriminant function plotted against the second DF are shown in Fig. 2: in this figure the same DF1 *versus* DF2 is plotted with the six different labels to ease interpretation. It can be seen

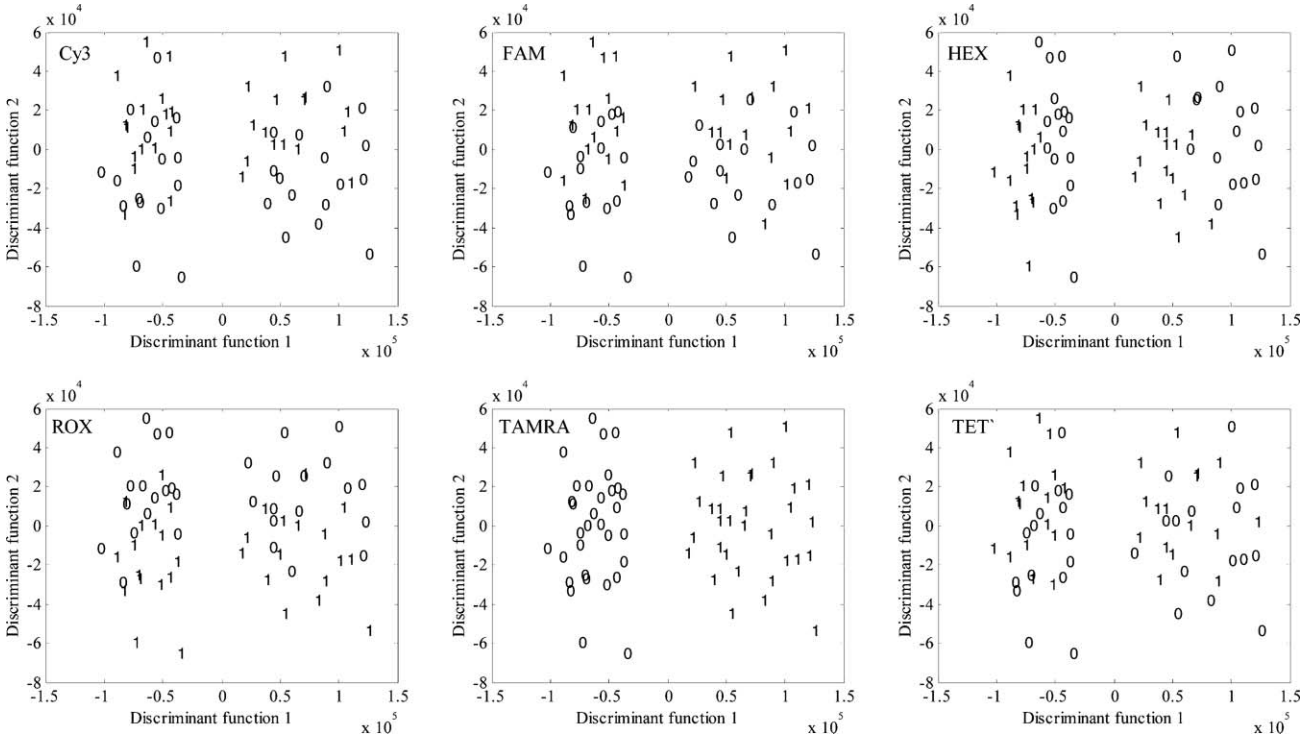


Fig. 2 PC-DFA ordination biplots of DF1 *versus* DF2. In these plots the different panels refer to whether each dye was present ('1') or absent ('0'). The first 20 PCs are used by DFA (which account for 99.99% of the explained variance) with the *a priori* class structure of the individual samples (64 classes) and not whether a dye was present or not.

from the PC-DFA plot that TAMRA is easily separated out in the first DF. The DFA algorithm is programmed to separate samples based on their class membership, and by definition DF1 is the most important ordinate generated by the DFA algorithm and each subsequent DF is ranked in order of importance to class separation; note here that 64 and not two classes were used so this PC-DFA plot is showing natural differences amongst the 64 samples analysed. Inspection of the other labels on the plots suggests that there may be some separation of ROX in DF2, and perhaps some separation of HEX away from non-HEX-labelled oligonucleotides in DF1 from those ±TAMRA.

Inspection of lower DFs shows that, again with an optimistic viewpoint, most of the dyes show some separation in a combination of DFs (Fig. 3). For example, oligonucleotides labelled with Cy3 are found in the upper left of DF 2 *versus* DF 3, whilst FAM are recovered in the lower right of DF2 *versus* DF4, and ROX in the lower left of the same plot. Clearly this is subjective, and a more objective robust chemometric approach is required; however, it demonstrates that the SERRS spectra were information-rich and contained some specific features about each of the particular dyes.

As discussed above, when the desired targets (*i.e.* the different dye-oligonucleotide labels) associated with each of the SERRS spectra are known, then the system may be supervised. The goal here is to teach a machine learning algorithm to associate a SERRS spectrum with whether it contains one of six dye-labelled oligonucleotides or not. Since five replicate SERRS spectra were collected from each of the 64 samples [with the exception of sample 51 (which contained FAM, ROX, TAMRA and TET but not Cy3 and HEX), for which only four spectra were obtained due to a single outlying spectrum which had to be excluded from the analysis], the resampling routine was programmed to select all five replicate spectra for each condition as either training or test samples in order to avoid model overtraining, *i.e.* we did not allow machine replicates to be in both the training and test data. The 200 PLS1 models were calibrated iteratively with a randomly selected training set from 44 of the possible 64 dye combinations and cross-validated by using the remaining 20 samples as a test subset, and for each resampled model the optimal number of latent variables was selected using the minimum error for the test predictions.

The results of PLS1 resampling for TAMRA-labelled oligonucleotides or DNA labelled with ROX are shown in Fig. 4 and Fig. 5, respectively. These two dyes were chosen as examples since TAMRA could be easily identified from PC-DFA and ROX was the hardest to detect in all of the samples using PLS1 (Table 2) (the remaining dyes can be viewed in ESI†). In each figure two plots are shown which seek to illustrate the reproducibility, robustness and accuracy of these models. These are a boxplot summarizing the PLS predictions for *all* training and test samples across the 200 resampled models: the boxes have lines at the lower quartile, median, and upper quartile values; the whiskers are lines extending from each end of the boxes to show the extent of the rest of the data, and outliers are marked by crosses. Secondly, a contingency matrix is shown giving the number of true positive (TP), false positive (FP), true negative (TN), false negative (FN) classifications based upon the mean of PLS1 test predictions for each sample across the 200 sub-models. This is based upon a hard classification boundary of $\geqslant 0.5$ (positive), <0.5 (negative). Some common metrics that

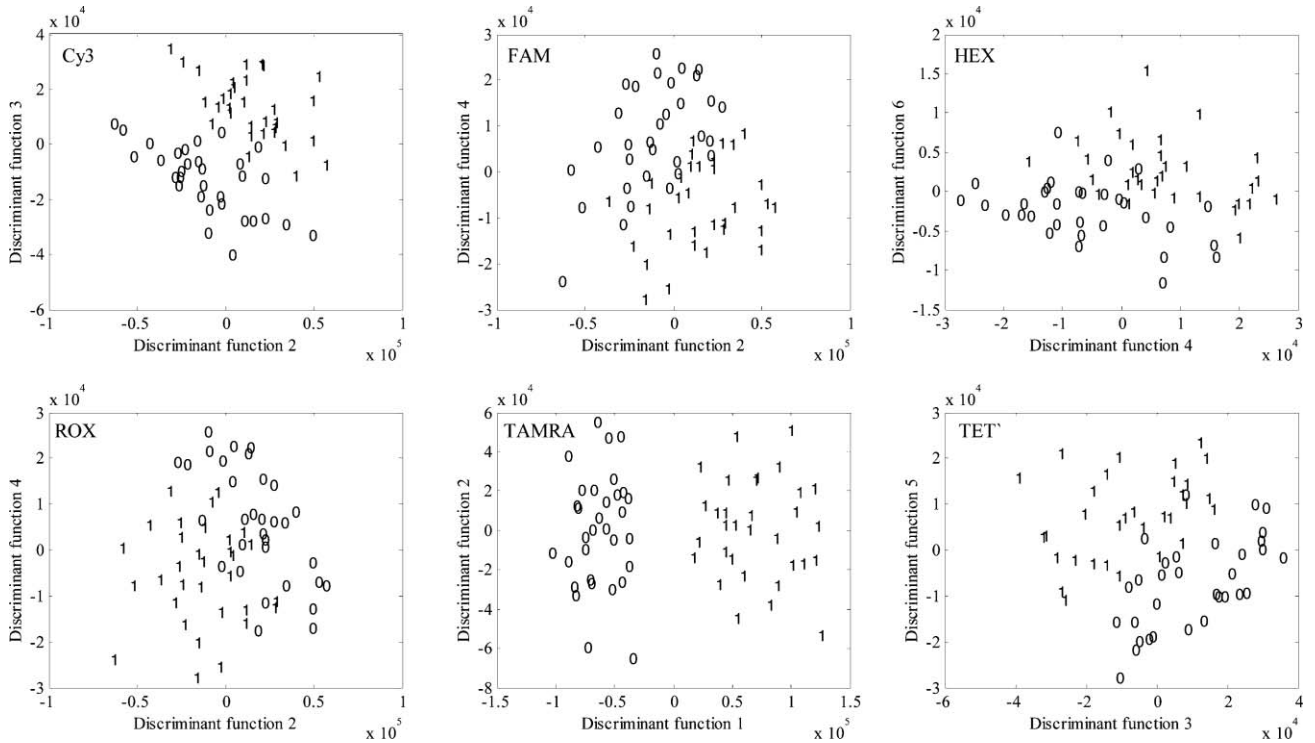


Fig. 3 PC-DFA ordination biplots of the selected DFs. In these plots the different panels are selected by eye to show the best separation between each of the dyes. The coding is whether a dye was present ('1') or absent ('0'). The first 20 PCs are used by DFA (which account for 99.99% of the explained variance) with the *a priori* class structure of the individual samples (64 classes) and not whether a dye was present or not.

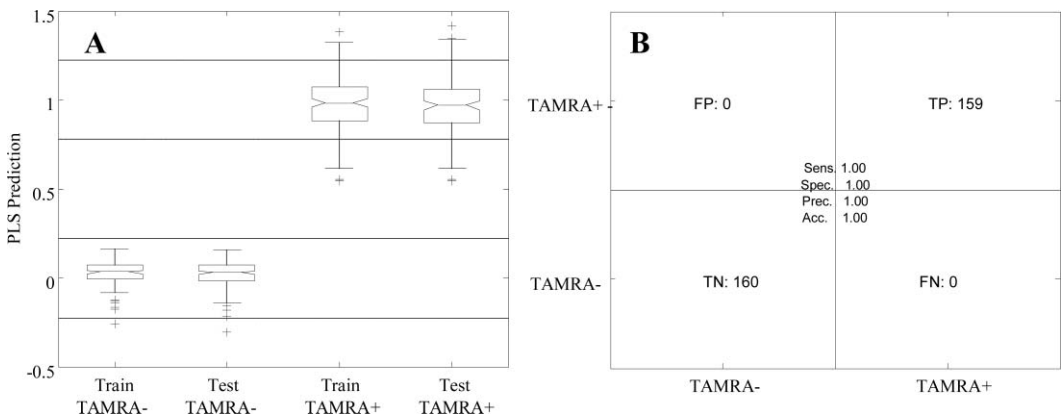


Fig. 4 PLS1 resampling results for DNA labelled with TAMRA (plus = present, minus = absent). (A) Boxplot summarizing the spread of all training and test predictions with the 95% confidence intervals represented by horizontal dotted lines. (B) A contingency matrix representing the model classification accuracy based upon the average predictions for the test samples and a classification boundary of $\geqslant 0.5$ (TAMRA present), <0.5 (TAMRA absent) (TP = true positive, FP = false positive, TN = true negative and FN = false negative).

can be derived from these calculations are also provided, these are:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{TN}}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}.$$

What is clear from these figures is that, even for the most challenging spectral deconvolution (with respect to ROX), PLS1 can be used to detect the presence of these multiplexing dyes in complex, low-concentration mixtures based upon their SERRS spectra, with an accuracy of almost 100%. In Fig. 4 and Fig. 5, one can see that for ROX there is a much smaller distance between the group centroids (0 and 1) representing the absence (0) or presence (1) of the dye when compared to TAMRA, although a majority of the predictions fall within

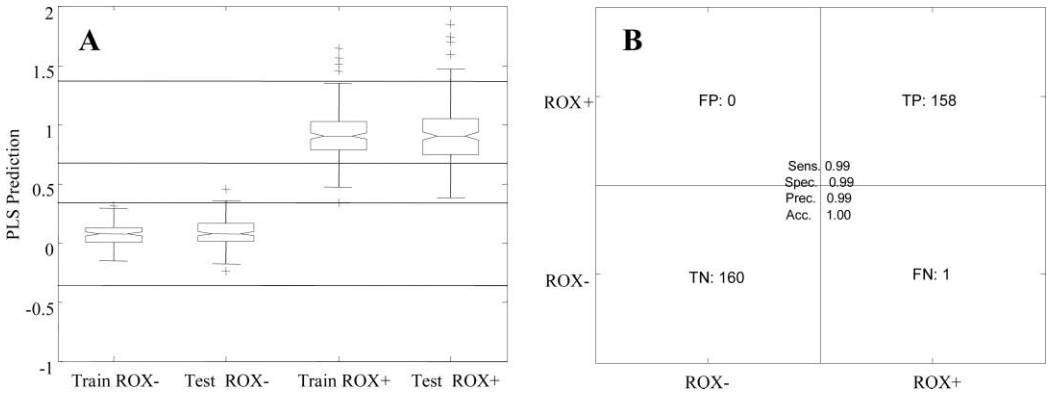


Fig. 5 PLS1 resample results for DNA labelled with ROX (plus = present, minus = absent). (A) Boxplot summarizing the spread of all training and test predictions with the 95% confidence intervals represented by horizontal dotted lines. (B) A contingency matrix representing the model classification accuracy based upon the average predictions for the test samples and a fixed classification boundary of $\geqslant 0.5$ (ROX present), <0.5 (ROX absent) (TP = true positive, FP = false positive, TN = true negative and FN = false negative).

95% confidence in both cases, indicating that there is simply a broader distribution of model predictions for ROX (as can be seen in the ESI†).

To summarize this information further, average classification accuracies across the 200 models (with one standard deviation of error about the mean) for each of the dye-labelled oligonucleotides based upon hard classification boundaries are collated in Table 2. In this table, different margins of error on the output are shown, which give an indication of how close to 0 or 1 the predictions have been, for the dyes being absent or present. For example, for the ROX-labelled oligonucleotides the average percentage of predictions showing the dye to be identified correctly is 97.3%. This means that when a positive identity of dye is taken as $\geqslant 0.5$, and absence as <0.5, over the 200 resampling iterations the samples are assessed correctly on average 97.3 times out of 100; this remains unchanged when a slightly tighter tolerance of defining presence as >0.6, and the absence of dye as <0.4; finally when a crisper cut off is used – presence is scored at >0.8, and absence is <0.2 – this falls slightly to 62%. Overall, from this table one can see that even

Table 2 A summary of classification accuracy based on hard thresholding for PLS predictions across all 200 resampled models

Dye [present (+), absent (–)]	Class boundary [ave.% ± std (train% ± std | test% ± std)]				
	0.5/0.5	0.4/0.6	0.2/0.8	0.1/0.9	0.05/0.95
Cy3– (%)	100 ± 0.1 (100 ± 0.1 | 100 ± 0)	99.8 ± 0.6 (100 ± 0.4 | 99.6 ± 0.7)	90 ± 3.9 (89.9 ± 4.3 | 90.1 ± 3.5)	66.3 ± 9.7 (65.4 ± 12.3 | 67.2 ± 7)	47.3 ± 7.6 (45.7 ± 7.5 | 48.8 ± 7.7)
Cy3+ (%)	99.2 ± 0.8 (99.4 ± 0.4 | 98.9 ± 1.1)	98 ± 1.2 (98.9 ± 0.6 | 97.1 ± 1.8)	84.8 ± 6.8 (86.1 ± 5.9 | 83.5 ± 7.7)	63.3 ± 6.8 (66.5 ± 8 | 60 ± 5.5)	50.8 ± 6.6 (52.1 ± 6.1 | 49.4 ± 7.1)
FAM– (%)	100 ± 0 (100 ± 0 | 100 ± 0)	100 ± 0 (100 ± 0 | 100 ± 0)	95.2 ± 3.1 (97.8 ± 1.4 | 92.6 ± 4.7)	66.7 ± 8.3 (74.9 ± 4 | 58.4 ± 12.5)	46.1 ± 7.9 (53.7 ± 5 | 38.5 ± 10.8)
FAM+ (%)	99.2 ± 1 (99.7 ± 0.4 | 98.6 ± 1.6)	96.5 ± 2.6 (98.7 ± 0.9 | 94.3 ± 4.2)	79.3 ± 4.2 (86.8 ± 4.1 | 71.8 ± 4.2)	59.7 ± 8.1 (68 ± 6.8 | 51.3 ± 9.3)	49.9 ± 7.6 (56.9 ± 6.2 | 42.8 ± 8.9)
HEX– (%)	100 ± 0 (100 ± 0 | 100 ± 0)	100 ± 0 (100 ± 0 | 100 ± 0)	95.6 ± 2 (97.2 ± 1.8 | 93.9 ± 2.2)	69.9 ± 7.5 (73.4 ± 9.3 | 66.3 ± 5.6)	47.6 ± 9.8 (49.1 ± 10.8 | 46.1 ± 8.8)
HEX+ (%)	97.1 ± 2.4 (98.2 ± 1.4 | 96 ± 3.3)	95.8 ± 2.3 (96.7 ± 1.9 | 94.8 ± 2.6)	82.6 ± 6.5 (87.5 ± 6.5 | 77.7 ± 4.5)	64.5 ± 8.2 (68.9 ± 1 1.3 | 60.1 ± 5)	50.8 ± 5.1 (54 ± 6.7 | 47.6 ± 3.5)
ROX– (%)	100 ± 0.1 (100 ± 0.1 | 100 ± 0)	99.4 ± 0.8 (99.8 ± 0.5 | 99 ± 1.1)	83.6 ± 7.3 (87.9 ± 5.5 | 79.3 ± 9)	53.8 ± 15.1 (56.8 ± 8.5 | 50.8 ± 21.7)	40.8 ± 14.8 (42.4 ± 7.5 | 39.1 ± 22)
ROX+ (%)	98.3 ± 0.7 (98.8 ± 0.4 | 97.7 ± 0.9)	93.8 ± 3.4 (96.1 ± 1.5 | 91.4 ± 5.3)	69 ± 7.2 (71 ± 6.6 | 66.9 ± 7.7)	50.5 ± 6.9 (51.1 ± 7.5 | 49.9 ± 6.2)	40.6 ± 8 (41.2 ± 6.7 | 39.9 ± 9.3)
TAMRA– (%)	100 ± 0 (100 ± 0 | 100 ± 0)	100 ± 0 (100 ± 0 | 100 ± 0)	100 ± 0.1 (100 ± 0.1 | 100 ± 0.1)	89.9 ± 5 (89.5 ± 4.5 | 90.2 ± 5.4)	63.4 ± 10.8 (61.5 ± 6.6 | 65.2 ± 15)
TAMRA+ (%)	100 ± 0 (100 ± 0 | 100 ± 0)	98.7 ± 1.4 (98.3 ± 1.1 | 99.1 ± 1.6)	90.3 ± 4.2 (89.8 ± 3.2 | 90.7 ± 5.2)	71.1 ± 2.6 (68.8 ± 2.1 | 73.4 ± 3.1)	58.8 ± 4.4 (57.4 ± 2.3 | 60.2 ± 6.4)
TET– (%)	100 ± 0 (100 ± 0 | 100 ± 0)	99.8 ± 0.4 (100 ± 0 | 99.6 ± 0.8)	91.6 ± 5.3 (96.1 ± 3.5 | 87.1 ± 7.1)	71.6 ± 5.6 (75.2 ± 6.2 | 67.9 ± 5)	55.2 ± 5.3 (57.4 ± 5.4 | 52.9 ± 5.1)
TET+ (%)	100 ± 0.1 (100 ± 0 | 100 ± 0.2)	98.1 ± 1.7 (99.3 ± 0.7 | 96.8 ± 2.6)	84 ± 3.2 (91.3 ± 3.2 | 76.7 ± 3.1)	63.3 ± 6.7 (75.3 ± 6.4 | 51.2 ± 6.9)	48.6 ± 4.7 (57.4 ± 5.8 | 39.8 ± 3.6)

under the most rigorous classification thresholds (<0.05 and >0.95) these models can still give correct predictions that are better than chance odds (*i.e.* >50%).

Conclusions

We have shown that the highly multiplexed detection of labelled oligonucleotides using SERRS is possible, but only when combined with chemometrics. It is clear from the results shown above that PLS1 can be used as a highly accurate method of calibrating these data to predict the presence or absence of a particular dye-labelled oligonucleotide within a complex mixture, and that this approach was validated by using random resampling so that statistical measures of classification accuracy and confidence limits can be generated.

In the future we shall investigate how far it is possible to multiplex DNA detection using this powerful combination of SERRS and machine learning algorithms. The obvious further advancement of the data analysis developed here is to move to more challenging and complex biological samples. We recognise that this a considerable advancement in the field; however, this preliminary study indicates that it should be possible.

Acknowledgements

K. F. and D. G. would like to thank the DTI Measurements for Biotechnology programme through their funding of the project. D. G. would also like to thank the RSC for the award of their Analytical Grand Prix fellowship. R. G. and R. J. are indebted to the Engineering and Biological Systems committee of the UK BBSRC for financial support.

References

1 H. Li, L. Ying, J. J. Green, S. Balasubramanian and D. Klenerman, *Anal. Chem.*, 2003, **75**, 1664.
2 M. Fleischmann, P. J. Hendra and A. J. McQuillan, *Chem. Phys. Lett.*, 1974, **26**(2), 163.
3 D. L Jeanmaire and R. P. Van Duyne, *J. Electroanal. Chem.*, 1977, **1**, 20.
4 A. M. Stacy and R. P. Van Duyne, *Chem. Phys. Lett.*, 1983, **102**, 365.
5 D. Graham, C. McLaughlin, G. McAnally, J. C. Jones, P. C. White and W. E. Smith, *Chem. Commun.*, 1998, 1187.
6 G. McAnally, C. McLaughlin, R. Brown, D. C. Robson, K. Faulds, D. R. Tackley, W. E. Smith and D. Graham, *Analyst*, 2002, **127**, 838.
7 D. Graham, W. E. Smith, A. M. T. Linacre, C. H. Munro, N. D. Watson and P. C. White, *Anal. Chem.*, 1997, **69**, 4703.
8 K. Faulds, W. E. Smith and D. Graham, *Anal. Chem.*, 2004, **6**(2), 412.
9 D. Graham, B. J. Mallinder and W. E. Smith, *Biopolymers*, 2000, **57**, 85.
10 D. Graham, R. Brown and W. E. Smith, *Chem. Commun.*, 2001, 1002.
11 L. Fruk, A. Grondin, W. E. Smith and D. Graham, *Chem. Commun.*, 2002, 2100.
12 R. Brown, W. E. Smith and D. Graham, *Tetrahedron Lett.*, 2003, **44**, 1339.
13 A. Ruperez, R. Montes and J. J. Laserna, *Vib. Spectrosc.*, 1991, **2**, 145.
14 K. Faulds, L. Stewart, W. E. Smith and D. Graham, *Talanta*, 2005, **67**, 667.
15 K. Faulds, R. P. Barbagallo, J. T Keer, W. E. Smith and D. Graham, *Analyst*, 2004, **129**, 567.
16 D. Graham, B. J. Mallinder, D. Whitcombe and W. E. Smith, *ChemPhysChem*, 2001, 746.
17 F. T. Docherty, P. B. Monaghan, R. Keir, D. Graham, W. E. Smith and J. M. Cooper, *Chem. Commun.*, 2004, 118.
18 K. Faulds, F. Mackenzie, W. E. Smith and D. Graham, *Angew. Chem., Int. Ed.*, 2007, **46**(11), 1829.
19 R. Brereton, *Chemometrics: data analysis for the laboratory and chemical plant*, John Wiley & Sons Ltd, Chichester, 2003.
20 W. Krzanowski, *Principles of Multivariate Analysis: A User's Perspective*, Oxford University Press, Oxford, 1988.
21 B. F. J. Manly, *Multivariate Statistical Methods: A Primer*, Chapman & Hall/CRC, New York, 1994.
22 C. Chatfield and A. J. Collins, *Introduction to Multivariate Analysis*, Chapman & Hall, London, 1980.
23 D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. DeJong, P. J. Lewi and J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier, Amsterdam, 1997.
24 R. O. Duda, P. E. Hart and D. E. Stork, *Pattern Classification*, Wiley, London, 2nd edn, 2001.
25 R. C. Beavis, S. M. Colby, R. Goodacre, P. B. Harrington, J. P. Reilly, S. Sokolow and C. W. Wilkerson, in *Encyclopedia of Analytical Chemistry*, ed. R. A. Meyers, Wiley, London, 2000, pp. 11558–11597.
26 R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan and D. B. Kell, *Trends Biotechnol.*, 2004, **22**, 245.
27 C. H. Munro, W. E. Smith, M. Garner, J. Clarkson and P. C. White, *Langmuir*, 1995, **11**, 3712.
28 P. C. Lee and D. Meisel, *J. Phys. Chem.*, 1982, **86**, 3391.
29 R. Goodacre, B. Shann, R. J. Gilbert, E. M Timmins, A. C. McGovern, B. K. Alsberg, D. B. Kell and N. A. Logan, *Anal. Chem.*, 2000, **72**, 119.
30 I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, 1986.
31 H. Martens and T. Næs, *Multivariate Calibration*, Wiley, Chichester, 1989.
32 R. K. H. Galvao, M. C. U. Araujo, M. D. N. Martins, G. E. Jose, M. J. C. Pontes, E. C. Silva and T. C. B. Saldanha, *Chemom. Intell. Lab. Syst.*, 2006, **81**, 60.
33 H. S. Basu and L. J. Marton, *Biochem. J.*, 1987, **244**, 243.

