Investigation of support vector machines and Raman spectroscopy for lymph node diagnostics

Martina Sattlecker,[a] Conrad Bessant,*[a] Jennifer Smith[b] and Nick Stone[b]

Received 29th September 2009, Accepted 19th February 2010
First published as an Advance Article on the web 5th March 2010
DOI: 10.1039/b920229c

This study concerns the combination of Raman spectroscopy and multivariate statistical analyses for the assessment of lymph nodes in the course of breast cancer diagnostics and staging. Axillary lymph node samples derived from breast cancer patients were measured by Raman microspectroscopy. The resulting Raman maps were pre-processed and cleaned of background noise and low intensity spectra using a novel method based on selecting spectra depending on the distribution of the mean of arbitrary units of all spectra within individual samples. The obtained dataset was used to build different types of Support Vector Machine (SVM) models, including linear, polynomial and radial basis function (RBF). All trained models were tested with an unseen independent dataset in order to allow an assessment of the predictive power of the algorithms. The best performance was achieved by the RBF SVM model, which classified 100% of the independent testing data correctly. In order to compare the SVM performance with traditional chemometric methods a linear discriminant analysis (LDA) model and a partial least square discriminant analysis (PLS-DA) model were generated. The results demonstrate the enhanced performance and clinical potential of the combination of SVMs and Raman spectroscopy and the benefits of the implemented filtering.

1 Introduction

Breast cancer is the most common cancer in women worldwide and, due to the increasing number of newly diagnosed cases, is a growing health care problem.

Most frequently breast cancer originates in the glandular elements of the breast, the lobules and the ducts. Malignant transformation includes such changes as nuclear enlargement, changes in the number of chromosome and variations in shape and size.[1] These changes affect the chemical composition and do not cause a large-scale production of new chemicals. One of the most significant changes in malignant disease development is the change of the nuclear-to-cytoplasm ratio. In this manner malignant tissue differs from benign tissue in the concentration of the main building blocks, nucleic acid, proteins, lipids and carbohydrates.[2]

Progressing breast carcinoma metastasizes to the regional lymph nodes over the efferent lymphatic vessels and enters the subcapsular sinus. For this reason early lymph node involvement is often found in the subcapsular sinus. An invaded lymph node may respond by displaying secondary follicles with reactive germinal centres, sinus histiocytosis and granulation. A further very specific change is desmoplasia, the change in the formation of collagenous fibrous stroma around the metastatic cells. With growing involvement genuine lymph node architecture gets increasingly replaced by metastases, in the majority of cases reflecting the features of the primary tumour.[3]

[a]*Cranfield University, College Road, Cranfield, Bedfordshire, UK MK43 0AL. E-mail: c.bessant@cranfield.ac.uk; Tel: +44 (0) 1234 758512*
[b]*Biophotonics Research Group, Gloucestershire Royal Hospital, Great Western Road, Gloucester, UK GL1 3NN. E-mail: n.stone@medical-research-centre.com; Fax: +44 (0) 8454 225485; Tel: +44 (0) 8485 225486*

Lymph node involvement is an important prognostic factor for breast cancer patients. Thus, breast cancer staging includes the assessment of the lymph nodes in the ipsilateral axilla. For this reason lymph node biopsy is carried out in order to determine the presence of metastasis.[4] A frequently applied method is sentinel lymph node biopsy, where the first node or nodes with direct lymphatic drainage from the tumour are identified. These lymph nodes are considered to be the first ones to be involved when a tumour metastasizes.[5] Lymph node involvement has a major impact on further treatment of the patient, including extensive dissection of axillary lymph nodes, chemotherapy and occasionally radiotherapy.

Current routine histopathology methods for lymph node assessment encounter several limitations. Traditional histological staining techniques are subjective, resulting in missed lesions and significant disagreement of inter- and intra-observers.[6] Alternatively, methods have been developed to allow faster intra-operative assessment of lymph nodes. These methods, including imprint cytology[7,8] and frozen section analysis,[9] also showed wide variations in the accuracy of detecting metastasis when compared with traditional methods.

Since current methods lack reliability, techniques are required which are more sensitive and objective. These requirements can be met by spectroscopic methods such as Raman spectroscopy. In this manner Raman spectroscopy has been extensively investigated for it capability to differentiate between malignant and benign tissue samples, including breast,[10–12] gastric,[13,14] colon,[15,16] bladder,[17,18] cerivical[19–21] and parathyroid tissue.[22]

A further step towards the clinical application of Raman spectroscopy includes the development of diagnostic algorithms, which allow reliable classification of tissue samples. For this purpose several multivariate statistical techniques, such as Linear

Discriminant Analysis (LDA),[13,14,17–19,21] Principal Component Analysis (PCA),[15,20] Artificial Neural Networks (ANN)[23] and Support Vector Machines (SVM),[16] have been investigated. However, LDA is the predominantly applied technique for classifying Raman data derived from tissue samples. In general, linear classification approaches achieve a diagnostic accuracy of around 90%, varying between different tissue types.[13,14,17–19,21] Nevertheless, there is space for improvement, which is crucial for future diagnostic application of Raman spectroscopy.

Support vector machines, developed by Vapnik[24] and Burges,[25] are considered to be superior over traditional linear approaches due to their capability to represent non-linear features within data. For this reason, SVMs increasingly attracted attention for classifying spectral data derived from tissue for diagnostic purposes.[16,26,27] At present, SVMs have not been investigated for metastasis detection in lymph nodes.

In this work we report the successful application of SVM for detecting metastasis in axillary lymph nodes based on Raman spectroscopic data. The developed algorithm comprises a novel spectral filter method, which removes noise and low intensity spectra. For the evaluation of the SVMs over traditional linear methods, LDA and PLS-DA models were generated. The performances of these models were compared.

2 Materials and methods

2.1 Samples

A total of 43 axillary lymph node samples were collected after surgical resection for breast cancer. All samples were obtained with the full consent of patients and approved by the Gloucestershire Research Ethics Committee. Each lymph node was cut into halves. One half was placed onto acetate paper and snap frozen in liquid nitrogen in order to maintain the freshness of the tissue. From the frozen sample a 7 μm section was cut and placed on a calcium fluoride slide and stored in a −80 °C freezer for Raman spectroscopy. The other half of the node was sent for the routine histopathology, which found that out of the 43 samples 13 were positive and 30 were negative for metastasis.

2.2 Raman microspectroscopy

A Renishaw System 1000® Raman microspectrometer coupled to a diode laser, a Leica® microscope, a Prior® electronic stage, a video viewer and a desktop computer with customized Grams® software was used for all measurements. The diode laser had an output of 350 mW and was set to a wavelength of 830 nm with the aim to reduce autofluorescence from tissue. Raman mapping was executed in steps of 100 μm in x and y directions across the sample surface. At each point the spectra were integrated for a total of 30 seconds.

2.3 Data preprocessing and spectra selection

All preprocessing and downstream data analyses were executed in Matlab (Mathworks, USA). For this reason the raw Raman maps were loaded into Matlab and converted into 3D hyperspectral matrices. For each individual map principal component analysis (PCA) was executed. The resulting first three principal components were used to transform the Raman map into a composite image. According to the composite images heterogeneous regions were selected manually by avoiding obvious fat or blood likely areas. Thus, spectra were collected, which represent homogenous regions of positive and negative nodal parenchyma. An example of this process is demonstrated in Fig. 1. Saturated spectra and spectra containing evidence of cosmic rays were removed from the dataset. The resulting dataset consisted of 10 477 spectra, where 3385 were from positive samples and 7092 were from negative samples. The number of spectra available for each sample varied from four to 1014 spectra per sample. The high variation of available spectra for each sample was caused by the fact that some nodes contained a large amount of fat, what resulted in an increased number of saturated spectra, which had to be removed before further analysis.

A filtering method was developed to select spectra in order to improve the performance of all classification models. This method filters spectra for every lymph node independently by calculating the mean of arbitrary intensity units for each single spectrum within the sample. By assessing the distribution of the mean values it was observed that data distribution is left skewed as illustrated in Fig. 2. For this reason it was decided to use the mean value which occurs most frequently (mode) within the sample because it is assumed this value represents spectra, which are diagnostically relevant. Spectra with a mean value smaller than the mode are dismissed with the aim to remove low intensity spectra from the dataset.

A second threshold was estimated by adding the standard deviation to the first limit. Spectra with a higher mean value than this threshold were dismissed due to the fact that it was observed that they often exhibited non-spectral features and artefacts without diagnostic value. Further these features are assumed to decrease the performance of a classification model. Fig. 2 illustrates the effect of the spectra selection method on the basis of two individual lymph node samples.

2.4 Support vector machines

Support vector machines (SVM) are a relatively young classification method and were first introduced by Vapnik.[24] The fundamental idea of this technique is to separate classes with a hyperplane by maximising the margin between them. The simplest approach is to plot the dataset into the N-dimensional input space. For spectral data N is equivalent with the amount of wavenumbers recorded in each spectrum or with the number of selected features. In the input space samples are considered as points that are separable by a hyperplane into distinctive groups (positive and negative in this application). More complex datasets are frequently not separable in the input space. In order to address this problem data are mapped into a higher dimensional space (feature space) by a kernel function. The most frequently applied types of kernel functions are:

Linear: $K(x_i, x_j) = x_i \times x_j + 1$

Polynomial kernel: $K(x_i, x_j) = (x_i \times x_j + 1)^d$

Gaussian radial basis function: $K(x_i, x_j) = \exp\frac{-\|x_i - x_j\|^2}{2\sigma^2}$

For the application of SVMs the kernel function must be selected *a priori*. Depending on the type of kernel, specific

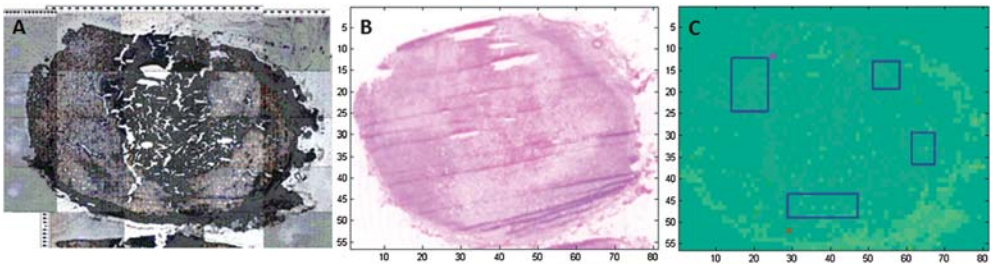


Fig. 1 (A) White light image of a lymph node effaced with metastatic tumour. (B) The H&E staining shows pale stained areas, which are a result of the marked desmoplastic reaction to the metastatic tumour. (C) Composite image of the lymph node sample. Each pixel represents one spectrum. Thus, the grid unities represent spectra for both axes (81 spectra in x direction and 56 in y direction). Areas boxed in blue represent the selected spectra, which were then used for further investigations.

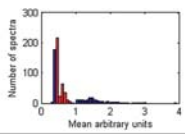
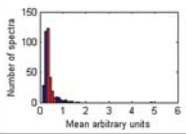
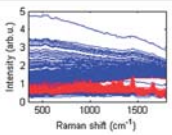
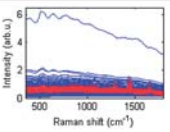
	Sample 1	Sample 2
Pathology group	Positive	Negative
Number of spectra from sample before filtering	697	376
Distribution of mean values		
Selected spectra		
Number of spectra from sample after filtering	140	75

Fig. 2 This illustration demonstrates the applied spectra selection method. Sample 1 represents one individual positive lymph node sample and sample 2 represents one individual negative lymph node sample out of 43. The histogram illustrating the mean intensity values of spectra for two individual samples, one positive lymph node and one negative lymph node. Spectra with a mean intensity value between the two estimated thresholds are illustrated in red. As a result spectra representing the majority of the sample are selected and non-spectral features as well as low intensity spectra are dismissed.

parameters must be set. In this manner, for a polynomial kernel the polynomial order d and for a Gaussian RBF kernel the Gaussian width σ must be specified.

Once the data are mapped into the feature space an infinite number of separating hyperplanes may exist, creating the risk of overfitting the hyperplane to the given data points. An overfit hyperplane might separate the training data perfectly, but perform poorly on unseen data. A soft margin was introduced in order to avoid overfitting the hyperplane by allowing some training data points to be misclassified. This misclassification rate is regulated by the penalty weight C. The higher C the lower the misclassification rate, thus for $C = \infty$ no misclassification is allowed.

2.5 Model construction and testing

Every individual sample set was filtered according to the method described earlier (Section 2.3). From the resulting data for each sample 50 spectra were selected randomly in order to balance the dataset. The resulting dataset was split randomly into a training set (31 samples) and a testing set (12 samples). Samples with less than 50 spectra available were also integrated into the test set. Thus the training set consisted of 1550 and the testing set of 355 spectra.

All SVM models were built in Matlab utilizing LIBSVM toolbox 2.88.1 by Chang and Lin.[28] For this study three different types of kernel functions were investigated: linear, polynomial and RBF. For every kernel the necessary parameters were optimised by a combined approach of leave one sample out cross validation (LOOCV) and a grid search. C is the only parameter which is required to be optimised for a linear SVM. The range of C was set from 2^{-5} to 2^{13}, increasing in steps of the power of two. Each individual value of C within the predefined range was used for LOOCV. For this step only a subset of the training data was used with the aim to speed up the parameter search. Thus, five positive and five negative samples were selected randomly out of the training data, which were considered sufficient for a first parameter approximation. The parameter C, which resulted in the highest LOOCV result was estimated and used for a second more rigid parameter search. This time LOOCV was executed with C values close around the previously estimated value. Additionally, the whole training set was used for this fine tuning step.

For the polynomial and the RBF kernel the parameter optimisation was similarly executed as for the linear kernel. The main difference is that for these kernel types two parameters have to be optimised, which can be done by a grid search. In a grid search systematically alternating pairs of parameters are used for LOOCV. The parameter combination that resulted in the best LOOCV result is estimated and in the following fine tuned. For the optimisation of a polynomial kernel changing combinations of C and the polynomial degree are applied for LOOCV. The range of C was set from 2^{-5} increasing in steps of the power of two up to 2^{15}. C and σ were the two parameters that have to be optimised for a RBF kernel SVM. For C the range was set from 2^{-5} to 2^{15}, increasing in steps of the power of two. The range for σ started at 2^{-15} up to 2^{3}, also increasing in steps of the power of two. For the first parameter approximation again the reduced training set was applied. The estimated parameter combinations were then used for a second grid search. In comparison, the complete training set was applied for the fine tuning procedure.

As a result of the grid search for each kernel function the best parameters were estimated. These parameters were then used to

build the final SVM model. In that manner, three SVM models, a linear, a polynomial and a RBF model were generated and finally tested with the unseen dataset.

In order to obtain further assessment of the performance of the SVM models the same dataset was used for the investigation of two other types of classification techniques, linear discriminant analysis (LDA) and a partial least squares discriminant analysis (PLS-DA). The LDA model was trained by LOOCV with an increasing number of PCs (1–25) fed into the model. The number of PCs leading to the best LOOCV result was used to build the final model, which was then tested with the independent testing set. For the PLS-DA model, the number of latent variables (LVs) was optimised by LOOCV. The number of LVs achieving the highest LOOCV result was used to build the final model, which was also tested with the unseen dataset. The PLS-DA model was built by utilizing PLS Toolbox 3.5 (Eigenvector Research).

2.6 Assessment of model fit

This approach applies Monte Carlo methods for the empirical assessment of the model fit. Specifically, all samples of the training set, which was used to build the original model are assigned randomly with a class membership, either positive or negative. These samples are then used to build a model using the same parameters as the originally estimated. The newly generated model is then tested with the testing set, which has also been randomly assigned with class membership. This procedure is executed multiple times, where each time the class membership of the data is randomly permuted. The achieved testing accuracies are then used to create a null distribution. The comparison of the null distribution with the observed testing performance allows empirical assessment of the model fit. This approach is inspired by similar workflows, such as that used by Wongravee *et al.*[29] to estimate the significance of variables for variable selection purposes.

2.7 Investigation of key features

For the model development the spectral region from 350–1850 cm^{-1} was used. In order to investigate spectral features that have the greatest impact on the model performance alternating intervals of 100 wavenumbers were eliminated from the dataset in a systematically manner. The remaining dataset, with a total of 1401 wavenumbers, was used to build and test a SVM model by using the previously optimised parameters. In this manner for a spectral range starting at 350 to 1850 cm^{-1} 15 models were built and tested. A decrease in testing performance is assumed to be caused by the fact that the left out spectral features have a high impact on the model performance.

3 Results and discussion

3.1 Model training

For the first step of the SVM optimisation only a subset of the training data was used in order to maintain a manageable training time. For the linear kernel the first parameter approximation showed that a C of 2^7 results in the highest LOOCV of 75.4%. A second parameter optimisation was carried for further optimising the previously estimated C. For this search the range of C was set from 2^6 to 2^8 with an increase of $2^{0.25}$ and the whole training set was used. The best parameter C was identified to be $2^{6.75}$ by achieving a LOOCV result of 88.7%. The training of the polynomial kernel included the optimisation of two parameters, C and the polynomial degree. The executed grid search showed that lower polynomial degree and an increased C result in an improved LOOCV performance. Thus, less complex polynomial models seem to be more appropriate for modelling these data. The best parameter was found to be a polynomial degree of 3 and C 2^{14} resulting in 87.4% LOOCV accuracy. A RBF kernel likewise a polynomial kernel requires the optimisation of two parameters, C and σ. The first grid search resulted in an approximated C value of 2^9 and σ of 2^{-13}, with a LOOCV accuracy of 95.6%. The fine tuned parameters were found to be $C = 2^{10}$ and $\sigma = 2^{-13.25}$ with a LOOCV accuracy of 89.4%. The LDA model was trained by optimising the number of PCs fed into the LDA. The highest LOOCV accuracy of 89.5% was achieved by feeding 13 PCs into the LDA model. For the PLS-DA model the optimised number of LVs was found to be 8 by resulting in a LOOCV accuracy of 82.7%.

3.2 Model testing

The optimised parameters and the whole training set were used to build the final classification models, which were then tested with the unseen testing set. The RBF SVM classified the entire testing set correctly and thus achieved a 100% classification accuracy. Almost a similar performance was achieved with the polynomial SVM by achieving a classification accuracy of 99.2%. In comparison the linear classification models showed a reduced classification performance. Thus, the linear SVM identified 92.4% of the testing set correctly, the LDA model 90.3% and the PLS-DA model 95.2%. All results including sensitivity and specificity are summarised in Table 1. These results show that for Raman spectral data from human lymph nodes non-linear techniques perform better on predicting classes of unseen data than linear ones, although, all applied methods performed similarly during the training by LOOCV procedure.

Table 1 Summary of all built classification model. All results are percentages of correctly classified spectra

Method	Parameters	Training (%)	Testing (%)	Sensitivity testing (%)	Specificity testing (%)
Linear SVM	$C = 6.75$	97.4	92.4	100	90.1
Polynomial SVM	$d = 3$, $C = 13$	95.8	99.2	100	98.9
RBF SVM	$\sigma = C = 10$	93.7	100	100	100
LDA	$PC = 13$	90.3	93.8	100	91.9
PLS-DA	$LV = 8$	94.5	95.2	100	93.8

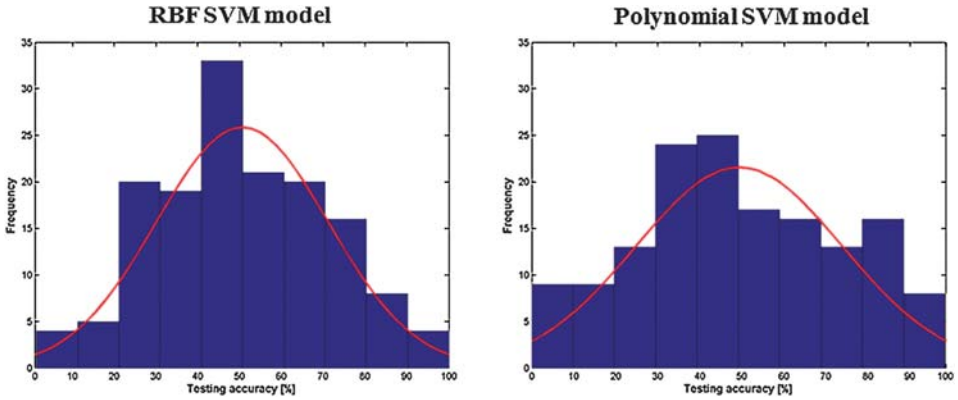


Fig. 3 Null distributions for RBF and polynomial SVM models were created by testing 150 random models. The null distribution generated by the polynomial SVM is significantly flatter than the null distribution generated by the RBF random model. Thus, the RBF SVM model is better suited to this application than the polynomial SVM model.

3.3 Assessment of model fit

The RBF SVM model, as well as the polynomial SVM model, was used to investigate the model fit based on Monte Carlo methods. These two models were used because they achieved the highest testing performance as illustrated in Table 1. For each of the SVM models 150 random models were generated. The resulting null distributions are illustrated in Fig. 3. The comparison of the two null distributions shows that the fitted curve for the polynomial model is flatter than for the RBF model. In this manner, the polynomial SVM model achieved random testing accuracies above 80% more frequently than the RBF SVM model. The comparison of the observed testing result (polynomial SVM 98.9% and RBF SVM 100%) with the null distributions shows that the polynomial model it is more likely to achieve a classification in the range of the observed result. Therefore, it can be said that the polynomial model might be slightly over fitted and due to that the RBF SVM model is more suited to this dataset.

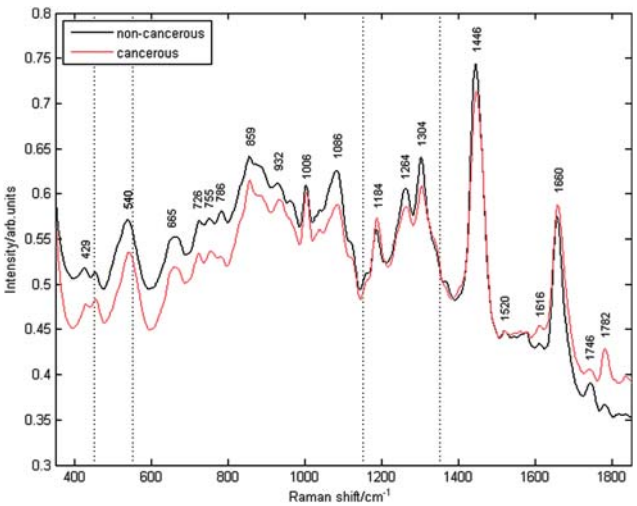


Fig. 4 The graph illustrates the mean Raman spectra for cancerous and non-cancerous samples including the assignment of major peaks.

3.4 Investigation of key features

Furthermore, the RBF SVM model was used for investigating the wavenumbers that have the highest impact on the model performance. For this 100 wavenumbers were eliminated from the data systematically. It showed that out of 15 generated models only two showed a decrease in testing performance. These regions were identified as 451–550 cm^{-1} (decreased testing accuracy: 99.2%) and 1151–1250 cm^{-1} (decreased testing accuracy: 98.9%). Nevertheless, the loss in performance is not significant and thus the impact of these spectral features cannot be considered as major. This becomes even more obvious by the fact that when these two regions were eliminated from the dataset the generated model still achieved a testing performance of 98.9%. On the other hand, creating a model by only using these two regions resulted in a relatively low testing performance of 76.9%. For this reason an assessment of whether the model performance could be improved by adding a third spectral region was performed. The two previously identified spectral regions (450–549 cm^{-1} and 1150–1249 cm^{-1}) were combined with one of the remaining ones. Thus, a total of 13 models were generated. As showed that an extract of 300 wavenumbers is sufficient to achieve a 100% testing result and a training result of 90.3%, what is close to the 93.7% testing result of the model build on the whole spectral range. This training and testing results were achieved by the combination of the spectral intervals starting from 450–550 cm^{-1} and 1150–1350 cm^{-1}. Peaks identified in these regions are 540, 1184, 1264, 1304 cm^{-1} (Fig. 4). A summary of all models including peaks and their major assignment is illustrated in Table 2. This demonstrates that 300 wavenumbers are sufficient to generate an accurate SVM model and further brings the advantage of reducing computational time.

4 Conclusions

This study demonstrated that SVM coupled with Raman spectroscopy is a superior approach over traditional methods for the classification of Raman spectral data derived from tissue. Especially, the RBF SVM shows high diagnostic

Table 2 Spectral intervals given in italics were shown to be the only ones to have a minor impact on the model performance. The combination of these two intervals with a third one was used to build a SVM classification model based on only 300 wavenumbers. These two intervals were rotationally combined with each of the remaining spectral regions and then used to generate a RBF SVM model. As the training and testing results show models built by only using 300 features are a good approximation to the original RBF SVM model applying 1500 features

Interval/cm^{-1}	Train (%)	Test (%)	Peak position/cm^{-1}	Major assignments
350–450	81.6	93.0	429	Calcium hydroxyapatite
450–550	—	—	*540*	*Disulfide bonds*
550–650	82.5	85.9		
650–750	85.3	85.9	665	Thiamine
			726	C–S(protein)/CH_2 rocking/adenine
750–850	90.3	97.8	755	Symmetric breathing of tryptophan
			786	DNA: O–P–O, cytosine, uracil, thymine
850–950	88.8	98.6	859	Tyrosine/collagen
			932	Skeletal C–C: α-helix
950–1050	87.6	96.9	1006	Phenylalanine, carotenoids
1050–1150	86.8	100	1086	Skeletal C–C stretch
1150–1250	—	—	*1184*	*Cytosine, guanine, adenine*
1250–1350	90.3	100	1264	Amide III (α-helix), =C–H in plane bending (lipid)
			1304	CH_2 deformation (lipid)/adenine, cytosine
1350–1450	90.1	100	1446	CH_2 bending modes of proteins
1450–1550	86.4	99.2	1520	–C=C–carotenoid
1550–1650	90.1	100.0	1616	C=C stretching mode of tyrosine and tryptophan
1650–1750	89.4	99.7	1660	Amide I (protein)
			1746	C=O stretch (lipid)
1750–1850	77.6	99.2	1782	Unknown assignment

potential for future application due to the fact that this achieved a 100% classification accuracy on previously unseen data. Further testing also proved the statistical significance of the model. Further it was found that not any spectral region has a major impact on the model performance. Nevertheless, an extract of 300 wavenumbers is sufficient to achieve the same testing performance as a model generated by using the whole Raman spectra consisting of 1500 wavenumbers. A lower number of spectral features results in a reduced computing time, which is desirable especially for future clinical application of Raman spectroscopy.

Acknowledgements

This work was funded by Cranfield University and Gloucestershire Hospitals NHS Foundation Trust. Nick Stone is funded by a National Institute of Health Research Senior Research Fellowship. Additional thanks for their contribution to the technical and administrative staff of the Department of Histopathology, Gloucestershire Royal Hospital, and to medical and administrative staff in the Department of Breast Surgery, Gloucestershire Royal Hospital.

References

1 V. Kumar, A. K. Abbas, N. Fausto, S. L. Robbins and R. S. R. p. b. o. d. Cotran, in *Robbins and Cotran Pathologic Basis of Disease*, ed. V. Kumar, A. K. Abbas and N. Fausto, J. A. Perkins (illustrations), Elsevier Saunders, Philadelphia, PA, London, 7th edn, 2005.
2 K. E. Shafer-Peltier, A. S. Haka, M. Fitzmaurice, J. Crowe, J. Myles, R. R. Dasari and M. S. Feld, *J. Raman Spectrosc.*, 2002, **33**, 552–563.
3 H. L. Ioachim, L. J. Medeiros and H. L. I. s. l. n. p. Ioachim, in *Ioachim's Lymph Node Pathology*, ed. H. L. Ioachim and L. J. Medeiros, Lippincott Williams & Wilkins, Philadelphia, PA, London, 4th edn, 2009.
4 S. Arnaud, G. Houvenaeghel, V. Moutardier, M. Butarelli, M. Martino, A. Tallet, A. C. Braud, J. Jacquemier, C. Julian-Reynier and I. Brenot-Rossi, *Eur. J. Surg. Oncol.*, 2004, **30**, 735–743.
5 D. L. Morton, D. R. Wen, J. H. Wong, J. S. Economou, L. A. Cagle, F. K. Storm, L. J. Foshag and A. J. Cochran, *Arch. Surg.*, 1992, **127**, 392–399.
6 G. Cserni, S. Bianchi, W. Boecker, T. Decker, M. Lacerda, F. Rank and C. A. Wells, *Cancer*, 2005, **103**, 358–367.
7 A. A. Salem, A. G. Douglas-Jones, H. M. Sweetland and R. E. Mansel, *Eur. J. Surg. Oncol.*, 2003, **29**, 25–28.
8 A. A. Salem, A. G. Douglas-Jones, H. M. Sweetland and R. E. Mansel, *Eur. J. Surg. Oncol.*, 2006, **32**, 484–487.
9 D. A. Grabau, F. Rank and E. Friis, *Apmis*, 2005, **113**, 7–12.
10 M. V. Chowdary, K. K. Kumar, J. Kurien, S. Mathew and C. M. Krishna, *Biopolymers*, 2006, **83**, 556–569.
11 A. S. Haka, K. E. Shafer-Peltier, M. Fitzmaurice, J. Crowe, R. R. Dasari and M. S. Feld, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 12371–12376.
12 N. Stone and P. Matousek, *Cancer Res.*, 2008, **68**, 4424–4430.
13 T. Kawabata, T. Mizuno, S. Okazaki, M. Hiramatsu, T. Setoguchi, H. Kikuchi, M. Yamamoto, Y. Hiramatsu, K. Kondo, M. Baba, M. Ohta, K. Kamiya, T. Tanaka, S. Suzuki and H. Konno, *J. Gastroenterol.*, 2008, **43**, 283–290.
14 S. K. Teh, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh and Z. Huang, *Br. J. Cancer*, 2008, **98**, 457–465.
15 M. V. Chowdary, K. K. Kumar, K. Thakur, A. Anand, J. Kurien, C. M. Krishna and S. Mathew, *Photomed. Laser Surg.*, 2007, **25**, 269–274.
16 E. Widjaja, W. Zheng and Z. Huang, *Int. J. Oncol.*, 2008, **32**, 653–662.
17 B. W. de Jong, T. C. Schut, K. Maquelin, T. van der Kwast, C. H. Bangma, D. J. Kok and G. J. Puppels, *Anal. Chem.*, 2006, **78**, 7761–7769.
18 P. Crow, A. Molckovsky, N. Stone, J. Uff, B. Wilson and L. M. WongKeeSong, *Urology*, 2005, **65**, 1126–1130.
19 P. R. Jess, D. D. Smith, M. Mazilu, K. Dholakia, A. C. Riches and C. S. Herrington, *Int. J. Cancer*, 2007, **121**, 2723–2728.
20 C. M. Krishnaa, N. B. Prathimaa, R. Malinia, B. M. Vadhirajab, R. A. Bhattc, D. J. Fernandesb, P. Kushtagic, M. S. Vidyasagarb and V. B. Karthaa, *Vib. Spectrosc.*, 2006, **41**, 136–141.
21 F. M. Lyng, E. O. Faolain, J. Conroy, A. D. Meade, P. Knief, B. Duffy, M. B. Hunter, J. M. Byrne, P. Kelehan and H. J. Byrne, *Exp. Mol. Pathol.*, 2007, **82**, 121–129.

22 K. Das, N. Stone, C. Kendall, C. Fowler and J. Christie-Brown, *Laser Med. Sci.*, 2006, **21**, 192–197.
23 S. Sigurdsson, P. A. Philipsen, L. K. Hansen, J. Larsen, M. Gniadecka and H. C. Wulf, *IEEE Trans. Biomed. Eng.*, 2004, **51**, 1784–1793.
24 V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, London, 1995.
25 C. J. C. Burges, *Data Min. Knowl. Discov.*, 1998, **2**, 121–167.
26 W. Lin, X. Yuan, P. Yuen, W. I. Wei, J. Sham, P. Shi and J. Qu, *J. Biomed. Opt.*, 2004, **9**, 180–186.
27 S. K. Majumder, N. Ghosh and P. K. Gupta, *J. Biomed. Opt.*, 2005, **10**, 024034.
28 C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines*, 2001.
29 K. Wongravee, G. R. Lloyd, J. Hall, M. E. Holmboe, M. L. Schaefer, R. R. Reed, J. Trevejo and R. G. Brereton, *Metabolomics*, 2009, **5**, 387–406.