RSC Advances

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances

Journal Name

ARTICLE

Identification of different tumor states in nasopharyngeal cancer using surface-enhanced Raman spectroscopy combined with Lasso-PLS-DA algorithm

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Guannan Chen,[a*] Xueliang Lin,[a] Duo Lin,[a,b**] Xiaosong Ge,[a] Shangyuan Feng,[a] Jianji Pan,[c] Juqiang Lin,[a] Zufang Huang,[a] Xi Huang,[a] and Rong Chen[a]

Identification of different states in cancer is of vital importance for cancer treatment and management. A powerful diagnostic algorithm based on Lasso-partial least squares-discriminant analysis (Lasso-PLS-DA) was developed here for improving blood surface-enhanced Raman spectroscopy (SERS) analysis, with the aim to classify different states in nasopharyngeal cancer (NPC). A total of 160 blood plasma samples were collected for this study, obtained from 60 normal volunteers, 25 T1 stage cancer and 75 T2–T4 stages cancer patients. Results show that a diagnostic sensitivity of 68% and a specificity of 84.0% can be achieved for separating T2-T4 stage from T1 stage cancer, which had a 20% improvement in diagnostic specificity compared with the previous work. This exploratory study demonstrates that the Lasso-PLS-DA can be integrated with blood SERS analysis as a promising clinical complement for different T stages detection in NPC.

1 Introduction

Early detection and accurate identification of different stages for cancer is crucial to improving patients' survival by making proper treatment strategy. In the past decades, several optical spectroscopic technologies have been comprehensively explored for non-invasive and objective cancer detection, mainly including infrared, fluorescent and Raman spectroscopy (RS) [1, 2]. In particular, RS is capable of probing 'fingerprints' information of specific biomolecular showing promising application for cancer diagnosis[3-6]. Specifically, the advent of surface-enhanced Raman spectroscopy (SERS) has further extended the biomedical application of conventional RS[7-9], as the Raman signals can be dramatically enhanced to enable single molecule detection by exploiting the interaction between the analytes of interest and metal nanoparticles (NPs) surface [10-11]. Compared with infrared and fluorescent spectroscopy technologies, SERS holds significant advantages in minimal photobleaching, minimal background signal from aqueous samples, and multiplexing capabilities under a single excitation light. With the ability to explore extremely subtle changes of biomolecular content and structure associated with cancer transformation, SERS has recently attracted increasing attentions as a potential tool for cancer screening. For instance, Wang et al. reported a specific and sensitive methodology using epidermal growth factor-SERS nanoparticles that can rapidly detect circulating tumor cells in peripheral blood specimens from patients with squamous cell carcinoma of the head and neck [12]. Additionally, a novel method based on SERS for human saliva analysis has been investigated for non-invasive nasopharyngeal and lung cancer detection [13, 14].

It should be noted that each raw Raman spectrum obtained from biological sample usually contains high dimension of the spectral space such as intensity variables, which will result in computational complexity and inefficiency in extracting the most diagnostically significant information. Besides, the Raman spectra belong to similar subjects are commonly similar, making it a challenge to differentiate them sensitively with simplistic band feature analysis. These main limitations will hinder further clinical applications of Raman spectroscopy in medical diagnosis. Numerous developments in multivariate analysis including principal component analysis (PCA), linear discriminant analysis (LDA), partial least-squares regression (PLS), artificial neural networks (ANNs), support vector machines (SVM) and genetic algorithm (GA), within the past decade have enabled significant progress of RS and other technologies in biomedical detection [1, 15-19]. For example, Huang et al. demonstrated the ability to identify dysplasia from normal gastric muscosa tissue using RS in conjunction with PCA-LDA [4]. Similar diagnostic algorithm was also used for cell and blood identification based on Raman spectra for cancer detection [20, 21]. Most previous researches focused on discriminate cancer from normal subjects using Raman method with multivariate analysis, however there is few study on identification of different cancer stages, which is of great importance for cancer treatment and management. Very recently, we have evaluated the feasibility of a label-free method based on blood plasma SERS with PCA-LDA for exploring variability of different tumor (T) stages in nasopharyngeal cancer (NPC) [22]. This

[a] Key Laboratory of Optoelectronic Science and Technology for Medicine, Ministry of Education and Fujian Provincial Key Laboratory for Photonics Technology, Fujian Normal University, Fuzhou 350007, China
[b] College of Integrated Traditional Chinese and Western Medicine, Fujian University of Traditional Chinese Medicine, Fuzhou, Fujian, 350122, China
[c] Fujian Provincial Cancer Hospital, Fuzhou, Fujian, 350001, China
Email: *edado@fjnu.edu.cn; **linduo1986@163.com

preliminary study showed high diagnostic accuracies of 83.5% and 93.3%, respectively, can be achieved for classification of T1 stage cancer and normal, and T2–T4 stage cancer and normal blood groups. However, the diagnostic accuracy is only 63.0% for classification of T1 stage cancer and T2–T4 stage cancer. Thus, the development of a more powerful diagnostic algorithm that could identify Raman spectra belong to different NPC stages would be of significant clinical value during blood SERS analysis.

In this work, a robust multivariate statistical method based on Lasso-partial least squares-discriminant analysis (Lasso-PLS-DA) was employed to develop efficient diagnostic algorithm for classification of SERS spectra between blood samples from different NPC stages.

2 Material and Methods

2.1 Preparation of Au Colloids and Human Blood Plasma Samples

The stable Au colloid solutions used for SERS were prepared following the protocol reported by Grabar et. al. [23] The obtained NPs of Au colloid follow a normal distribution with a mean diameter of 43 nm and standard deviation of 5 nm. A total of 160 blood plasma samples were collected in this study, from 60 normal volunteers, 25 T1 stage cancer and 75 T2–T4 stages cancer patients. After 12 hours of overnight fasting, a single 3 mL blood samples were collected from the study subjects between 7:00–8:00 A.M. with the use of EDTA anticoagulant. This study was performed in compliance with the relevant laws and institutional guidelines, and was approved by ethical committee of Fujian provincial cancer hospital. In addition, the informed consent was obtained from all subjects. Finally, a drop of plasma-Au NPs mixture (20 μL blood plasma and 20 μL Au colloid) was transferred onto a rectangle aluminum plate for SERS measurement. More detail information has been described in our previous paper [22].

2.2 SERS Measurement

In brief, a Renishaw Raman micro-spectrometer (Great Britain) was employed for blood plasma SERS measurement using a 785 nm laser excitation source. The system acquires SERS spectra in the wavelength region of 400–1750 cm^{-1}, and each spectrum was acquired within 10 s integration time and with ×20 objective (NA = 0.4). The spectral resolution of the system is 2 cm^{-1}. The software package WIRE 2.0 was used for spectral acquisition and analysis.

2.3 Data Processing

Least absolute shrinkage and selection operator

Lasso is fundamentally based on the familiar expression for multiple linear regression, which is of the form:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n$$

Where Y is a combination of the parameters, β_0 is the regression coefficient for the residual and the β_i values are the regression coefficients (for predictor variables 1 through n) computed from the data. The data X is highly correlated in a multi-dimensional space (wavelength bins). Different group of statistical methods can be used to shrinkage regression in different ways.

Lasso regression is a regularization technique by reducing the number of predictors in a regression model. It uses the original data matrix X to constrain the values of the correlation coefficients values of the multiple linear regression. It produces shrinkage estimates with potentially lower predictive errors than ordinary least squares (the model parameter of lasso should be adaptively chosen to minimize an estimate of expected prediction error.). Under this constraint, the model weighs the importance of each channel for prediction, and unimportant channels are driven to β values equal to 0 by the optimization process. The formulae of Lasso is expressed as [24]:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \le t$$

Where, $\hat{\beta}^{lasso}$ represent the estimated coefficient, arg $\min_{\beta}$ is the vector β with minimal mean squared error, N is the number of observations, x_{ij} is data, a vector of p values at observation ij, y_i is the response at observation i, the parameters β_0 and β_j are scalar and p-vector respectively.

In general, the advantage of Lasso is to drive the parameters to zero deselects the features from the regression. Thus, Lasso automatically selects more relevant features and discards the others in an iterative process. This advantage also has the effect of making the Lasso robust restrain noise. A sparser model with smaller number of non-zero coefficient (called β values) could be produced by Lasso model most significantly.

After obtaining the useful spectral variables using the Lasso, we noticed that not all of the useful variables were distributed in each SERS bands, and some bands had more useful variables than others. In order to avoid the uncertainty of prediction result due to the interference part variables by the noise, we choose some integral SERS spectral ranges which contain one or more useful spectral variables to establish further prediction model. PLS-DA is employed to classify the cancer stages detection based on the spectral bands selected above in this study. Two block regression is made by the PLS. Firstly, the dependent block (X) can predict the independent block (Y). The Y block represents the class labels and each X block represents each spectrum. PLS-DA integrates the basic principle of PCA, and maximizes the covariance between group affinity and spectral variation in order to rotate the components further. Therefore, the diagnostically relevant variation could be explained by the PLS components. However, the number of model components causes the complexity of the PLS-DA model. The performance of the PLS-DA is measured by comparing the root mean square error in

prediction (RMSEP) of the model proposed by PLS-DA with the RMSEP of the model containing all the variables.

RMSEP is defined as

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}}$$

Where, N is the number of objects in the evaluation set. Due to overfitting, an external validation set is used to avoid overoptimistic results. Moreover, it often does not allow us to perform statistical tests on the significance of the difference in RMSEPs for the limited size. The prediction results based on Lasso-PLS-DA with RMSEP analysis are always better than that using complex full-spectrum model based on PLS-DA with RMSEP.

3 Results and Discussion

Using Au-NPs as substrate, we have successfully acquired blood plasma SERS spectra from 160 subjects. In these samples, 60 were histopathological normal and 100 were NPC. According to TNM classification, 25 cancers were of T1 stage and 75 T2–T4 stages. The fluorescence background of the original SERS data was removed using a modified multi-polynomial fitting algorithm [25], then each spectrum was normalized by the integrated area under the curve, and after that the whole normalized SERS data sets were fed into Matlab for analysis. All the analysis method was coded and run in Matlab (MATLAB R2013b, MathWorks, Natick, MA, US).

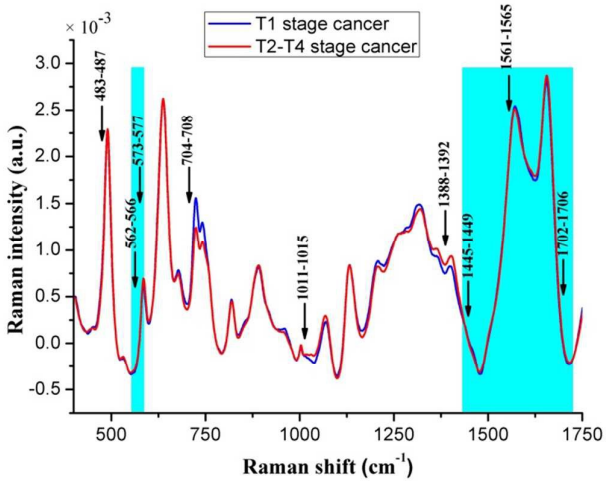


Fig. 1 The selection variables of SERS bands using Lasso algorithm. The black arrows show nine significant bands selected by Lasso algorithm. Two integral SERS spectral ranges (550-585 and 1435-1730 cm^{-1}) which contain more useful spectral variables are marked by the cyan shadow area.

The Lasso algorithm with Leave-one-out cross-validation (LOOCV) was employed to seek the significant SERS spectral features that were immediately bound up with different stage cancer pathologies firstly. The latter convention was used as the Lasso's model parameter in this paper because this parameter can contact to the useful features more intuitively and directly than others. In LOOCV, one cancer sample (i.e., one spectrum) was taken out from all of these 100 cancer samples, and then the rest of blood spectra were used to reconstruct by the Lasso algorithm for classifying the selected spectrum. This procedure was iterated until all withheld cancer sample were classified [4]. The features for SERS spectrum obtained by Lasso algorithm were shown in Fig. 1. Both LOOCV and 10-fold cross-validation got the same nine significant band regions of spectral variables. Nine significant band regions of spectral variables (483-487, 562-566, 573-577, 704-708, 1011-1015, 1388-1392, 1445-1449, 1561-1565 and 1702-1706 cm^{-1}) were selected from the SERS band regions. Two integral SERS spectral ranges (550-585 and 1435-1730 cm^{-1}) which contain more useful spectral variables were also marked by the cyan shadow area. According to previous literatures [22], the selected spectral ranges were possibly related to DNA/RNA bases and Amide I. It can be seen that the spectral features (band positions, intensities and bandwidths) of the two regions between T1 and T2-T4 cancer plasma are very similar, whereas some significantly diagnostic variables can be extracted by Lasso algorithm from them. The reason may be that cancer belongs to part of a widely accepted multistep, continuum progression cascade from normal to cancer, and it suggests subtle and vague molecular distinction, making it a challenge to identify different cancer stages by simplistic spectral features analysis. This result confirms a potential role of the proposed method based on Lasso algorithm for classification of different cancer stages. Similarly, Huang et al. applied genetic algorithm to select significant spectral variables from the Raman band regions for providing clinically discrimination between normal and precancer cervical tissues [15]. Different from their work, the selected spectral range in this work is wider aim to avoid interference to variables from the noise.

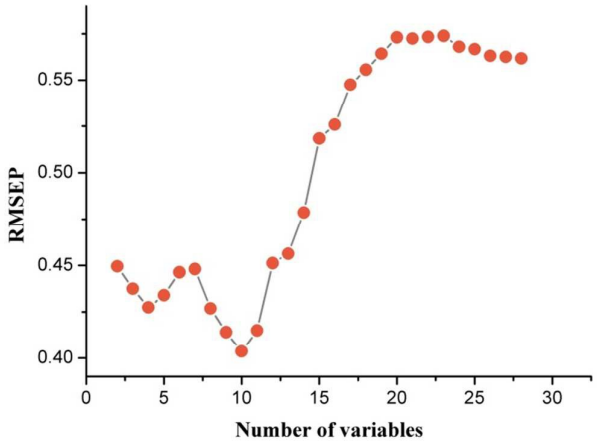


Fig. 2 Root mean standard error of prediction (RMSEP) as a function of the number of variables.

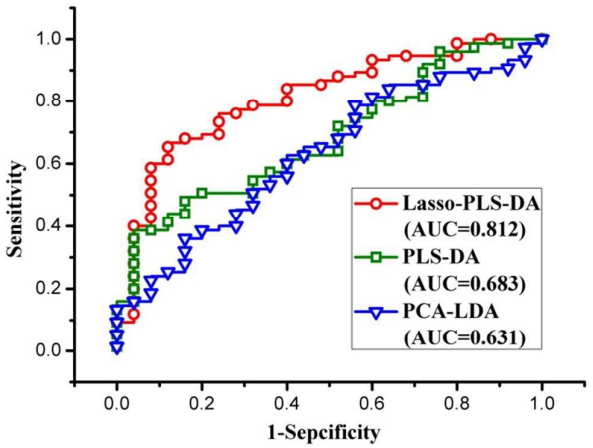


Fig. 3 Receiver operating characteristic (ROC) curves of discrimination results for different T stages of NPC generated from Lasso-PLS-DA, PLS-DA and PCA-LDA with leave-one-out cross-validation algorithms. The integrated areas under the ROC curves (AUC) for Lasso-PLS-DA, PLS-DA and PCA-LDA are 0.812, 0.683 and 0.631.

The PLS-DA model was then further used for the selected SERS band regions. The optimum number of variables was determined with leave-one-out cross-validation using root mean standard error method. Fig. 2 represents that the number of variables was generated by RMSEP. The minimum value for the optimum number of variables was showed in RMSEP, and due to overfitting it raises with the increasing number of variables. To assess the predictive accuracy of the Lasso-PLS-DA based diagnostic algorithms, the receiver operating characteristic (ROC) curve (Fig. 3) was produced. The ROC curves for Lasso-PLS-DA was generated by calculate the selected two spectral ranges, with the integration area under the ROC curve (AUC) of 0.812 (the optimum number of components was 10). The ROC curve for PLS-DA and PCA-LDA by calculate the full-spectrums was 0.683 (the optimum components = 4), and 0.631 (the optimum components = 4), respectively. It was found that the Lasso-PLS-DA algorithm was capable of achieving greater efficiency in comparison to conventional algorithm based on PLS-DA and PCA-LDA. This is explainable. Using the full-spectrum variables, the diagnostic efficiency of PLS-DA and PCA-LDA may be interfered by non-significant variables and noise. For Lasso-PLS-DA, spectral regions including the selected most significant spectral variables, were employed as an optimal input for further PLS-DA, allowing a reliable way to solve these limitations. Posterior probability values were also used to predict the response (Fig. 4). The posterior probability scatter plot yielded a diagnostic sensitivity of 68% (51/75) and a specificity of 84.0% (21/25) for separating T2-T4 stage from T1 stage cancer with a threshold line, which had a remarkable improvement compared with the previous work (a sensitivity of 62.7% (47/75) and a specificity of 64.0% (16/25)) [22]. Excitingly, the diagnostic specificity was increased by 20% in this work. From the result above, we can find that the Lasso-PLS-DA algorithm renders a powerful way to identify different stages cancer by developing a classification model from the significant Raman features.

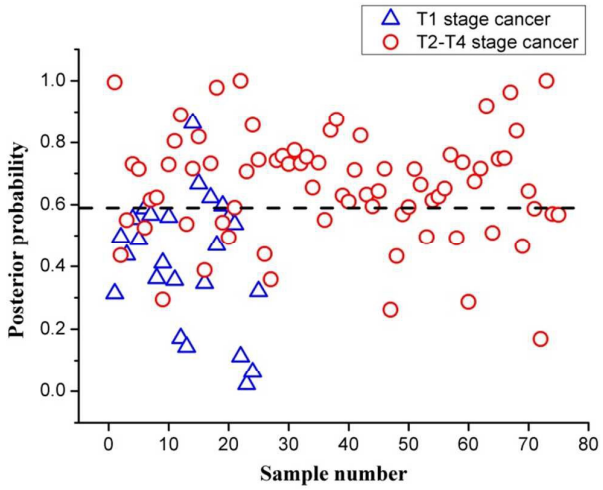


Fig. 4 Scatter plot of the posterior probability values for different stages cancer using the Lasso-PLS-DA with the leave-one-out, cross-validation. The separate line provides a diagnostic sensitivity of 68% (51/75) and a specificity of 84.0% (21/25) for discriminating T2-T4 stage from T1 stage cancer.

We also calculated the ROC curves of discrimination results for normal and cancer (Table 1). The results using PLS-DA and PCA-LDA with leave-one-out cross-validation for full-Spectrum were 0.931 and 0.919, whereas the result was 0.930 using Lasso-PLS-DA. Results indicate that the Lasso-PLS-DA used to efficiently distinguish different NPC stages is also can be used to distinguish normal and cancer groups.

Journal Name

ARTICLE

Table 1 Classification results of SERS prediction using Lasso-PLS-DA, PLS-DA and PCA-LDA together with the leave-one-out, cross-validation.

Diagnostic combinations	The integration area under the ROC curve		
	Lasso-PLS-DA	PLS-DA	PCA-LDA
T1 stage vs. T2-T4 stage Cancer	0.812	0.683	0.631
Normal vs. Cancer	0.930	0.931	0.919

4 Conclusions

In summary, a powerful diagnostic method based on SERS combined with Lasso-PLS-DA algorithm was developed for differentiating blood plasma obtained from NPC patients in different states. Lasso algorithm is capable of extracting some significantly diagnostic variables for classify the SERS spectra which have similar spectral features. Further, these valuable variables can be used for determining some SERS spectral ranges in order to avoid interferences to the variables from the noise. Thus, Lasso algorithm is suitable and reliable for evaluation of SERS spectra. Besides, results shows the diagnostic efficiency can be significantly increased by Lasso-PLS-DA in comparison to conventional algorithm based on PLS-DA and PCA-LDA, demonstrating that plasma SERS analysis with Lasso-PLS-DA algorithms has great potential to be a clinical complement for NPC detection, especially for the classification of different tumor stages. Our next step is to conduct more detailed prospective studies and obtain more data set to verify the novelty and reliability of this potential diagnostic algorithm.

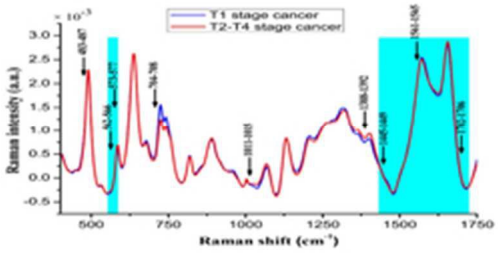
Acknowledgements

This work was supported by the Program for Changjiang Scholars and Innovative Research Team in University (IRT15R10), and the National Natural Science Foundation of China (Grant Nos. 61575043, 61178090, 61210016, 81101110 and 61405036).

References

1. C. Krafft, G. Steiner, C. Beleites and R. Salzer, *J. Biophotonics*, 2009, **2**, 13-28.
2. J. Hegyi, V. Hegyi, T. Ruzicka, P. Arenberger and C. Berking, *JDDG: Journal der Deutschen Dermatologischen Gesellschaft*, 2011, **9**, 368-372.
3. Q. Tu and C. Chang, *Nanomedicine*, 2012, **8**, 545-558.
4. S. Teh, W. Zheng, K. Ho, M. Teh, K. Yeoh and Z. Huang, *Br. J. Cancer*, 2008, **98**, 457-465.
5. H. J. Lee, W. Zhang, D. Zhang, Y. Yang, B. Liu, E. L. Barker, K. K. Buhman, L. V. Slipchenko, M. Dai and J.-X. Cheng, *Sci Rep*, 2015, **5**, 7930.
6. E. Brauchle, S. Thude, S. Y. Brucker and K. Schenke-Layland, *Sci rep*, 2014, **4**, 4698.
7. D. Zhu, Z. Wang, S. Zong, H. Chen, P. Chen and Y. Cui, *RSC Adv*, 2014, **4**, 60936-60942.
8. J. Wang, R. Liu, C. Zhang, G. Han, J. Zhao, B. Liu, C. Jiang and Z. Zhang, *RSC Adv*, 2015, **5**, 86803-86810.
9. D. Pissuwan, A. Hobro, N. Pavillon and N. Smith, *RSC Adv*, 2014, **4**, 5536-5541.
10. L. Ou, Y. Chen, Y. Su, Y. Huang, R. Chen and J. Lei, *J. Raman Spectrosc.*, 2013, **44**, 680-685.
11. J. Lin, Z. Huang, S. Feng, J. Lin, N. Liu, J. Wang, L. Li, Y. Zeng, B. Li and H. Zeng, *J. Raman Spectrosc.*, 2014, **45**, 884-889.
12. X. Wang, X. Qian, J. J. Beitler, Z. G. Chen, F. R. Khuri, M. M. Lewis, H. J. C. Shin, S. Nie and D. M. Shin, *Cancer Res.*, 2011, **71**, 1526-1532.
13. S. Feng, D. Lin, J. Lin, Z. Huang, G. Chen, Y. Li, S. Huang, J. Zhao, R. Chen and H. Zeng, *Appl. Phys. Lett.*, 2014, **104**, 073702.
14. X. Li, T. Yang and J. Lin, *J. Biomed. Opt.*, 2012, **17**, 0370031-0370035.
15. S. Duraipandian, W. Zheng, J. Ng, J. J. Low, A. Ilancheran and Z. Huang, *Analyst*, 2011, **136**, 4328-4336.
16. V. L. Tsang, A. X. Wang, H. Yusuf-Makagiansar and T. Ryll, *Biotechnol. Prog.*, 2014, **30**, 152-160.
17. H. Kuang, Y. Xia, J. Liang, B. Yang, Q. Wang and Y. Sun, *Carbohydr. Polym.*, 2011, **84**, 1258-1266.
18. J. A. Etzel, N. Valchev and C. Keysers, *Neuroimage*, 2011, **54**, 1159-1167.
19. M. Z. Martin, M. A. Mayes, K. R. Heal, D. J. Brice and S. D. Wullschleger, *"Spectrochim. Acta, Part B "*, 2013, **87**, 100-107.
20. J. W. Chan, D. S. Taylor, S. M. Lane, T. Zwerdling, J. Tuscano and T. Huser, *Anal. Chem.*, 2008, **80**, 2180-2187.
21. S. Feng, R. Chen, J. Lin, J. Pan, G. Chen, Y. Li, M. Cheng, Z. Huang, J. Chen and H. Zeng, *Biosens. Bioelectron.*, 2010, **25**, 2414-2419.
22. D. Lin, J. Pan, H. Huang, G. Chen, S. Qiu, H. Shi, W. Chen, Y. Yu, S. Feng and R. Chen, *Sci rep*, 2014, **4**, 4751.
23. K. C. Grabar, R. G. Freeman, M. B. Hommer and M. J. Natan, *Anal. Chem.*, 1995, **67**, 735-743.
24. T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman and R. Tibshirani, *The elements of statistical learning*, Springer,

2009.
25. Z. Huang, A. McWilliams, H. Lui, D. I. McLean, S. Lam and H. Zeng, *Int. J. Cancer*, 2003, **107**, 1047-1052.



21x10mm (300 x 300 DPI)