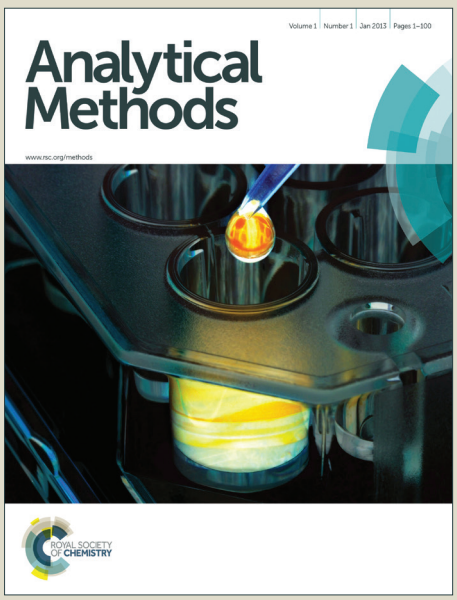
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Journal Name

RSCPublishing

ARTICLE

Raman peak recognition method based automated fluorescence subtraction algorithm for retrieval of Raman spectra of highly fluorescent samples

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Kun Chen,[a] Haoyun Wei,[a*] Hongyuan Zhang,[a] Tao Wu,[a] and Yan Li[a]

Intense fluorescence background is a major problem in the application of Raman spectroscopy. An appropriate algorithm which can faithfully retrieve weak tissue Raman signals is required. In this article, we propose a new algorithm for automated and artifact-free recovery of Raman spectra which combines a novel Raman peak recognition method (RPR method) with an improved iterative smoothing method (SG-SR method). SG-SR method, based on the modified Savitzky-Golay iterative process, substantially improve its convergence speed. By applying a novel negative relaxation factor to the Successive Relaxation iterative method, an automatic recognition of Raman peak is realized. In the proposed algorithm (RIA-SG-RPR algorithm), a real Raman peak position is firstly detected by RPR method to serve as the intrinsic criterion of convergence for the SG-SR method to avoid human interference. Then, real Raman signals are recovered from the iterative procedure of SG-SR method. This algorithm has been optimized and validated with mathematically simulated Raman spectrum as well as experimentally measured Raman spectra from varied fluorescent samples, resulting in a significant improvement on the rejection of both high fluorescence background and direct human intervention. This algorithm drastically avoids false Raman features to benefit the utilization of Raman spectroscopy to characterize molecular specifics in a more challenging Raman applications.

Introduction

In the past decades, Raman spectroscopy has been experiencing a period of growing interest in characterizing chemical agents, materials and biomedinice,[1-4] as an invaluable analytical tool. However, during the application of Raman spectroscopy , fluorescence , sometimes several orders of magnitude more intense than the weak Raman scattering, severely interferes with the Raman signals. Both instrumental[5–13] and computational methods[14–23] have been developed to subtract the intense background of Raman features for the application of in-vivo Raman spectroscopy. On one hand, the instrumental method based technique includes shifted excitation[5-11] and time gating.[12,13] Recently, thanks to the compact tunable laser source[5,9-10] and the robust extracted algorithm[7,8], shifted excitation Raman difference spectroscopy (SERDS) has become an applicable tool for the fluorescence subtraction. On the other hand, the computational method based techniques extract the Raman signal by carrying mathematical post-processing, including frequency-domain filtering such as Fourier transform (FFT),[15] iterative moving averaging technique,[16] wavelet transforms (DWT),[17] Shifted-Spectra Technique and first- and second-order derivatives[21] and polynomial fitting.[14,19] Each of these methods has its own advantage when used in certain situations. Due to its high efficiency and simplicity[18,19] the polynomial fitting is the most popular and widely used method, especially in biomedical applications since it was developed. After that, a modified multi-polynomial fitting (ModPoly)-based iterative algorithm was proposed by Lieber and Mahadevan-Jansen,[14] which largely reduced the dependence of the spectra on the polynomial order in Raman spectra processing. This algorithm was further improved by Zhao *et al*.[20] However, the polynomial fitting method is sensitive to the choice of the spectral region. Recently, Krishna *et al*[24] proposed a Range-independent background subtraction algorithm (RIA) based on the Savitzky-Golay smoothing method. However, Savitzky-Golay iterative smoothing suffers from its low speed of convergence, which costs huge amount of the computation time in Raman processing, especially in the case of large amount of spectral

data. In our previous work, we presented a novel modification of a Savitzky-Golay based fluorescence subtraction algorithm (RIA-SG-SR algorithm)[25] which drastically improves processing speed. RIA-SG-SR algorithm uses the SG-SR method based iteration instead of Savitzky-Golay iteration to achieve faster convergence. Real-time chemical analysis benefit from the convergence improvement since RIA-SG-SR can provide an efficient and rapid recovery of Raman Spectra within a few of milliseconds.

However, both RIA and RIA-SG-SR algorithm are still subject to two major limitations, especially in low signal-to-Fluorescence ratios (SFR) Raman spectra for real-time and artifact-free Raman processing systems. The first is that two extra artificial Raman peaks must be created as the criterion of convergence for these two algorithms, which means the retrieval of Raman signatures still suffers from their dependence on direct human intervention to manually specify the height and width of the two artificial Raman peaks. Nevertheless, subjective human intervention must be minimized for practical automated fluorescence subtraction. The second limitation is that the subtractions of the two algorithms are found to lead to distortions and false peaks in the recovered spectrum in the low Signal-to-Fluorescence ratio situation. Considering the typically low SFR of the raw Raman spectrum collected from biological tissue and other highly fluorescent samples, false peaks will significantly limit the application of RIA and RIA-SG-SR in clinic.

In this article, we present a novel fluorescence subtraction algorithm, named RIA-SG-RPR algorithm, based on an initial Raman peak recognition method to avoid or minimize certain shortcomings of RIA and RIA-SG-SR algorithms. The fundamental idea of this peak recognition method is that high-frequency components (Raman peaks) grow much faster than the baseline in a specific iterative process and the peaks can be determined subsequently. Compared with our previous work in Ref. 25, the novelties and findings of this manuscript are addressed to the improvement on automated and artifact-free subtraction while yielding consistent rejection of the intense fluorescence in low SFR Raman spectra. RIA-SG-RPR hardly needs direct human intervention to manually specify the convergence criterion. More importantly, RIA-SG-RPR substantially avoids false peaks and distortions added on recovered Raman signals in the case of highly fluorescent samples. This algorithm has been optimized and validated with mathematically simulated Raman spectrum as well as experimentally measured Raman spectra from varied fluorescent sample.

Materials and Methods

Fluorescence Subtraction Problem

Measured spectral data obtained on N detector channels can be represented as

$$O^0(v)=R(v)+B(v)+n(v) \quad (1)$$

where v is the Raman shift in cm^{-1}. The background fluorescence $B(v)$ and random noise $n(v)$ are added on the real Raman signal $R(v)$. $O^0(v)$ is the measured spectra. The purpose of fluorescence subtraction is to produce an estimation of $R(v)$ from $O^0(v)$. Actually, some filtering methods should be firstly applied on the raw spectrum to reduce interference by noise prior to fluorescence subtraction algorithm[26]. And developments in hardware system can also largely suppress noise and provide better signal-to-noise ratios. After that, Eq. 1 can be simplified as

$$O(v)=R(v)+B(v) \quad (2)$$

The concerns with fluorescence subtraction problem are addressed to isolate the Raman features from intense fluorescence. Raman signals always include some sharp Raman lines (~ 10 to 30 cm^{-1} or less in spectral width) which are high-frequency components compared with the broad underlying continuum of fluorescence background. A feasible method is to remove those high-frequency features through robust low-pass filters. From this point of view, Savitzky-Golay based smoothing method has been successfully applied to RIA[24] and RIA-SG-SR[25] algorithm. In details, the RIA and RIA-SG-SR algorithms are iterative smoothing of the measured raw Raman spectrum ($O(v)$) in such a manner that the high-frequency Raman peaks ($R(v)$) are gradually eliminated, finally leaving the underlying broad baseline ($B(v)$) which can be subtracted from the raw spectrum to yield the true Raman signal.

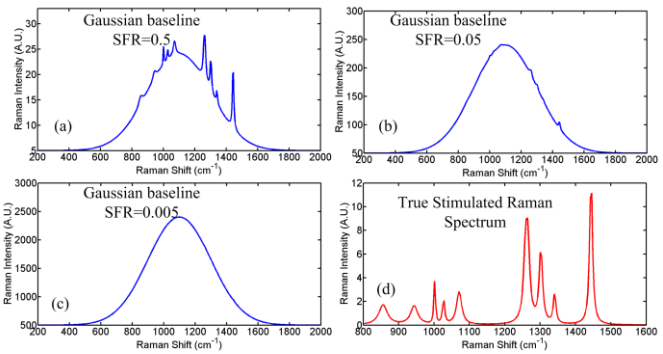


Fig.1 (a), (b), and (c) are the example simulated Raman spectra superimposed on Gaussian baseline with three different Signal-to-Fluorescence Ratios (SFR) 0.5, 0.05 and 0.005, respectively. (d) is the mathematically simulated Raman spectra.

In general, the line-profile of $R(v)$ is considered to be Lorentzian in nature[27]. Hence, Lorentzian expression is also utilized in this article to model the simulated Raman signals, as seen in Fig. 1(d). Four types of baselines, including fifth-order polynomial, exponential, Gaussian and sigmoidal distributions are applied to simulate the complicated background ($B(v)$), which are depicted in Ref. 25 and also discussed in other reports[16,18]. Signal-to-Fluorescence Ratios (SFR) are varied in our paper to evaluate the reconstruction performances of different algorithms, mathematically and experimentally. It is defined as $SFR=R_{max}-R_{min}/F_{max}-F_{min}$, where $R_{max(min)}$ represents the maximum (minimum) intensity of Raman peaks and $F_{max(min)}$ represents the maximum (minimum) intensity of fluorescence. As an example, Fig. 1 (a), (b) and (c) show the Raman signal added to Gaussian baseline with different SFRs: SFR=0.5, SFR=0.05 and SFR=0.005. It is pertinent to mention here that the rest of three types have the similar trends qualitatively, in other words, the Raman peaks gradually disappeared with the decrease of the SFR.

Raman Peak Recognition (RPR) method

The Savitzky-Golay smoothing method is a type of filter, first described in 1964 by Abraham Savitzky and Marcel J. E. Golay[28], which yields the smoothing value by use of the polynomial least square,

$$\hat{X}_i = GX_i \quad (3)$$

where, $X_i=[x^i_{-k},x^i_{-k+1},\cdots x^i_0,\cdots x^i_{k-1},x^i_k]^{\mathrm{T}}$ is the raw spectrum data. $\hat{X}_i=[\hat{x}^i_{-k},\hat{x}^i_{-k+1},\cdots \hat{x}^i_0,\cdots \hat{x}^i_{k-1},\hat{x}^i_k]^{\mathrm{T}}$ is the result of smoothing. G is determined by the window length ($2k$+1). In RIA-SG-SR[25] algorithm, we proposed an improvement on the Savitzky-Golay smoothing method to accelerate the iterative procedure and obtain great reduction of the computation time. This smoothing method is denoted as SG-SR method and its iterative form are as follows,

$$\hat{X}_i=(1-w)X_i+w(L\hat{X}_i+DX_i+UX_i) \quad (4>w>0) \quad (4)$$

$$\hat{X}_i=(E-wL)^{-1}((1-w)+D+U)X_i \quad (4>w>0) \quad (5)$$

where, E is identity matrix. $G=D+U+L$; D, U and L are strictly lower triangular matrix, diagonal matrix and strictly upper triangular matrix, respectively. From the mathematical demonstration procedure in Ref. 24, the range of $4>w>0$ is the condition for the convergence of SG-SR method. In that case, high-frequency Raman peaks ($R(v)$) are gradually eliminated and $O(v)$ gradually converges to $B(v)$.

It is important to point out that, the convergence or divergence of Eq. 5 is determined only by the relaxation factor w. Here, we firstly attempt to apply SG-SR method to an appropriate negative relaxation factor.

$$\hat{X}_i=(E-wL)^{-1}((1-w)+D+U)X_i \quad (w<0) \quad (6)$$

Unlike the convergence in Eq. 5, Eq. 6 is divergent and $\hat{X}_i$ would head for infinity with the increase of iteration. Contrary to SG-SR method, high-frequency components (Raman peaks) grow much faster than the baseline in the iterative process. And on this basis, an innovative Raman Peak Recognition (RPR) method is proposed, which could be used to suppress the broad background and extract high-frequency Raman peaks, with an appropriate negative relaxation factor w after an appropriate iteration number.

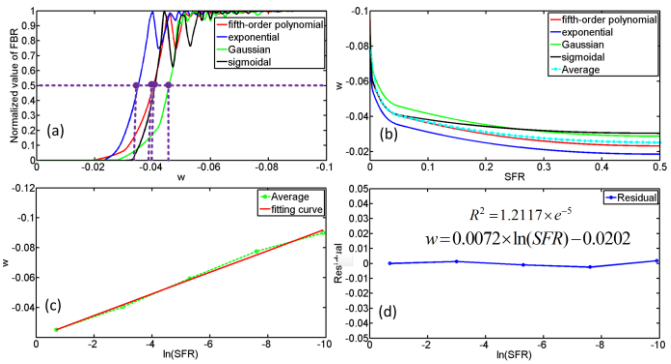


Fig.2 The determination of the optimal relaxation factor based on different baselines (a) and different SFRs (b), respectively. (c): One-order polynomial fitting result with logarithmic coordinates. (d): The residual error is within ±0.002.

Typically, the RPR method with an optimal value of the negative relaxation factor results in a more reliable recognition of Raman peaks. So, the next discussion should be addressed to the determination of the optimal w. At first, peak-to-background ratio (PBR), which is defined as $PBR=Peak-Background/Peak+Background$, of processed curve with different baselines after 100 iterations are calculated, shown in Fig. 2(a). A larger PBR means that one can recognize the tallest Raman peak more easily. It can be seen from Fig. 2(a) that, PBR gradually approaches to 1 with the increase of the absolute value of w quickly. Here, the w, corresponding to the PBR of 0.5 is chosen as the optimal relaxation factor for RPR method. In this case, the maximum peak is three times the intensity of background to ensure the accurate recognition of the most prominent Raman peak. It is noticed that the optimal valves for four different baselines are slightly different (-0.040, -0.035, -0.043 and -0.041, respectively). Fig. 2 (b) illustrates the choice of w with different SFRs and different baselines. The absolute value of w decreases as the increase of SFR. When the SFR is large enough, w is chosen as zero and $\hat{X}_i$ would remain unchanged. It means that Raman peaks can be determined directly, instead of the iterative procedure. Even though w is slightly different between different baselines, all the four types have the similar trends qualitatively. Given that the measured background is the combination of various baseline types, the average value of different baselines of w can be used in the practical processing. Furthermore, using polynomial fit technique, a more accurate equation of the average w and SFR can be obtained with logarithmic coordinates in Fig. 2(c). Obviously, it is a mode of linear relation as:

$$w = 0.0072\times\ln(SFR) - 0.0202 \quad (7)$$

$$SFR = \exp(\frac{w+0.0202}{0.0072}) \quad (8)$$

And the residual error is within ±0.002 in Fig. 2(d). In fact, since the SFR of a raw Raman spectrum can be approximatively estimated by visual inspection firstly, one can determine an approximate w with small tolerance from Eq. 7. It is necessary to point out that iteration number brings as great influence as w to the peak recognition result. PBR increases with the increase of iteration number. However, excessive iterations should be avoided from the perspective of reducing computational cost. Given that, 100 iterations is chosen for RPR method in this paper. Then, one can also determine the optimal w to obtain adequate PBR to find real Raman peaks as mentioned above.

Fig. 3 illustrates the recognition results of Raman peaks. Fig. 3(a) gives a specific example of this method. The Raman peak parts of high-frequency on the curve are gradually extracted after 100 iterations when the relaxation factor is set as -0.040. What is important is that the maximum points coincide with the positions of real Raman peaks in the simulation, which means the iterative result with negative relaxation factor can be used to determine the positions of Raman peaks. To be specific, the envelop curve is firstly achieved by fitting and interpolation of spline curve; and then peak positions can be accurately detected from this curve, shown in Fig. 3(a) with green filled circles. Then, the position of the maximum Raman peak, P, is picked as the criterion of convergence to avoid human interference for the subsequent iterative procedure. The performance of RPR method with respect to different SFRs and different baselines are also demonstrated in Fig. 3(b)-(f).

Other Raman peak detection methods such as second-order derivatives and ridge lines mentioned in Ref. [17], have been practically proven to attain reliable positions of Raman peaks, good peak-width estimation and also the true Raman signals. These methods can be considered as integrated processing

system. In contrast, RPR method cannot subtract baseline from the raw spectrum by itself, but provide a simpler and quicker process. Therefore, it finds its application on detecting the positions of Raman peaks, one of which can be served as the intrinsic criterion of convergence for the subsequent subtraction algorithm.

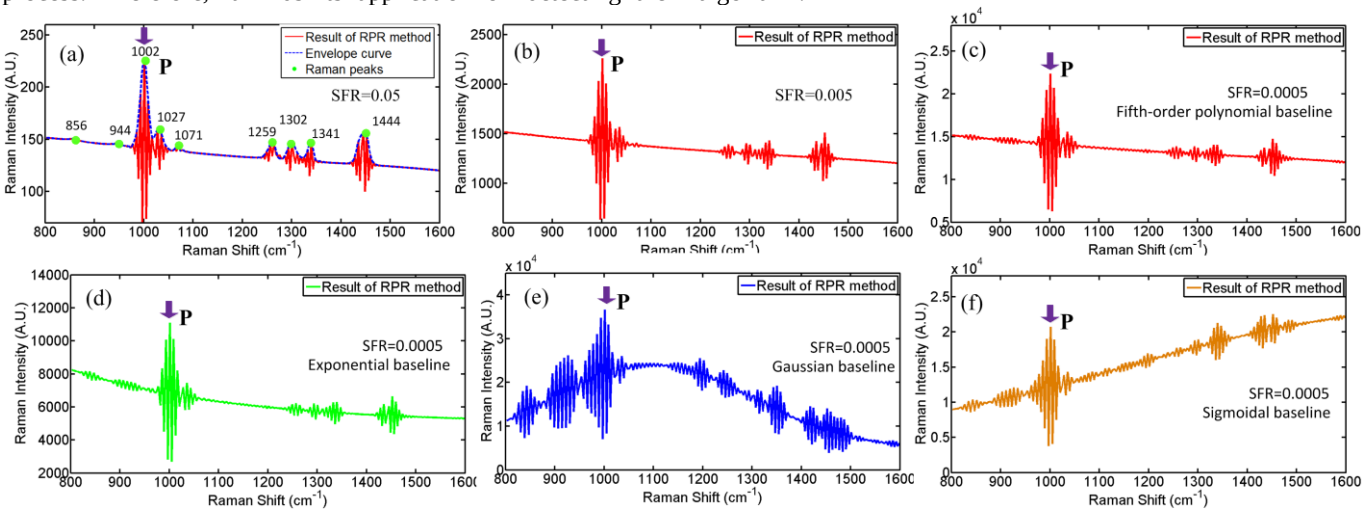


Fig.3 (a), (b) and (c): Recognition results of RPR method after 100 iterations based on fifth-order polynomial baseline with SFR=0.05, SFR=0.005 and SFR=0.0005, respectively. (d), (e) and (f): Recognition results of RPR method after 100 iterations based on different baselines with SFR=0.0005. P represents the optimal Raman peak position which is chosen to serve as the intrinsic criterion of convergence for the subsequent iterative procedure.

RIA-SG-RPR algorithm

RIA and RIA-SG-SR algorithms are iterative smoothing of the measured raw Raman spectrum. Compared with the RIA algorithm, the initial Savitzky-Golay smoothing method is replaced by the SG-SR methods in the RIA-SG-SR algorithm to achieve additional improvement in the convergence speed over the Savitzky-Golay procedure. The details of RIA and RIA-SG-SR algorithms can be found in Ref. 24 and Ref. 25.

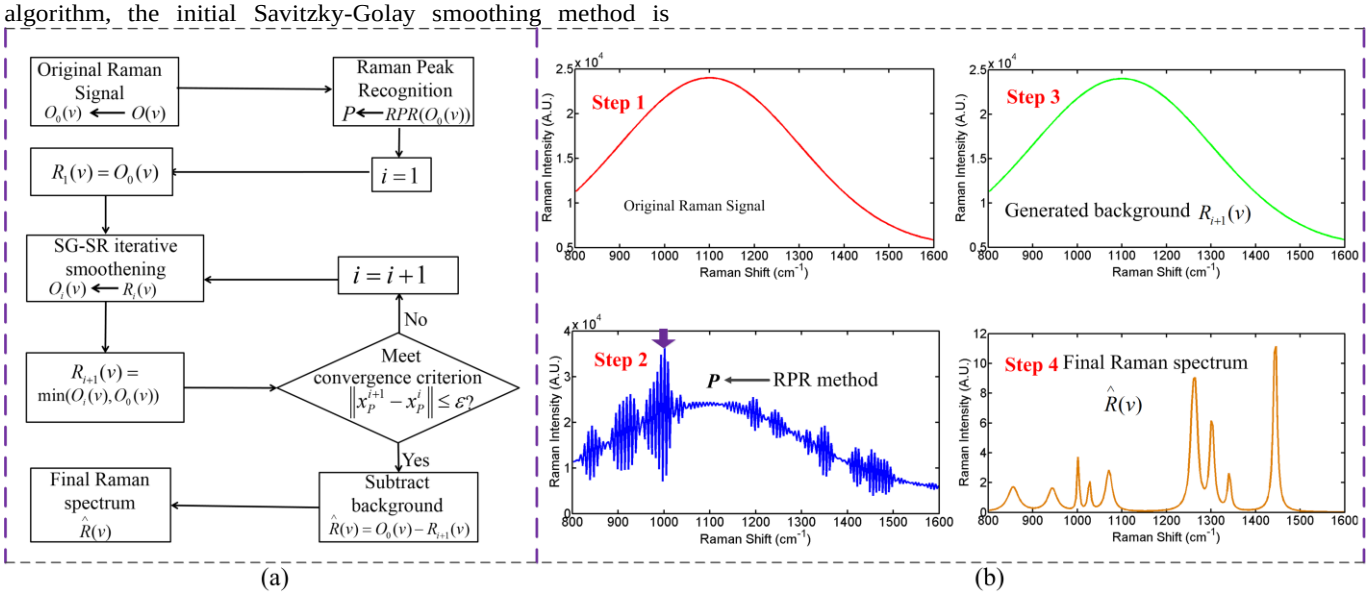


Fig.4 (a): Flowchart of the RIA-SG-RPR algorithm for background subtraction. (b): Pictorial demonstration of the working of the RIA-SG-RPR based on Gaussian baseline with SFR of 0.0005.

For practical automated fluorescence rejection, subjective direct human intervention must be minimized, and thus the concerns with the two algorithms must be addressed to obtain more truly representative Raman spectra. However, it's impossible for RIA and RIA-SG-SR algorithm to avoid human interference because of the two extra artificial Raman peaks served as the criterion of convergence in the iterative procedure. An initial Raman peak recognition can find a way to the artifact-free recovery of Raman spectra in RIA-SG-RPR algorithm as well as good rejection of low Signal-to-Fluorescence spectra. The underlying basis of the RIA-SG-RPR is iterative smoothing of the measured raw Raman spectrum. The algorithm uses a model based on the initial Raman peak recognition (RPR) method and the improved SG-SR iterative smoothing of the measured Raman spectrum. A detailed layout of the RIA-SG-SR algorithm is shown in Fig. 4(a). The first step is to perform the Raman peak recognition (using the RPR method discussed before) of the input spectral data to derive the positions of the maximum Raman peak, P, as the criterion of convergence. Following Raman peak recognition, the whole of

the input spectrum is subjected to a modified iterative smoothing based on the SG-SR method. The SG-SR method is equivalent to a low-pass filter, which tends to filter out the high-frequency components of a signal leaving the low-frequency baseline intact. The process of smoothing by SG-SR is iterated until the convergence criterion is met, which means $\left\| x_P^{i+1} - x_P^i \right\| \le \varepsilon$. Here, x_P^i represents the value of P position after the i-th iteration. Then this output curve, representing the background, is subtracted from the raw spectrum to obtain the true Raman spectrum with zero background. In this article, ε is chosen as 0.00001 to ensure that the Raman signal has been completely removed that has been demonstrated in our practical experiment. Fig. 4(b) provides a pictorial demonstration of the working of the RIA-SG-RPR applied on the mathematically generated raw Raman spectrum based on Gaussian baseline with SFR of 0.0005.

Results and Discussions

Rejection of low SFRs based on different baselines

A background subtraction algorithm is desired to accurately recover the Raman signal from the raw Raman spectrum irrespective of the SFRs. However, in practice, low SFRs always pose challenges for background subtraction. Both RIA and RIA-SG-SR result in distortions when they are applied on low SFR spectra, especially in the applications on biological samples with tense fluorescence. In contrast, RIA-SG-RPR algorithm provides a more powerful rejection of low SFRs.

To analyse the performance of the three different algorithms (RIA, RIA-SG-SR, and RIA-SG-RPR) with respect to SFRs, they were applied on the mathematically generated raw Raman spectra of various SFRs, ranging from relatively high to much lower values. Fig. 5(a)–(c) shows the recovered Raman spectra after processing. It is apparent from the figures that in all the situations the Raman peaks subtracted by RIA-SG-RPR algorithm have been faithfully retrieved without any distortions. However, the performances of the RIA and RIA-SG-SR vary with SFR. It is found that Raman peaks could be faithfully retrieved without any distortions up to SFR of ≥0.05, shown in Fig. 5(a) and (b), using RIA and RIA-SG-SR. With SFR below ~0.05, the subtractions of the two algorithms are found to lead to distortions and false peaks in the recovered spectrum, which is similar to the discussions in Ref. 24. The similar results are also found in Fig. 5(d) and (e) and (f), based on the three other baselines with low SFRs. It should be noted that the major Raman lines after RIA-SG-RPR processing show a broad shoulder on the left side. These little artifacts can be ascribed to the modified small-window moving technique. In the small-window moving process, subsequent values shows a better approximation to the actual solutions compared with the previous ones. These little artifacts are negligible, because the presented method substantially obtains the consistent removal of Raman peak with correct height, width and locations which play a more important role in characterizing molecule.

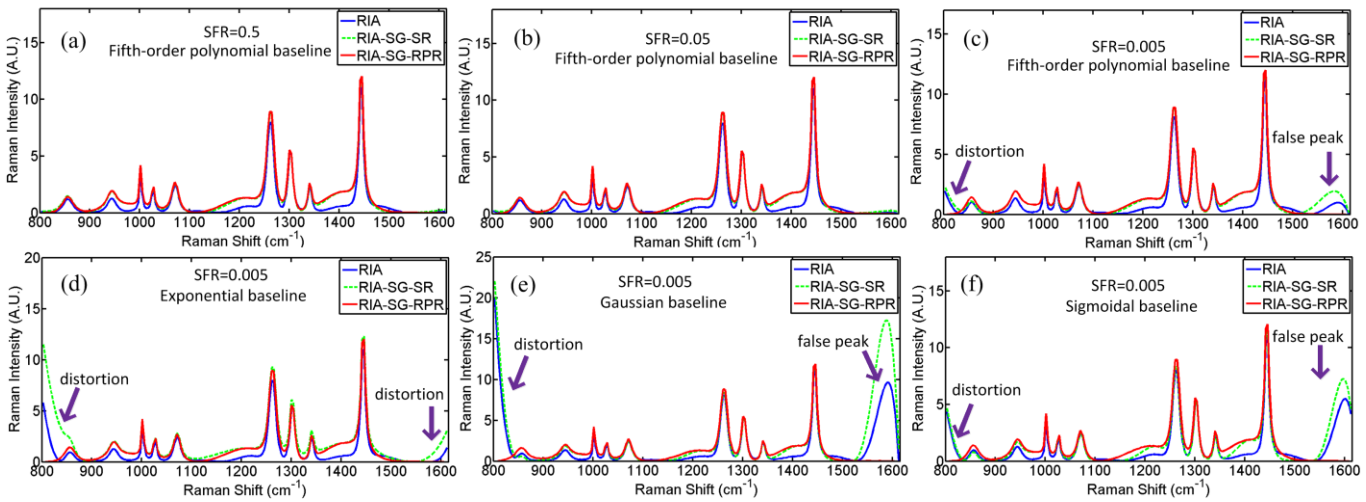


Fig.5 Comparison of different subtraction algorithms with respect to different SFRs and different baselines. (a), (b) and (c) are recovered Raman spectra of three different raw spectra (fifth-order polynomial baseline with three different SFRs), using RIA, RIA-SG-SR and RIA-SG-RPR, respectively. (d), (e) and (f) are recovered Raman spectra of different raw spectra based on three different baselines (exponential, Gaussian and sigmoidal) with SFR of 0.005, using RIA, RIA-SG-SR and RIA-SG-RPR, respectively.

To measure the merits of the recovered spectra, the quantitative performance assessments were evaluated on the basis of the following merits between the true spectrum R^o and the recovered spectrum R: the root of mean square error $RMSE=\sqrt{\sum_{i=0}^{N}(R_i^o - R_i)^2 / N}$, the Pearson's correlation coefficient $CC=\frac{\sum_{i=1}^{N}(R_i-\bar{R}_i)(R_i^o-\bar{R}_i^o)}{\sqrt{\sum_{i=1}^{N}(R_i-\bar{R}_i)^2}\sqrt{\sum_{i=1}^{N}(R_i^o-\bar{R}_i^o)^2}}$ and the self-weighted correlation coefficient[29] $WCC=\frac{\sum_{i=1}^{N} w_i(R_i-\bar{R}_i)(R_i^o-\bar{R}_i^o)}{\sqrt{\sum_{i=1}^{N} w_i(R_i-\bar{R}_i)^2}\sqrt{\sum_{i=1}^{N} w_i(R_i^o-\bar{R}_i^o)^2}}$. $\bar{R}_i$ and $\bar{R}_i^o$ are the mean of corresponding spectra in *CC*, while they are weighted mean defined as $\bar{R}_i=\sum w_i R_i / \sum w_i$ and $\bar{R}_i^o=\sum w_i R_i^o / \sum w_i$ in *WCC*. *RMSE* represents the average difference between the two spectra, with a small *RMSE*

corresponding to a good match. *Pearson's CC* represents the average similarity between the trends of the true and the recovered spectra, and the larger values denote a better match. As an improved version of *Pearson's CC*, *Griffiths' WCC* places emphasis on those Raman bands in the spectrum and thus can obtain a more reliable measure of the similarity, which is also consistent with visual comparison. Table I shows the resulting *RMSE*, *CC*, and *WCC* for each method on the simulated data with different SFRs. When the SFR is equal to 0.05, all the three methods achieve almost the same evaluations. However, for SFR below 0.05, the case is opposite. It shows that RIA-SG-RPR suppresses distortions and false peaks better than RIA and RIA-SG-SR, especially based on the Gaussian and sigmoidal baselines where the *RMSE* of the latter two methods are much larger than that of the proposed method while the *CC* and *WCC* is in the opposite case. These quantitative evaluations are consistent with visual assessment and show superiority over RIA and RIA-SG-SR, especially with complicated non-linear background. As mentioned above, the higher *WCC* means a better match between the Raman bands in the two compared spectra which actually reveal the "Molecular fingerprint". That is, RIA-SG-RPR provide more exact spectral details. Considering the possible erroneous judgment leading by the false peaks, one can benefit from the high *WCC*, which is of crucial importance in the application of Raman spectroscopy.

Table 1 Comparison of the performance of RIA, RIA-SG-SR and RIA-SG-RPR in two SFR conditions. For CC and WCC, higher values imply superior performance, while lower RMSE values imply improved performance.

Merits	Method	RMSE		CC		WCC	
		0.05	0.005	0.05	0.005	0.05	0.005
Fifth-order polynomial	RIA	0.3525	0.4292	0.9846	0.9708	0.9830	0.9811
	RIA-SG-SR	0.3586	0.7592	0.9553	0.9242	0.9812	0.9798
	RIA-SG-RPR	0.3342	0.3321	0.9844	0.9845	0.9870	0.9871
Exponential	RIA	0.3612	**0.6814**	0.9837	**0.9204**	0.9828	**0.9769**
	RIA-SG-SR	0.3829	**1.6905**	0.9541	**0.7326**	0.9813	**0.7603**
	RIA-SG-RPR	0.3345	0.3351	0.9844	0.9846	0.9870	0.9870
Gaussian	RIA	0.3626	**2.8488**	0.9825	**0.3872**	0.9830	**0.1057**
	RIA-SG-SR	0.3932	**4.4622**	0.9530	**0.1976**	0.9809	**0.1252**
	RIA-SG-RPR	0.3442	0.3237	0.9845	0.9867	0.9870	0.9876
Sigmoidal	RIA	0.3628	**1.2088**	0.9821	**0.7711**	0.9829	**0.8574**
	RIA-SG-SR	0.3741	**1.6569**	0.9520	**0.6571**	0.9814	**0.6942**
	RIA-SG-RPR	0.3349	0.3403	0.9843	0.9843	0.9870	0.9870

As is known to all, a linear least square fit (using a polynomial of degree one) of linear regression of the input spectral data is used to extrapolate two sets of linear data at the two ends of the selected portion of the spectral range in RIA and RIA-SG-SR. After that, two Gaussian peaks, one on each side, are added to the extended linear portion to serve as the criterion of convergence for the subsequent iterative procedure. This extension can be used to explain these distortions and false peaks on each side of the recovered spectrum. Linear fit of the input spectral data is only a crude approximation of the baseline, which causes the discontinuity of the hybrid curve. This discontinuity is further amplified in the iterative process and regarded as a "Raman peak" because of its high frequency. This hypothesis is clearly manifested in Fig. 5 and Table I, especially under the condition of complicated background (exponential, Gaussian and sigmoidal baselines in Fig. 6(d)-(f)), because it's more difficult to fit these non-linear baselines using just a linear polynomial. One can further deduce that, results will get worse when these two methods are applied to a more complicated input data experimentally. However, RIA-SG-RPR is totally free of this problem.

Artifact-free and automated fluorescence subtraction

In the case of the RIA and RIA-SG-SR, successful use of the two algorithms require appropriate selection of a number of different parameters used for the criterion of convergence, including the heights and widths (FWHM) of the two added Gauss peaks. However, this involves user intervention and is clearly a disadvantage for its automated use. According to Ref. [24], a simple criterion is that the height of the Gaussian peak is chosen as equal to the maximum of the ordinate values of the raw spectrum to guarantee that the recovery of all the Raman peaks is completed. However, it doesn't mean this selection is optimal, because the maximum of the ordinate values is generally much larger than the real heights of Raman peaks and the computing cost multiplies at an astonishing rate with the increase of the heights in our processing. While, a smaller Gauss peak means that the iterative smoothing operation would be terminated before the Raman peaks are recovered fully.

Fig. 6 illustrates that human intervention to specify the height and width of the two artificial peaks have direct severe impact on the recovered Raman signals in the practical application of RIA and RIA-SG-SR. Normal height and width were determined after several prospective attempts to guarantee that the Raman peaks are recovered fully. However, it's difficult to determine appropriate height and width facing with measured Raman spectra since no one knows the real Raman peaks. Deviation from normal values can lead to great distortions, shown in Fig. 6(b) and (c). Whereas the results of RIA-SG-RPR algorithm can keep stable without the human interference. In addition, the position where the peaks should be added on is also depended on the user intervention which can further lead to inaccuracy of background subtraction.

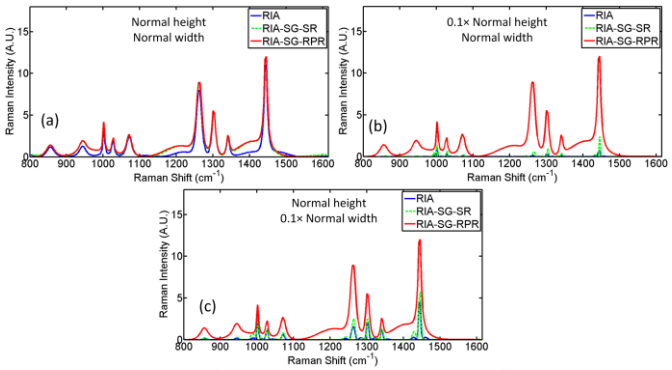


Fig.6 The variations of the recovered spectra with different heights and widths of the two false Gaussian peaks for RIA and RIA-SG-SR. Normal height and width were determined after several prospective attempts to guarantee that the Raman peaks are recovered fully.

In our proposed algorithm, only one parameter, the SFR should be specified by direct human intervention in preprocessing. It can be approximatively estimated by visual inspection. After that, *w* is obtained from Eq. 7. Actually, since different SFR leads to little difference of *w*, this user intervention will not bring as great influence as the RIA to the recovered Raman signals. So, it is pertinent to mention here that RIA-SG-RPR algorithm can provide a more robust, automated and artifact-free recovered Raman spectra compared with RIA and RIA-SG-SR.

Effects of noise on RIA-SG-RPR

Synthetic spectra were generated with a signal-to-noise ratio (SNR) of 10, which is a reasonably challenging level of noise, by adding a constant level of Gaussian (white) noise with an original standard deviation of 1.0 to the spectrum. These baselines permitted us to investigate the RIA-SG-RPR's performance on a variety of dissimilar baselines to gain quantitative estimates of its performance. First, we want to demonstrate the robust rejection of noise of RPR method. Fig. 7(a)-(d) give the results of peak recognition with respect to different baselines. It is clear that, all the optimal Raman peak (P) have been found, exactly the same as Fig. 3. Then, an automated smoothing[26] was applied on the raw spectrum to reduce interference by noise. Followed by the RIA-SG-RPR algorithm, the recovered Raman signal can be seen in Fig. 7(a), (b), (c) and (d). Due to the smoothing process for reducing noise, the intensity of each Raman peak decreases, compared to the true Raman signals in Fig. 1 (d). What's more, the noise removal prior to fluorescence subtraction algorithm is found to lead to distortions and the emergence of small false peaks in the recovered spectrum. So, controlling the noise is very important in the previous measurement. In spite of the interference by noise, this algorithm is found to substantially obtain the consistent removal of Raman peak with correct profile and locations.

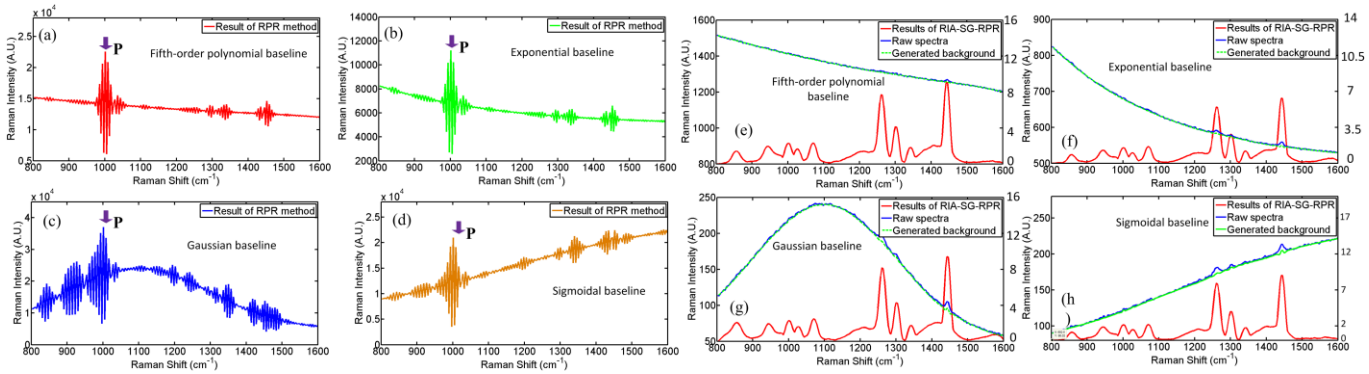


Fig.7 (a)-(d): the Raman peak recognition results based on four baseline types with constant SNR of 10. (e)-(f): the recovered Raman signal based on four baseline types with constant SNR of 10.

Recovered Raman signals from measured spectra

In order to demonstrate the utility of the proposed fluorescence removal algorithms, experiments were carried out to measure the Raman spectra of chemical agents.

First of all, qualitative demonstrations of the applicability of this propose method is shown in Fig. 8 by applying it to carbon tetrachloride, solid triacontanol and porcine skin in vitro, respectively. CCl_4 hardly exhibits fluorescence with the emission wavelength at 785nm. The subtraction algorithm just correct baseline drift (e.g., shot noise) slightly, but to keep the original spectral contours and intensities over the entire range in Fig. 8 (a), which proves that, RIA-SG-RPR can be used in the common case where the fluorescence is not strong. In contrast, triacontanol and porcine are two highly fluorescent samples. Porcine skin is a very difficult sample to measure not only due to its intense fluorescence, but also due to the complex Raman bands emitted from diversified proteins and amino acids.

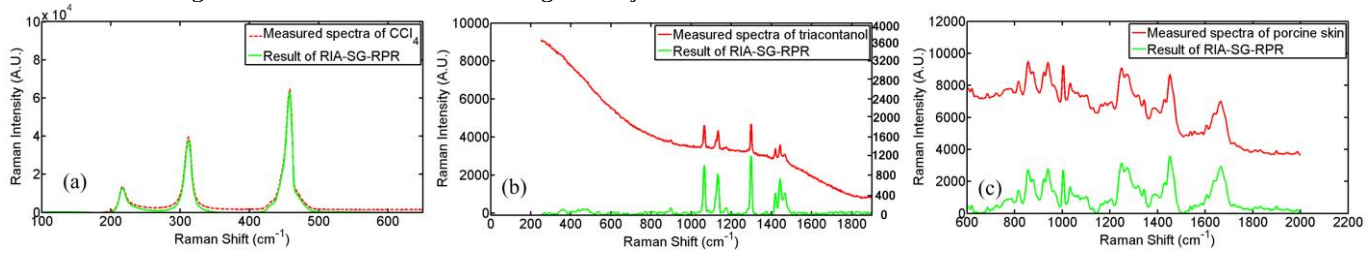


Fig.8 The experimentally measured raw Raman spectra (red line) and the recovered Raman spectra using RIA-SG-RPR(green line) of (a) carbon tetrachloride, (b) solid triacontanol and (c) porcine skin, respectively.

In Fig. 8 (b) and (c), the recovered Raman spectra both show dramatically reduction of fluorescence and Raman signals are more resolvable. More importantly, further quantitative analysis based on the Raman intensity can be implemented after this subtraction. In the more challenging Raman application, all the Raman characteristics of porcine skin[30] have been faithfully retrieved without any artifacts being introduced.

The most important point of RIA-SG-RPR in the article is the rejection of high fluorescence background. So, a well-designed experiment is unitized to illustrate this issue quantitatively, based a synthetic sample. IR775 is a near-IR laser dye with absorption maximum of 775 nm and strongly fluorescent when excited with a 785 nm laser. A trace amount of the IR775 was dissolved in pure ethanol to form a fluorescent sample with a known Raman spectrum. Since the concentration of the ethanol is many orders of magnitude larger than the dye, it is assumed that the dye's contribution to the Raman spectrum is negligible. The mixed solution of Ethanol and IR775 serve as samples with different SFRs, from relatively high to low by changing the concentration of IR775 in the mixed solution, with no immediately resolvable Raman peaks. This experimental system was also used in our previous work[25].

The experimentally measured raw Raman spectra from the mixed solutions are shown in Fig 9(a), (b) and (c). As the concentration increased from 10^{-6}M to 10^{-4}M, the SFRs decreased from 0.0666 to 0.00515. The recovered Raman spectrum of ethanol following processing with RIA and RIA-SG-RPR are also shown in Fig. 9. RIA leads to the emergence of false peaks on the two sides of the recovered spectra when the SFR drops. Contrary to RIA, it is clear from the figures that all the Raman characteristics of ethanol[31] have been faithfully retrieved without any artifacts being introduced. It can be concluded that RIA-SG-RPR algorithm can provide better rejection of the low SFRs compared to other fluorescence subtraction methods. Although it is claimed that RIA could faithfully retrieve without any distortions up to an SFR value of ≥ 0.005 in Ref. 24, false peak also occurs when the SFR decreases to ~0.0106 in Fig. 9 (b). Considering the typically low SFR of the raw Raman spectrum collected from biological tissue or other fluorescent samples, false peaks will significantly limit the use of RIA, and RIA-SG-SR, whereas RIA-SG-RPR algorithm can provide robust and faithful Raman processing.

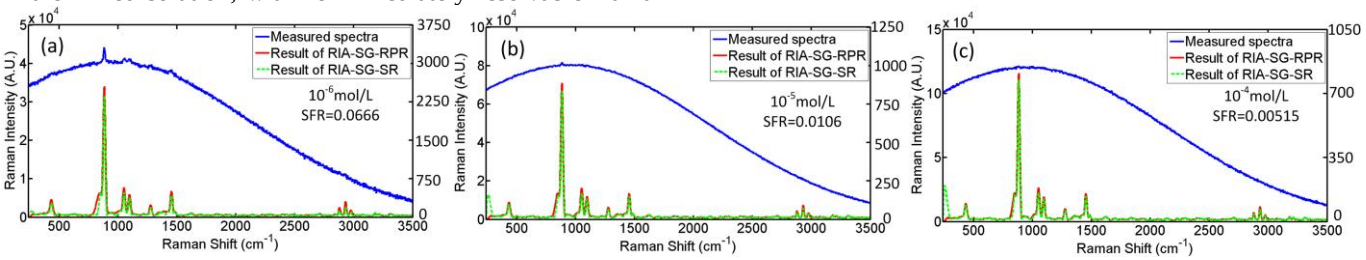


Fig.9 The experimentally measured raw Raman spectra (blue line) and the recovered Raman spectra using RIA-SG-RPR(red line) and RIA (green line) for solution of (a) 1×10^{-6} M, (b) 1×10^{-5} M and (c) 1×10^{-4} M near-IR laser dye IR775 dissolved in ethanol, respectively.

For various applications in general and tissue diagnostic applications and fluorescent chemical agents in particular, it is often required to background subtract a large number of Raman spectra measured experimentally from highly fluorescent samples. In such situations, it is always desirable to have a background subtraction algorithm that can provide an efficient, automated and rapid recovery of Raman spectra with respect to different SFRs. Besides, false Raman peaks should be drastically removed which reveal the nonexistent "molecular fingerprints" and bring confusion and misunderstanding to analytical chemistry, materials and biomedicine. Though there exists little artifacts, RIA-SG-RPR algorithm shows significant improvement on the artifact-free, automated subtraction and rejection of false Raman features to benefit the utilization of Raman spectroscopy to characterize molecular specifics, compared with RIA, RIA-SG-RPR algorithm in these challenging Raman applications. Besides, this algorithm still achieves two orders of magnitude reduction in computation time compared with RIA thanks to the improved iterative procedure of SG-SR method. It can achieve a rapid recovery of Raman Spectra within a few of milliseconds, from which real-time chemical analysis application can benefit a lot.

Conclusions

A novel Raman peak recognition method based fluorescence subtraction algorithm (RIA-SG-RPR) that substantially improves the rejection of low signal-to-fluorescence ratio is presented. The fundamental idea of this peak recognition method is to extract the high-frequency components (Raman signals) through a specific iterative process. This innovation also avoids direct human intervention to obtain more reliable and robust recovered Raman signals. This algorithm is optimized and validated with mathematically simulated Raman spectrum as well as experimentally measured Raman spectra from varied fluorescent samples. In the simulation, the Raman signals have been faithfully retrieved with almost fairly consistent spectral contours and intensities when the SFR drops to 0.005 with different type of baselines. Furthermore RIA-SG-RPR algorithm recovered the real Raman signals in the experimentally measured raw Raman spectra of biological samples with intense fluorescence. Compared with RIA and RIA-SG-SR, RIA-SG-RPR can be applied to those highly fluorescent samples with low SFRs. Considering the typically low SFR of the raw Raman spectrum collected from biological tissue or other fluorescent samples in general, RIA-SG-RPR algorithm can provide more automated and faithful Raman processing.

Acknowledgements

This work is funded by the State Key Lab of Precision Measurement Technology & Instrument of Tsinghua University, Tsinghua University Initiative Scientific Research Program and the National Natural Science Foundation of China (Grant No. 61205147).

Notes and references

[a] State Key Lab of Precision Measurement Technology & Instrument, Department of Precision Instrument, Tsinghua University, Beijing 100084, China. E-mail: luckiwei@mail.tsinghua.edu.cn

1 G. Clemens, J. R. Hands, K. M. Dorling and M. J. Baker, *Analyst.* 2014, **139**, 4411.
2 K. Buckley, P. Matousek, *Analyst.* 2011, **136**, 3039.
3 A. Walter, S. Erdmann, T. Bocklitz, E. Jung, N. Voqler, D. Akimov, B. Dietzek, P. Rosch, E. Kothe and J. Popp, *Analyst.* 2010, **135**, 908.
4 S. K. The, W. Zheng, D. P. Lau and Z. Huang, *Analyst.* 2009, **134**, 1232.
5 M. Maiwald, G. Erbert, A. Klehr, H.D. Kronfeldt, H. Schmidt, B. Sumpf and G. Tr¨ankle, *Appl. Phys. B.* 2006, **85**, 509.
6 J. Zhao, M. M. Carrabba, F. S. Allen, *Appl. Spectrosc*. 2002, **56**, 834.
7 S. T. McCain, R. M. Willett, D. J. Brady, *Opt. Express*. 2008, **16**, 10975.
8 J. B. Cooper, M. Abdelkader, K. L. Wise, *Appl. Spectrosc*. 2013, **67**, 973.
9 J. B. Cooper, S. Marshall, R. Jones, M. Abdelkader, K. L. Wise, *Applied Optics*, 2014, **53**, 3333.
10 M. Maiwald, H, Schmidt, B, Sumpf, G. Erbert, H.-D. Kronfeldt, G. Tränkle, *Applied Optics* 2009, **48**, 2789.
11 M. A. Martins, D. G. Ribeiro, E. A. P. Santos, A. A. Martin, A. Fontes, H. S. Martinho, *Biomed. Opt. Express*. 2010, **1**, 617.
12 P. P. Yaney, *JOSA*. 1972, **62**, 1297.
13 P. Matousek, M. Towrie, A. Parker, *J. Raman Spectrosc*. 2002, **34**, 238.
14 C. A. Lieber, A. Mahadevan-Jansen, *Appl. Spectrosc*. 2003, **57**, 1363.
15 P. A. Mosier-Boss, S. Lieberman, R. Newbery, *Appl. Spectrosc.* 1995, **49**, 630.
16 B. D. Prakash, Y. C. Wei. *Analyst.* 2011, **136**, 3130.
17 Z. M. Zhang, S. Chen, Y. Z. Liang, Z. Liu, Q. Zhang, L. Ding, F. Ye and H. Zhou, *J. Raman Spectrosc.* 2010, **41**, 659.
18 Z. M. Zhang, S. Chen and Y. Z. Liang, *Analyst.* 2009, **135**, 1138.
19 T. J. Vickers, R. E. Wambles, C. K. Mann, *Appl. Spectrosc.* 2001, **55**, 389.
20 J. Zhao, H. Lui, D. I. McLean, H. Zeng, *Appl. Spectrosc.* 2007, **61**, 1225.
21 M. N. Leger, A. G. Ryder, *Appl. Spectrosc*. 2006, **60**, 182.
22 H. G. Schulze, R. B. Foist, K. Okuda, A. Ivanov, R. F. B. Turner, *Appl. Spectrosc*. 2012, **66**, 757.
23 P. J. Cadusch, M. M. Hlaing, S. A. Wade, S. L. McArthur, P. R. Stoddart, *J. Raman Spectrosc.* 2013, **44**, 1587.
24 H. Krishna, S. K. Majumder, P. K. Gupta, *J. Raman Spectrosc.* 2012, **43**, 1884.
25 K. Chen, H. Y. Zhang, H. Y. Wei, Y. Li, *Applied Optics*. 2014, **53**, 5559.
26 H. G. Schulze, R. B. Foist, A. Ivanov, R. F. B. Turner, *Appl. Spectrosc*. 2008, **62**, 1160.
27 R. J. Meier. *Vib. Spectrosc*. 2005, **39**, 266.
28 A. Savitzky, M. J. E. Golay, *Anal. Chem*. 1964, **36**, 1627.
29 P.R. Griffiths and L. Shao. *Appl. Spectrosc*. 2009. **63**, 919.
30 D. Huang, W. Zhang, H. Zhong, H. Xiong, X. Guo, Z. Guo, *J. Biomed. Opt.* 2012. **17,** 015004
31 C. Meneghini, S. Caron, A. Proulx, A. Proulx, F. Emond, P. Paradis, C. Pare, A. Fougeres, *Sensors Journa., IEEE.* 2008, **8**, 1250.