Iterative morphological and mollifier based baseline correction for Raman spectra

Matthias Koch, Christian Suhr, Bernhard Roth, Merve Meinhardt-Wollweber *

June 6, 2016

Abstract

In vivo Raman spectroscopy with low signal-to-noise ratio and strong, irregularly shaped fluorescence background imposes a challenge for automatic baseline correction methods. In this work, an approach that enables fast and efficient batch baseline correction has been developed which is based on a morphological operation in combination with a mollifier algorithm. As this algorithm relies only on three parameters which are determined by the given experimental conditions, it can be used for automatic and objective processing of many Raman spectra. The applicability of the baseline correction is demonstrated on resonance Raman spectra of beta-carotene mixed with fluorescent red ink as model system, on carotenoids in human skin and on an excitation-emission map of the green alga Haematococcus pluvialis. In the future, the algorithm opens the potential for wide application in Raman spectra analysis in biological contexts. In particular, it greatly facilitates data processing in cases where special photochemical sample preparation or complex experimental baseline removal were required before. Similarly, processing data of experiments using resonant excitation techniques yielding strong fluorescence background is possible.

Keywords: Baseline correction; Fluorescence removal; Morphological; Mollifier; *In vivo*

*Hannover Centre for Optical Technologies, Leibniz University Hannover, Hannover, Germany

1 Introduction

Raman spectroscopy is a powerful and established analytic tool for many disciplines. It specifically probes the molecular vibrations of a given sample by detection of the inelastically scattered photons, also called molecular fingerprints, and is widely used for label-free analysis of technical or biological samples. Background correction is essential to extract Raman signals from the raw spectra measured which are usually compromised by background signals most often originating from fluorescence.

There are several approaches that strive to improve background conditions already at the experimental stage. Experimental methods involving shifted excitation spectral differences[1–3] follow the idea to compare Raman spectra at two slightly different excitation wavelengths. This approach takes advantage of the fact that fluorescence excitation is broadband and independent of the exact excitation wavelength while the Raman lines will follow the excitation shift. One drawback of this method is a relatively complex instrumentation because two excitation wavelengths are required. The background-free, baseline corrected Raman spectrum then can be reconstructed from the difference of both raw spectra, but the underlying assumptions may not always hold true for *in vivo* samples because each spectrum taken could have another baseline which changes with accumulated irradiation, time and wavelength. Furthermore, when applying this method under resonance conditions, the intensity of Raman lines may change even for a small excitation wavelength shift[4].

Bleaching of the fluorophores[5–7] with high beam power over long time periods cannot regularly be used for baseline correction because concurrent bleaching of a set of fluorophores as present in *in vivo* samples can hardly be achieved. Besides, it is highly probable that such strong irradiation would change the sample also affecting the Raman active components and the further behaviour of the sample.

Time gated approaches[3,8], which differentiate fluorescence background and Raman scattering by their different interaction times[9], need very high pulsed intensities which may alter the sample and are, due to their need for gated detection systems, very expensive.

Apart from experimental approaches, mathematical postprocessing is also possible and even much more in use due to lower experimental demands. Schulze *et al.*[10] give an overview on classic baseline correction algorithms and their limitations. Manually tuned or assisted methods cannot be applied to a large data set within reasonable time and are difficult to evaluate because of the strong operator dependence. Some algorithms such as artifical neural networks or polynomial curve fitting fit a mathematical model of the baseline, but as very little is known *a priori* on the possibly irregularly shaped baseline in general, algorithms that do

not use a model of the baseline shape are desirable. Also, noise may vary considerably, so that noise median based methods cannot be applied either. Using first derivative methods, one difficulty is to choose an appropriate threshold to reliably detect peaks, which is especially problematic in low signal-to-noise spectra. To avoid severe artifacts, Fourier-based algorithms need well tuned filter parameters, which are difficult to determine when little is known on the signals to be expected.

In our own work with biological samples, we obtained resonance Raman spectra with very low signal-to-noise ratio due to small *in vivo* concentrations of Raman active molecules such as carotenoids and with strong arbitrarily shaped fluorescence backgrounds which pose a challenge on common baseline removal algorithms. In addition, studies on samples with unknown content require searching for unknown lines, which need to be objectively quantified and compared across a large data set.

More advanced iterative polynomial algorithms like the approaches as reported in[11] and[12] make sure that Raman signals present in the spectra will be preserved under all circumstances, which renders them an excellent choice for human inspection afterwards. However irregularly shaped experimental baselines that cannot be approximated well by a polynomial make comparisons within large datasets complicated. There are many variations based on polynomial fits, which, for example, use variable polynomial orders[13]. Peak recognition based methods like[14] or peak stripping involving methods like[15] depend heavily on peak characteristics or statistical properties, which may not hold true under all conditions, as will be shown in figure 7b. Furthermore, methods that involve highly advanced mathematical methods like wavelet transformation[16] have been proposed, but the lack of geometrical descriptiveness may cause the spectroscopists to rather rely on simpler algorithms whose action and possible artifacts can be understood more easily.

One new and interesting approach is the application of morphology operators[17] on Raman baselines, which decide by the width of features whether they are treated as signal or baseline. The parameters of these operators depend mostly on slit width, resolution, line coalescing, and molecular line broadening mechanisms. The paper also describes an algorithm to automatically determine the feature width parameter, which is given by the width of the broadest line or combination of lines. This may eliminate the need for user intervention if the iterative process of determining the feature width succeeds. By its mathematical definition, the described solely morphological approach also can distort the shape of the lines, which can be problematic for some types of Raman analysis. However, it preserves the peak locations precisely.

The algorithm developed in this work is inspired by the work of Bukvic *et al.*[18] and Perez-Pueyo *et al.*[17]. The first approach[18] describes an algorithm for baseline determination in the case of large datasets when most of the points belong

to the baseline which can be described mathematically and only a few data points belong to actual signal lines. It relies on calculating histograms and determining baseline location and noise amplitude by normal distribution properties. Its main strength besides simplicity is that it will give error bounds for the baseline noise and identify signal locations, but it fails if the shape of the baseline cannot be modelled with a pre-known mathematical function or if many points in a spectrum belong to Raman lines. This idea is not applicable to the spectra captured in our experiments, but nevertheless clearly demonstrates the need to take care of the noise of the fluorescent baseline itself.

The second approach[17] is based on morphology operators alone and allows for complete automated baseline removal, but does not always lead to a differentiable calculated baseline and has difficulties determining the feature width on spectra with baseline features with width close to that of coalescent Raman lines. For this reason, the method described in our work combines a morphological operation with mollification and works with an initial user input given by experimental conditions to set the feature width which is automatically detected by an iterative approach in[17].

Our work also has some aspects in common with the approach in[15], which uses an iteratively applied Savitzky-Golay filter, with an effect comparable to a mollifier kernel. However, that algorithm employs peak-stripping and varies the smoothed width with encountered spectral features for a complete automated baseline correction without user-chosen parameters. In contrast, our algorithm requires the user to set the feature width manually, having the benefit that all spectral data sets are processed exactly in the same way, regardless of the actual peaks or baseline features contained therein.

2 Iterative morphological and mollifier based baseline correction algorithm

Before baseline correction can be applied, defective pixels of the CCD / CMOS camera used for the measurements or Rayleigh filter edges need to be removed as these could significantly affect the resulting data in the surrounding pixels through the noise mollification step by their extreme amplitude and, therefore, influence the shape of the baseline or distort actual signals. As hot, cold or defective camera pixels and filter locations are known for most experimental setups, the removal can be done automatically. The feature width, which is used to separate Raman signals from baseline features, should be set to slightly more than the footpoint to footpoint line width of coalescent peaks.

As Raman spectra are typically available as discrete point sampled data, pixels

have been chosen for discretization instead of wavenumbers for algorithmic simplicity. The pixel-to-wavenumber relation is not affine for wavelength calibrated spectrographs and differs slightly for signals on low and high wavenumbers in a single spectrum. However, as the pixels per wavenumber relation does not change by a lot within typical spectral ranges of about 3000 cm^{-1}, this simplification is useful for the calculation task of baseline removal. Additionally, the width of *in vivo* Raman lines mostly depends on the experimental setup and the feature width in the algorithm can be chosen large enough for the widest coalescent lines in the largest possible pixel-per-wavenumber range. If necessary for another type of wide-range measurements with very narrow lines, a wavenumber dependent feature width could be introduced into the algorithm.

The algorithm developed in this work is an iterative approach in which every step involves a morphological operation together with mollification. The mollification is used for two purposes, the reduction of random noise in the Raman spectra and - as detailed later - for smoothing of the calculated baseline. In both cases, spectral intensity data I[x] with x meaning pixel number is convolved with a mollifier kernel

$$m(x) = \begin{cases} exp\left(\frac{-1}{1-x^2}\right) & \text{for } |x| < 1 \\ 0 & \text{else.} \end{cases}$$

The smoothed spectral intensity is calculated as

$$I_{smoothed}[x] = \frac{\sum_{k=1}^{n} I[k] * m\left(\frac{x-k}{w}\right)}{\sum_{k=1}^{n} m\left(\frac{x-k}{w}\right)}$$

with k running from the first to the last pixel (n) of the dataset and w being the width of the mollifier.

In a first step this mollification is applied to the raw spectral data to reduce random noise, so w should be chosen according to the noise found for the experimental setup. Typically, this is a few pixels wide. On the borders of the spectra the mollifier kernel runs into void and has to be renormalized, which is done in the denominator.

In the second step, a pre-baseline is determined by choosing the pixels with lowest intensity within the feature width in the noise-smoothed experimental data set. In this way, the lowest pixel intensity within half the feature width to the left or to the right determines the intensity of the pixel in the pre-baseline. This can be imagined as fitting a horizontal line into the experimental spectrum from the bottom, with the contact point height giving a single point in the pre-baseline. The pre-baseline will always be lower or of same height as the experimental data and if the width of this line is chosen accordingly, it will not penetrate into and affect Raman lines or multiple coalescent Raman lines, as illustrated in detail in Fig. 1.

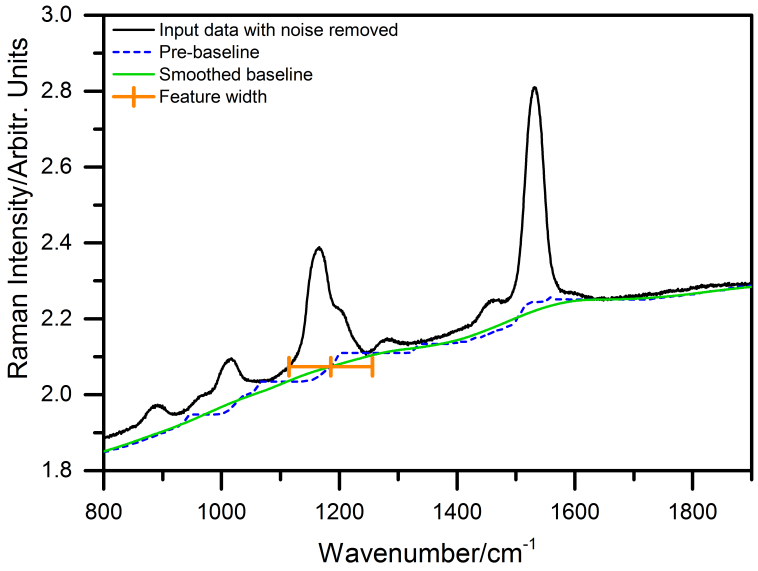


Figure 1: In each iteration step, pre-baseline values are determined as the lowest values within the chosen feature width of 180 pixels. Smoothing the resulting pre-baseline gives an approximation of the true baseline that is subtracted afterwards. Data from sample (1).

As the pre-baseline is rough and contains jumps, in a third step, it is convolved with a mollifier kernel as described above whose width w is equal to the feature width, i.e. the length of the horizontal line chosen before. Care has to be taken that the kernel is renormalized while running into void at both ends of the data set. This gives a mollified baseline which is lower or equal in most cases than the experimental pixel intensities excluding noise and - most importantly - does not affect the signals.

The mollified baseline (see Fig. 1) is then subtracted from the original noise-containing pixel intensities to preserve the original noise shape. The algorithm is iterated with the latest subtraction result as input data to reach the final baseline. Fig. 2 shows five iteration steps with a feature width of 180 pixels.

A flowchart of the individual processing steps is shown in Fig. S1. The complete source code of this algorithm in Pascal and its implementation in Matlab are available under GPL3 as a download (ramanbaseline.sourceforge.net).

3 Experimental aspects

In order to test and evaluate the proposed algorithm under controlled experimental conditions, samples containing the well-known Raman active molecule beta-

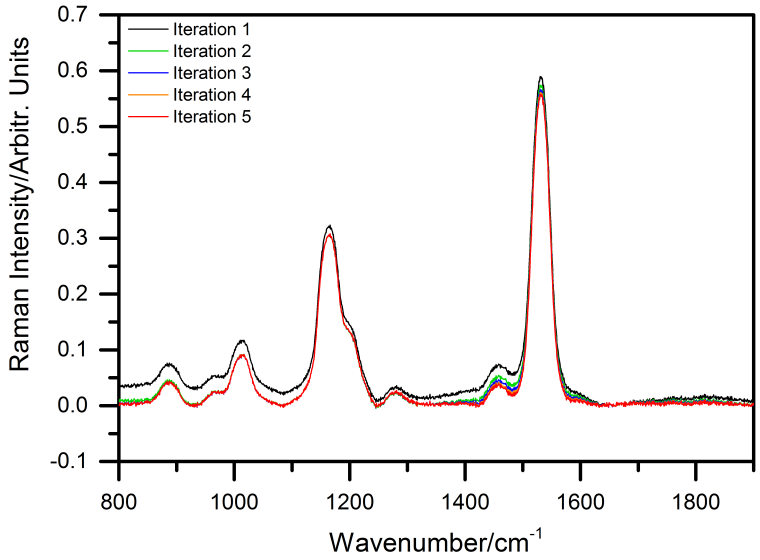


Figure 2: Convergence of the first five iteration steps applied to experimental data. The original raw data is shown in Fig. 1.

carotene and different amounts of ink showing very strong fluorescence were created. Beta-carotene was purchased from Sigma-Aldrich ($> 97.0\%$), ethanol from Carl Roth ($\geq 99.8\%$) and red fountain pen ink from Lamy. As can be concluded from its absorption spectrum, the ink probably contains eosin. All samples were prepared by dissolving 30 µg beta-carotene in 3 ml ethanol with selected amounts of ink added: Sample (1) contains no ink, sample (2) 0.4 µl ink, sample (3) 1.6 µl ink and sample (4) 8.0 µl ink.

Raw spectra were taken with an Andor SR500 spectrograph, equipped with a Semrock 473 nm RazorEdge long-pass filter, a 20 µm slit, a 1200 lines/mm grating and an Andor Newton DU940P camera. On average, the setup maps 0.8 cm^{-1} on every pixel. A Spectra Physics Excelsior single mode 473 nm laser with 50 mW continous wave output was used for excitation. Spectra were taken with 5 s accumulation time through a custom built fiber bundle.

The raw spectral range from 592 cm^{-1} to 2161 cm^{-1} has been used for all calculations, as baseline correction algorithms are sensitive to input wavenumber range, but for clarity, the figures shown here have been clipped to the region from 800 cm^{-1} to 1900 cm^{-1} (shown in Fig. 3), as there are no visible Raman peaks outside of this region. The amount of fluorescence increases with the amount of ink added to the beta-carotene solution. The two strongest Raman lines of beta-carotene at 1158 cm^{-1} and 1527 cm^{-1} are visible in the pure beta-carotene sample - even though they appear weak at the linear scale. With increasing fluorescence background, they become more difficult to perceive.

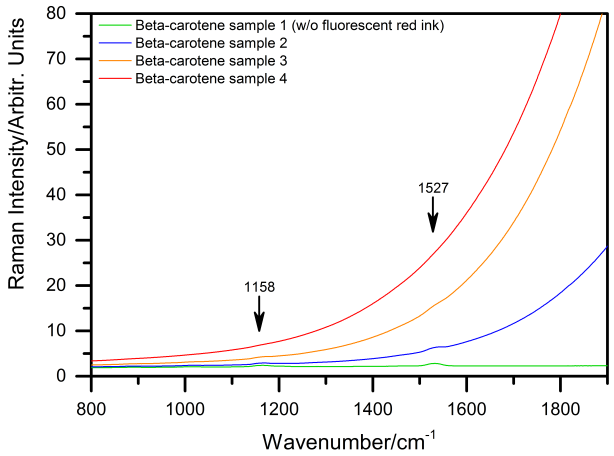


Figure 3: Raw Raman spectra from four samples of beta-carotene in ethanol with increasing amounts of fluorescent red ink used for validation of the developed algorithm.

4 Results and Discussion

4.1 Example spectra

All four raw spectra of beta-carotene in ethanol mixed with different amounts of fluorescent red ink (as shown in Fig. 3) were baseline corrected with the developed algorithm to obtain the data shown in Fig. 4. The noise removal mollifier width was set to six pixels, the feature width to 180 pixels and five iterations steps were applied. In direct comparison of the raw and corrected spectra of sample (1) which contains beta-carotene without ink, both peak locations and intensities are properly preserved. Comparison of this Raman spectrum with the background corrected samples (2), (3) and (4) shows that while the peak positions are preserved, the intensity of the lines decreases significantly with increasing ink concentration because the ink partially absorbs the excitation beam, so that the illuminated volume of beta-carotene in ink-containing samples is smaller than in the sample without ink. Furthermore, the ink reabsorbs the emitted Raman light. Therefore, one cannot assume a simple superposition of the beta-carotene signals when mixing beta-carotene with fluorescent ink. Strong fluorescence also leads to an increased amount of noise. Sample (4) has a very low signal-to-noise ratio, but the two strongest Raman lines of beta-carotene at 1158 cm^{-1} and 1527 cm^{-1} are still identified after background correction. Altough spectra of this quality usually would be discarded, the ability to process spectra with very low signal-to-noise ratio can be useful when measuring concentration curves or when seeking

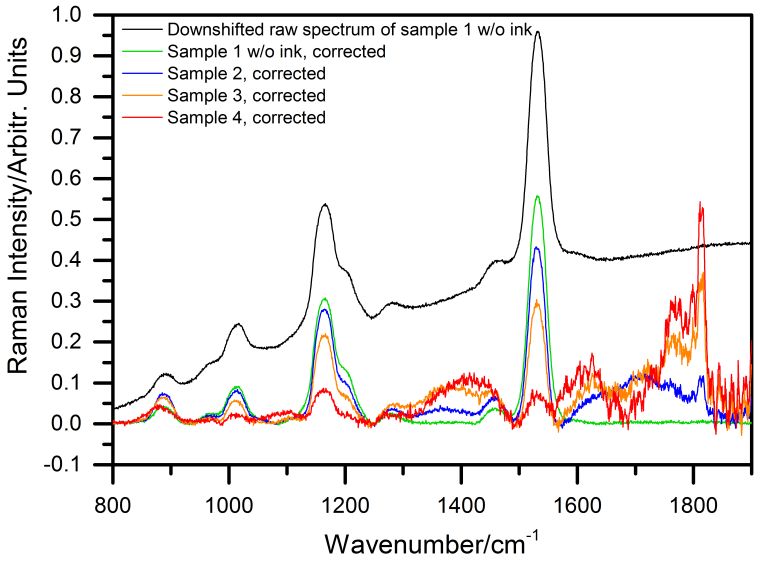


Figure 4: Raman spectra from Fig. 3 after applying the baseline correction algorithm with five iteration steps respectively. The raw spectrum (black) of sample (1), shifted down by a constant intensity offset, is included for comparison with the baseline corrected spectra. All spectra show the beta-carotene Raman lines. The signals at 1400 cm^{-1} and around 1800 cm^{-1} are unidentified spectral features of the ink used.

for signals of known molecular content.

4.2 Considerations regarding iteration count

As we employ an iterative algorithm, the most useful number of iterations needs to be determined. More iterations increase computation time, so a small number of steps is desirable. Furthermore, mollification - like any other smoothing step - has the effect of gradually smearing out data values over the whole spectrum, so no stationary solution usually can be found with this type of algorithm (for comparison,[15] employs a Savitzky-Golay filter and reports the same problem). Unlike polynomial baseline corrections, which can be numerically stable, this non-convergent behaviour implies that a very large number of iteration steps may introduce additional artifacts.

Fig. 2 already suggests that five iteration steps should be sufficient. To substantiate this statement, Fig. 5 shows the relative area changes between consecutive iteration steps for example spectra of samples (1) and (4). The area change decays more rapidly for sample (1) with low fluorescence than for sample (4) with strong fluorescence, but both spectra reach relative area changes below five per-

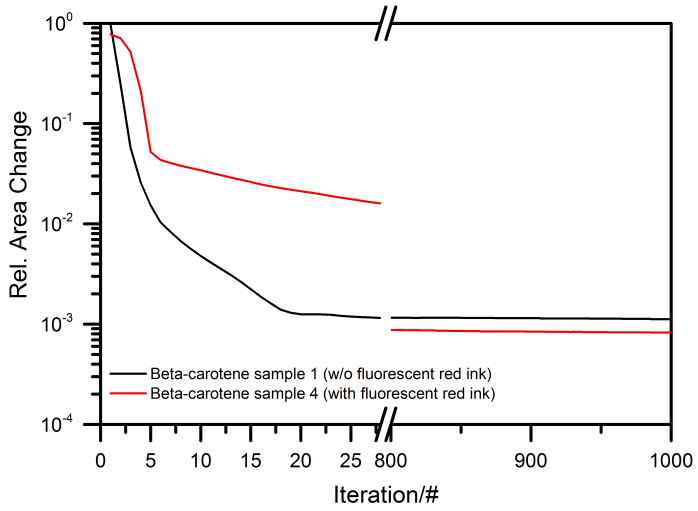


Figure 5: Relative area change between two successive iteration steps relative to initial spectral area plotted as function of the number of iteration steps for sample (1) with weak fluorescence and sample (4) with strong fluorescence.

cent after five iteration steps, which is therefore chosen as default step number, resulting in calculation times below one second for each spectrum.

For special cases, the best number of iteration steps can be determined experimentally by increasing iteration count step by step until a solution appears which does not change qualitatively with more iterations within the desired level of accuracy. However, even for that case, the total number of iterations should stay in the same order of magnitude compared to the pre-set value, to avoid artifacts.

4.3 Possible artifacts of the algorithm

If the feature width is chosen too narrow, parts of the peaks in the Raman spectrum under study will be considered as baseline and falsely removed. This can easily be avoided for spectra with separate peaks, but should be considered carefully if coalescent Raman lines appear in the experimental data. Similarily, the baseline correction will not give useful results when the feature width is chosen too large so that the fitting line cannot follow the baseline shape properly. The best feature width value that can deal with both narrowest baseline features and widest coalescent Raman lines depends on experimental conditions such as resolution, slit width and line broadening mechanisms and should be adapted accordingly. Usually, the feature width should be set to the footpoint to footpoint line width of

coalescent peaks or slightly more. According to our own experience, a fixed set of the parameters can be kept for measurements on very different samples with the same experimental setup.

Furthermore, a steep slope without signals at the border of the spectrum can be transformed into a large peak after correction if it is too narrow for the feature width chosen to approximate the background (as described above), as the algorithm cannot determine whether a peak at the border of the spectrum or a steep sloped part of the baseline is responsible for a rise of the curve at the end of the spectrum (not shown). Generally, the algorithm cannot treat curve slopes that are half the feature width or less away from the border.

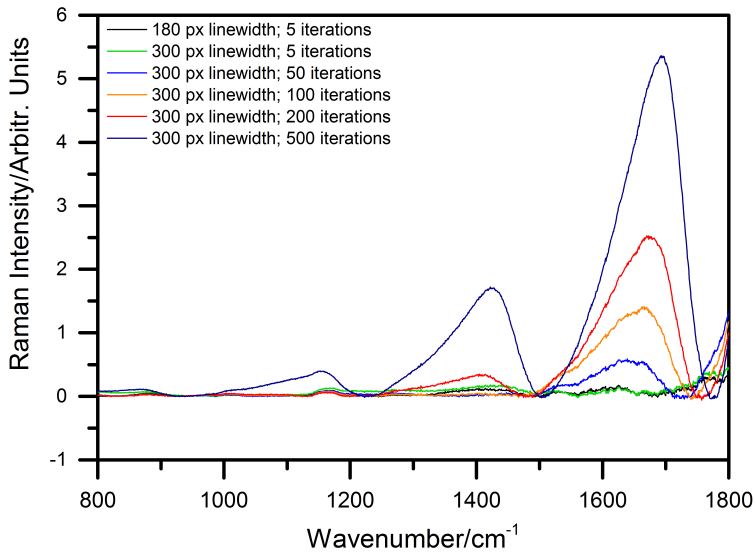


Figure 6: Artifacts appearing with increasing number of iteration steps on sample (4). See Fig. 4 for comparison with the case of properly chosen background subtraction parameters.

As the relative area change does not converge to zero with increasing iteration count, some sort of artifact is expected. Although strongly fluorescent baselines sometimes include unidentified spectral features as obvious around 1400 cm^{-1} and above 1600 cm^{-1} in the spectra shown in Fig. 4, these spectral features may also artificially arise when the number of iterations is very large. To illustrate this behaviour, which can further be enhanced by chosing a too large feature width, Fig. 6 compares cases where different iteration numbers were used. Most notably, the spectral features around 1100 cm^{-1}, 1400 cm^{-1} and 1700 cm^{-1} are artificially growing with increasing iteration step number. This usually happens when the chosen feature width is wide, but still slightly smaller than a spectral feature present in the spectrum. In this case, the fitting line will approximate the

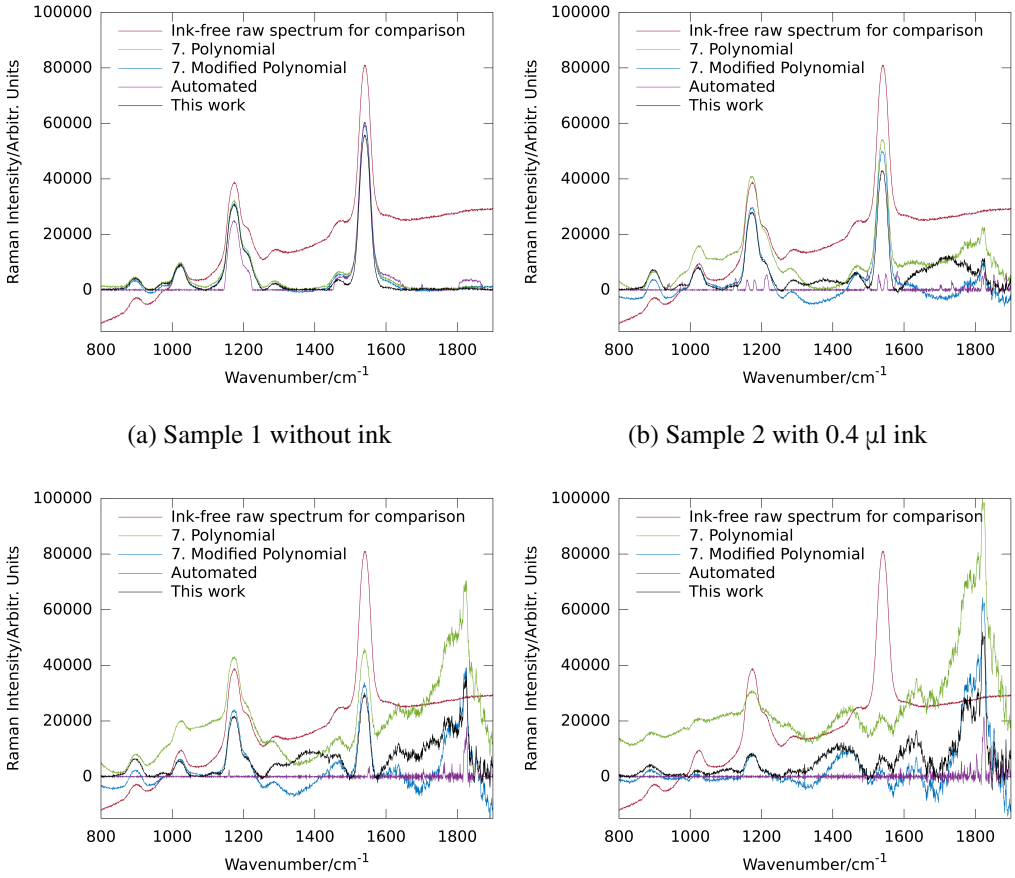
spectral feature in the center of the pre-baseline, which then will be smoothed in a way that it is smeared across the smoothed baseline which can then have a slightly higher intensity value than the original data on the slopes of the spectral feature. This causes the difference between the signal and the smoothed baseline to be negative after the iteration step, which in turn causes the spectral feature to rise when this difference curve is subtracted. Only a very small area is added in single iteration steps by this mechanism, but with many iteration steps, the summed artifact contribution can completely distort the original shape of the spectrum.

4.4 Comparison with other algorithms

To illustrate the performance of our algorithm, we compare the baseline correction results obtained with the ones from three other algorithms for which implementations are available. The iterative polynomial method as described in[11] and its modified variant, optimized for tissue fluorescence from[12] are available as part of the Biodata toolbox[19]. The implementation of the automated baseline correction algorithm[15] has been kindly made available by the authors. The individual algorithms have been used with the following settings: The polynomial based algorithms are employed by using a seventh order polynomial, the fully automated as given, and our algorithm with default settings of five iteration steps, a noise width of 6 pixels and a manually chosen feature width of 180 pixels. The polynomial corrections also have been tried by the authors with different polynomial orders from three to nine. Fit quality improved with order from three to seven; ringing occurred in order eight and got worse with a ninth order polynomial. Therefore, a seventh order polynomial has been chosen.

As can be seen in Fig. 7a, the different algorithms perform similiarly on the first sample without fluorescent ink, with the exception of the fully automated one that completely removes the peaks at 900 cm^{-1}, 1050 cm^{-1} and 1280 cm^{-1}. Additionally, the fully automated algorithm generates a plateau which is not present in the raw data at 1850 cm^{-1}.

For the second sample with 0.4 µl ink (see Fig. 7b), both polynomial corrections, altough preserving signals, start to show a slightly wiggling baseline because the irregular shape of the fluorescence cannot be fully approximated with a seventh order polynomial. An eighth and a ninth order polynomial were also employed, but they introduce ringing leading to even worse results. The fully automated algorithm still removes the genuine signals at 900 cm^{-1} and 1050 cm^{-1} and additionally transforms the two strong beta-carotene peaks into four small spikes respectively. Our algorithm gives a nearly flat baseline with all signals preserved. Around 1700 cm^{-1} all algorithms exhibit a bump shaped baseline distortion. Overall Raman line intensities are smaller than in the first sample, as already discussed.



(a) Sample 1 without ink

(b) Sample 2 with 0.4 µl ink

(c) Sample 3 with 1.6 µl ink

(d) Sample 4 with 8.0 µl ink

Figure 7: Comparison of the performance of different algorithms. The raw spectrum of sample 1 minus a constant offset, plotted in red, is included as reference for the beta-carotene spectral peak positions. Spectra corrected with the polynomial algorithm[11] are plotted in green, with the modified polynomial algorithm[12] in blue, with the fully automated algorithm[15] in purple and with our new algorithm in black.

For the third sample with 1.6 µl ink (see Fig. 7c), the fully automated algorithm shows no valid Raman peaks at all after correction. The polynomial baseline corrections still preserve the signals on an increasingly wiggling baseline, the modified polynomial correction giving a slightly better result. Our algorithm still achieves a nearly flat baseline and preserves all signals. Above 1700 cm^{-1}, the baseline distortions become stronger for all algorihms.

Finally, for the fourth sample, the addition of 8.0 µl ink leaves very weak

Ramans signals only (see Fig. 7d). The fully automated algorithm shows no carotenoid Raman peaks, but a sharp artifact peak from the measurement which is appearing at 1820 cm^{-1} is preserved correctly. Both polynomial algorithms give similiar shapes preserving the weak Raman signals, but the modified polynomial algorithm is performing slightly better. Our algorithm achieves a nearly flat baseline and preserves the weak Raman signals, altough a spectrum of this low signal to noise ratio would be discarded for most cases.

4.5 In-vivo Resonance Raman spectra examples

Carotenoids are often used as reference molecules because they exhibit very strong Raman signals on resonance excitation and can be found among others in human skin and in algae.

A real *in vivo* spectrum of a human finger (of one of the authors) was taken with the same experimental setup but with 10 μm slit width and is shown in Fig. 8. The result of the fully automated algorithm has been left out in this comparison, as the artifacts produced would reduce clarity of the image. All three algorithms considered produce similiar results, with the algorithm introduced in this work giving the flattest baseline.

For determination of the wavelength which is best suited to excite a sample under resonance conditions and to probe the excitation profile, excitation-emission maps in which each line consists of a Raman spectrum taken with another wavelength are required. These allow comparison of the Raman intensities taken with many different excitation wavelengths, for example, by using a tunable laser. Fig. 9 shows an *in vivo* excitation-emission map taken from a culture of green algae, Haematococcus pluvialis, SAG strain number 34-1a, with green motile cells. The raw data shown in Fig. 9a is then processed with our algorithm to generate the map in Fig. 9b. Although the individual raw spectra exhibit very different fluorescence backgrounds, the corrected map reveals consistent and detailed Raman spectra showing a smooth and authentic resonance behaviour in the signal intensity as expected for carotenoids. It should be noted that the slight wiggle in the line position is a mechanical artifact of our spectrograph, already present in the raw data, and not an artifact of the algorithm. Further examples of Raman excitation-emission maps of tissue phantoms containing carotenoids which are baseline-corrected with our algorithm can be found in[20], together with details on the experimental setup used.

As runtime performance might be of interest, the dataset for the excitation-emission map which consists of 45 different Raman spectra has been processed with the Matlab implementations of all four algorithms mentioned here within GNU Octave 4.0.0[21] on the same machine and with the settings given in the section before. Three runs have been averaged. The fastest has been the modified

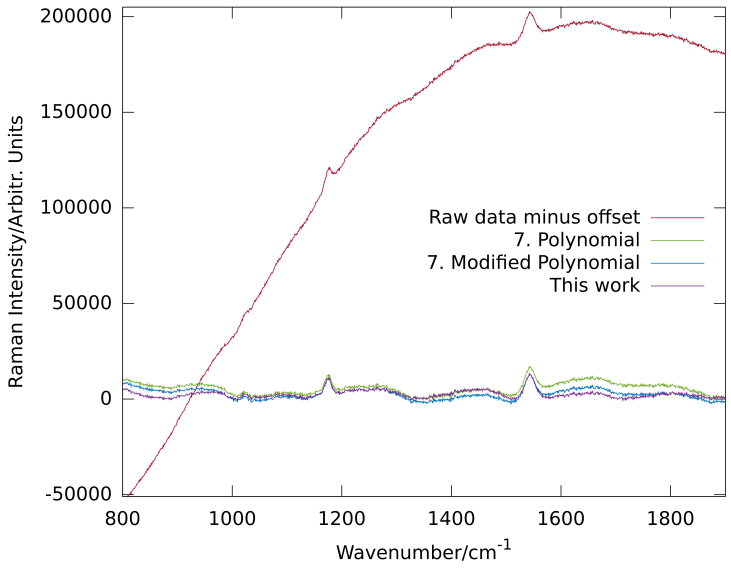
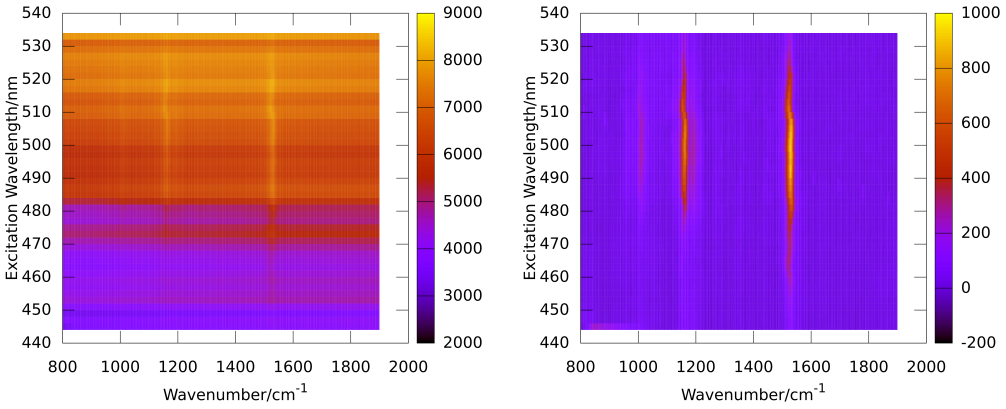


Figure 8: Comparison of the performance of different algorithms for a carotenoid spectrum taken from a human finger. The raw spectrum minus a constant offset, plotted in red, is included for comparison. Spectra corrected with the polynomial algorithm[11] are plotted in green, with the modified polynomial algorithm[12] in blue and with our new algorithm in purple.

polynomial algorithm with 0.8 s runtime, the polynomial algorithm needed 16 s, our new algorithm described here processed the data set in 27 s and the fully automatic algorithm took 322 s. In general, convolution is a computationally costly operation, which causes our new algorithm to be slower than traditional polynomial algorithms, but code optimisation with focus on performance is likely to shorten the runtime. The implementation which we made freely available is developed with source code readability in mind. Nevertheless, one strong point of our algorithm is that its runtime is independent of the actual data by design, as no data-dependent convergence criterions are employed.

5 Conclusion

In this work, we have presented a new fast and very efficient baseline correction algorithm for the particular needs of *in vivo* resonance Raman spectroscopy. The configuration parameters required are determined by the experimental setup and its largest coalescent Raman line widths only. The algorithm is well suited for batch processing of large sets of irregularly shaped *in vivo* Raman spectra



(a) Raw data.

(b) After baseline correction.

Figure 9: Resonance Raman measurement of Haematoccoccus pluvialis (SAG strain 34-1a).

with strong contribution from fluorescence taken under similar experimental conditions. Applications include sampling of carotenoids in human skin, long-time observations of bioreactors and comparison of multiple resonance Raman measurements taken with broadly tunable laser sources as reported in[20].

Our approach is well suited for baselines which cannot be handled by shape model based traditional algorithms and where fully automated, solely morphological algorithms like the one presented in[17] fail. One important aspect of our approach is that spectral features smaller than a chosen width are preserved under all circumstances to allow quantitative comparisons, which is important to evaluate coalescent Raman lines even with very low signal-to-noise ratio. This makes the approach well suited to search for unknown signals present in a sample. The approach also allows for calculation times of less that one second per spectrum and additionally, it is simple to implement and its function is geometrically descriptive and well understandable.

6 Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the VIP-project MeDiOO (grant no. 03V0826).

References

[1] A. P. Shreve, N. J. Cherepy, R. A. Mathies, *Appl. Spectrosc.* **1992**, *46*, 707.

[2] K. Sowoidnich, H.-D. Kronfeldt, *ISRN Spectrosc.* **2012**, *2012*.

[3] P. Matousek, M. Towrie, A. W. Parker, *J. Raman Spectrosc.* **2002**, *33*, 238.

[4] D. A. Long, *The Raman Effect - A Unified Treatment of the Theory of Raman Scattering by Molecules*, Wiley, Bradford, **2002**.

[5] I. V. Ermakov, M. R. Ermakova, R. W. McClane, W. Gellermann, *Opt. Lett.* **2001**, *26*, 1179.

[6] M. Darvin, N. Brandt, J. Lademann, *Opt. Spectrosc.* **2010**, *109*, 205.

[7] K. Golcuk, G. S. Mandair, A. F. Callender, N. Sahar, D. H. Kohn, M. D. Morris, *Biochim. Biophys. Acta* **2006**, *1758*, 868.

[8] P. Matousek, M. Towrie, C. Ma, W. M. Kwok, D. Phillips, W. T. Toner, A. W. Parker, *J. Raman Spectrosc.* **2001**, *32*, 983.

[9] J. M. Harris, R. W. Chrisman, F. E. Lytle, R. S. Tobias, *Anal. Chem.* **1976**, *48*, 1937.

[10] G. Schulze, A. Jirasek, M. M. L. Yu, A. Lim, R. F. B. Turner, M. W. Blades, *Appl. Spectrosc.* **2005**, *59*, 545.

[11] C. A. Lieber, A. Mahadevan-Jansen, *Appl. Spectrosc.* **2003**, *57*, 1363.

[12] J. Zhao, H. Lui, D. I. McLean, H. Zeng, *Appl. Spectrosc.* **2007**, *61*, 1225.

[13] A. Cao, A. K. Pandya, G. K. Serhatkulu, R. E. Weber, H. Dai, J. S. Thakur, V. M. Naik, R. Naik, G. W. Auner, R. Rabah, D. C. Freeman, *Journal of Raman Spectroscopy* **2007**, *38*, 1199.

[14] K. Chen, H. Wei, H. Zhang, T. Wu, Y. Li, *Anal. Methods* **2015**, *7*, 2770.

[15] H. G. Schulze, R. B. Foist, K. Okuda, A. Ivanov, R. F. B. Turner, *Applied Spectroscopy* **2011**, *65*, 75.

[16] Z.-M. Zhang, S. Chen, Y.-Z. Liang, Z.-X. Liu, Q.-M. Zhang, L.-X. Ding, F. Ye, H. Zhou, *Journal of Raman Spectroscopy* **2010**, *41*, 659.

[17] R. Perez-Pueyo, M. J. Soneira, S. Ruiz-Moreno, *Appl. Spectrosc.* **2010**, *64*, 595.

[18] S. Bukvic, D. Spasojevic, *Spectrochim. Acta B* **2005**, *B 60*, 1308.

[19] K. De Gussem, J. De Gelder, P. Vandenabeele, L. Moens, eng, *Chemometr. Intell. Lab.* **2009**, *95*, 49.

[20] M. Meinhardt-Wollweber, C. Suhr, A.-K. Kniggendorf, B. Roth, *Proc. SPIE* **2014**, *8945*, 89450B.

[21] S. H. John W. Eaton, David Bateman, R. Wehbring, *GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations*, **2015**.

7 Supplementary material

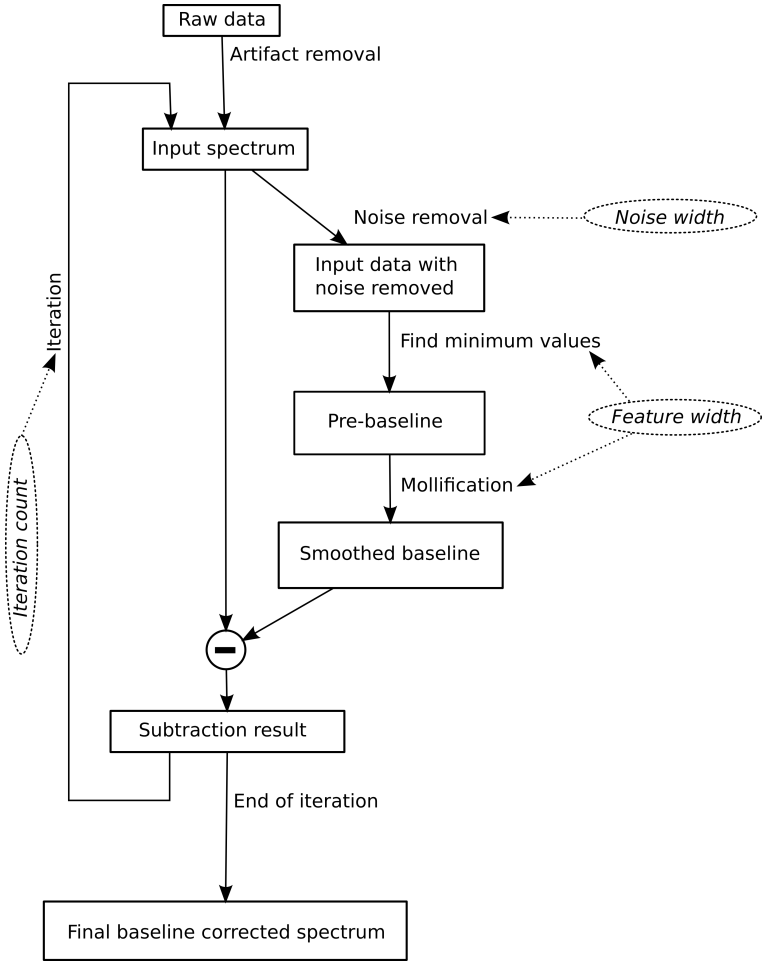


Figure S1: Flowchart illustrating the individual processing steps of our new baseline correction algorithm.