Analyst

PAPER

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2014, **139**, 5755

A new algorithm for identification of components in a mixture: application to Raman spectra of solid amino acids

Sylwester Gawinkowski, Agnieszka Kamińska, Tomasz Roliński* and Jacek Waluk

Received 27th June 2014
Accepted 1st September 2014

DOI: 10.1039/c4an01159g

www.rsc.org/analyst

The procedure for identifying components in a mixture was developed and tested on Raman spectra of mixtures of solid amino acids, using the spectra of single amino acids as templates. The method is based on finding the optimum scaling coefficients of the linear combination of template spectra that minimize the Canberra distance between measured and reconstructed spectra. The Canberra distance, used here as a measure of dissimilarity between spectra, defines the non-convex objective function in the related optimization process. In view of the possibility of the presence of local minima, *differential evolution*, which is a non-gradient stochastic method for finding the global minimum, was chosen for optimization. The method was tested on twenty measured spectra of mixtures of solid powders containing one to eight amino acids taken from the collection of twenty that are coded in living organisms. The results show that the procedure can successfully identify several amino acids, and, in general, several components in a mixture. The method was shown to compare favorably against the *least squares* and *partial least squares* methods, the procedures used in commercially available chemometrics packages.

1. Introduction

Many papers have been devoted to identification of substances by their spectra and the specific issues they address are diverse. For example, the problem may concern looking for a single substance by comparing its measured spectrum with successive entries in a library of spectra. Tanabe and Saëki[1] examined the possibility of identifying single substances by their IR spectra and the Pearson correlation coefficient. Several factors were investigated influencing the efficiency of the procedure, such as the wavenumber range and the spacing between adjacent data points, both related to the number of sampling points, as well as wavenumber accuracy and sample purity. Another problem concerns the case of assigning reference spectra in a library as components of a measured spectrum representing a mixture. Mallick *et al.*[2] compared several methods of calculating coefficients of components of a mixture spectrum assumed to be a linear combination of reference Raman spectra. They included all library reference spectra into the combination, which implied solving one problem of high dimension. Thus the efficiency of a numerical procedure was very important in this case. The methods were tested with simulated measurements obtained from a statistical model with the most important error sources. The work by Drake *et al.*[3] dealt with the case of linear dependence of some reference spectra in the library, which must take place in case the number of spectra exceeds the number of points in the spectrum, and found that *non-negative least squares with the active set* method described by Lawson and Hanson[4] was suitable for this task. Another possible scenario is looking for a particular substance in a mixture containing excipients or contaminants. O'Connell *et al.*[5] first pre-processed a large number of spectra by normalization to the strongest peak and calculating the first derivative. Both steps were justified by Principal Component Analysis (PCA).[6,7] Then they used several classification methods to discriminate the target analyte. These included Principal Component Regression (PCR),[6] Support Vector Machine,[8] *K*-Nearest Neighbors,[9] a decision tree,[10] and others.

Beyond correlation coefficient there are several other similarity or dissimilarity measures for matching spectra, such as the Euclidean distance, city-block distance,[11] Tanimoto coefficient (Jaccard index),[12] and cosine of an angle between the spectra. Li *et al.*[13] dealt with the general analysis of the correlation coefficient, Euclidean cosine and their first-difference counterparts applied to simulated spectra of one peak and ten peaks. The authors studied the influence of changing peak width and peak position on the similarity (dissimilarity) indices. They recommended that such indices should be used locally in predefined windows of significant intensities to increase the reliability of the results. Varmuza *et al.*[14] studied correspondence between spectral similarity, measured by the correlation coefficient, mean of the absolute and squared differences or Euclidean cosine, and structural similarity measured by the Tanimoto index. The authors performed random queries to a compound database, retrieving hit lists of

Institute of Physical Chemistry, Polish Academy of Sciences, Kasprzaka 44/52, 01-224 Warsaw, Poland. E-mail: rolinski@ichf.edu.pl; Fax: +48 22 343 3333

compounds with similar IR spectra and found the average for the Tanimoto coefficient between query and hit list compounds. The method was used to characterize the performance of a spectral similarity search.

In this work we propose a nonstandard dissimilarity measure between spectra, the so-called Canberra distance,[15] which is the sum of relative errors in intensities for successive wavenumbers. This index has been studied theoretically,[16] and it is presently applied in genomics as a measure of similarity.[17] It is assumed that the mixture is a linear combination of reference spectra. Then one tries to minimize the objective function, defined by the Canberra distance between the reconstructed and measured spectra, by varying the coefficients in the combination. Since the objective function is not a convex function of its coefficients, the uniqueness of the minimum is not guaranteed. Thus, some optimization procedure is required that is capable of finding the global minimum in the presence of local minima. Non-gradient stochastic optimization methods are suitable for this task. One method from this class, *differential evolution*, has recently gained popularity.[18,19] It is a representative of a wider class of genetic algorithms that finds the global solution with a high degree of probability, which proved to be the case in our present analysis. An example of the application of a genetic algorithm was presented by Forshed *et al.*[20] for the peak alignment procedure for NMR metabonomic data. The authors divided two spectra into common segments, and tried to shift sideways and stretch or shrink one of them by linear interpolation to fit the other one. The optimum values for this segment transformation were found by means of a genetic algorithm. The dissimilarity function (Canberra distance) of our paper can be viewed as a weighted city-block distance (sum of absolute differences), the weights being the inverses of the sums of absolute values of second derivatives of compared spectra for successive wavenumbers. The idea of using weights to cope with the problem when the range of values is wide was proposed by Liu *et al.*[21] The authors used the weighted Pearson product–moment correlation coefficient to compare high-performance liquid chromatograms and obtained better results than in the case of the non-weighted coefficient. It is methodologically appropriate to precede any identification process by the correlation analysis between template spectra. In our previous work[22] we performed comparison of the same set of spectra of twenty amino acids as in this paper using the intensities of the strongest peaks and their positions, as well as Pearson correlation coefficient as measures of similarity.

2. Methodology

The method was analyzed using as templates twenty measured Raman spectra of proteinogenic amino acids enumerated in Table 1. For a detailed description of amino acid samples, as well as the measurement conditions and results we refer the reader to the work by Roliński *et al.*[22] The samples were purchased from Aldrich and Merck and used without any additional purification. Raman spectra of single amino acids

Table 1 Collection of twenty amino acids taken into spectral analysis with their acronyms

	Amino acid	Acronym
1	Arginine	Arg
2	Proline	Pro
3	Alanine	Ala
4	Phenylalanine	Phe
5	Cysteine	Cys
6	Asparagine	Asn
7	Glutamine	Gln
8	Leucine	Leu
9	Threonine	Thr
10	Valine	Val
11	Isoleucine	Ile
12	Glutamic acid	Glu
13	Glycine	Gly
14	Aspartic acid	Asp
15	Lysine	Lys
16	Methionine	Met
17	Histidine	His
18	Serine	Ser
19	Tryptophan	Trp
20	Tyrosine	Tyr

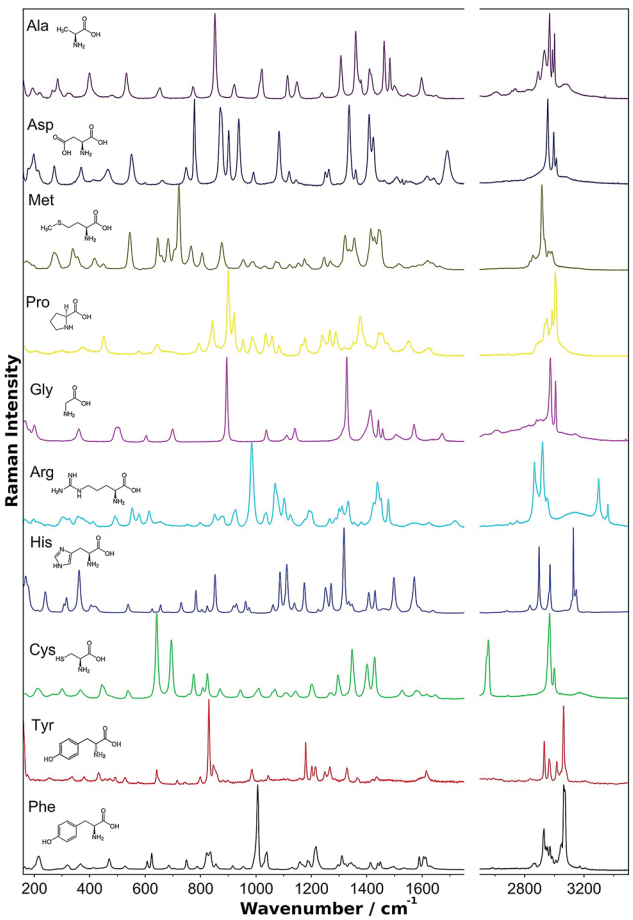

Fig. 1 Experimental spectra of solid amino acids measured at 293 K using the 632.8 nm laser line.

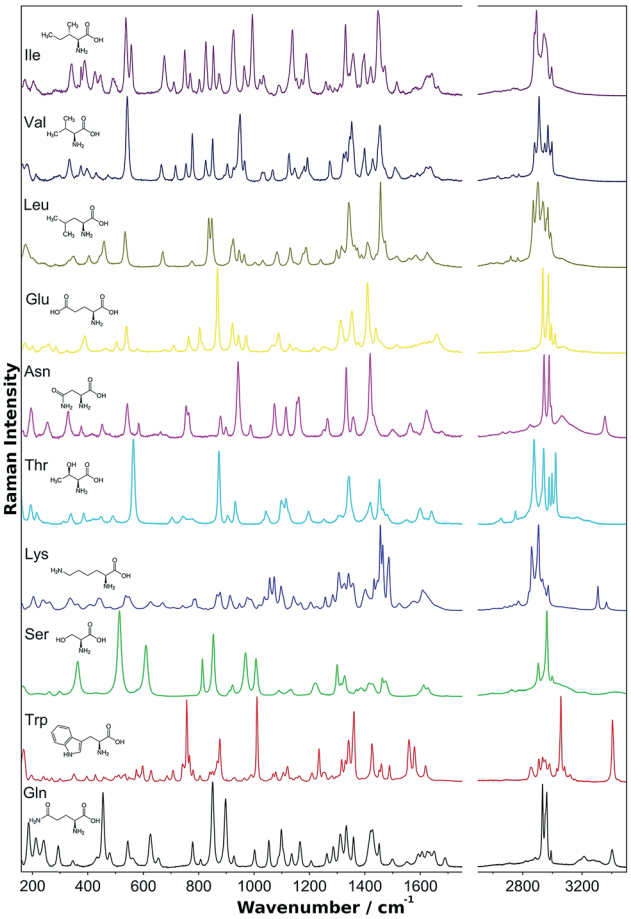

Fig. 2 Experimental spectra of solid amino acids measured at 293 K using the 632.8 nm laser line.

and their mixtures were recorded with a Renishaw InVia Raman microscope using the 632.8 nm line of the HeNe laser and a 20× objective. The laser power at the sample was 50 mW or less. The microscope was equipped with 1200 grooves per mm grating,

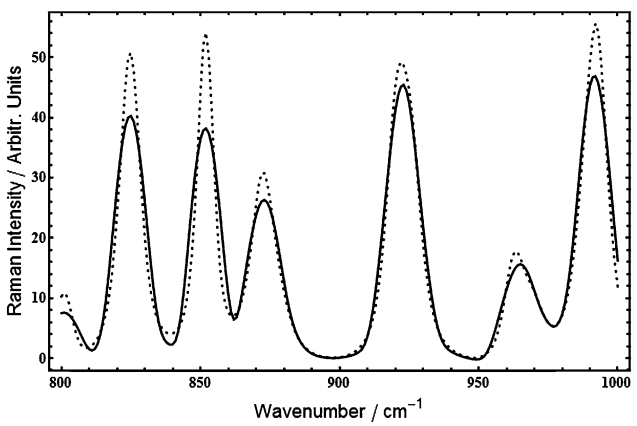

Fig. 3 The result of Savitzky–Golay filtering for the interpolated data (second order polynomial, 21 point window). Dashed line, interpolated data; solid line, result of filtering. For clarity, the data were restricted to the range 800–1000 cm^{-1}.

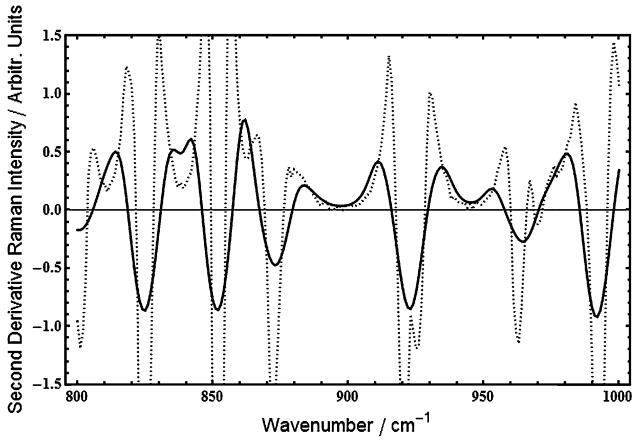

Fig. 4 Comparison of numerical calculation of the second derivative of interpolated data. Dashed line, second derivative of data obtained by finite differences for adjacent points; solid line, second derivative of data obtained by Savitzky–Golay filtering of the interpolated data (second order polynomial, 21 point window). For clarity, data were restricted to the range of 800–1000 cm^{-1}.

cutoff optical filters, and a 1024 × 256 pixel Peltier-cooled RenCam CCD detector, which allowed registering the Stokes part of Raman spectra with 5–6 cm^{-1} spectral resolution and 2 cm^{-1} wavenumber accuracy. To exclude the possibility of the orientational dependence of the signal on the polarization of the laser beam, the samples were finely pulverized and at least 100 spectra were recorded for each sample using an automatic translation stage and then averaged.

The measured spectra of solid amino acids are shown in Fig. 1 and 2. The high wavenumber range, 2500–3500 cm^{-1}, was not included in the identification process, providing a bigger challenge for the analytical algorithm since a region was neglected for which one observes substantial differences between the spectra of mixture components.

The spectra of mixtures and templates were first limited to the range 300–1700 cm^{-1} and then scaled so that the strongest peak within each spectrum equals to 100. This can be viewed as a simple preprocessing step.

Visual comparison of the spectra of mixtures with those of successive templates showed that the shifts of the corresponding peaks were very small, which greatly simplified the analysis since otherwise one would have to devise a measure of similarity that could compensate for this.[23] Since mixture intensities are sums of component intensities, one can expect good correspondence for intensities in some spectral regions only where a given component spectrum is dominant. If the fit was good for all considered energies (wavenumbers), this would mean that the mixture spectrum is trivial, *i.e.* containing one component. Moreover, even in the case of comparison of one component mixture and the corresponding template, the differences in intensities can be attributed to the measurement bias and the error in the definition of baselines for different spectra. It is also known that even under ideal measurement conditions the error in the value of intensity is theoretically proportional to the square root of the

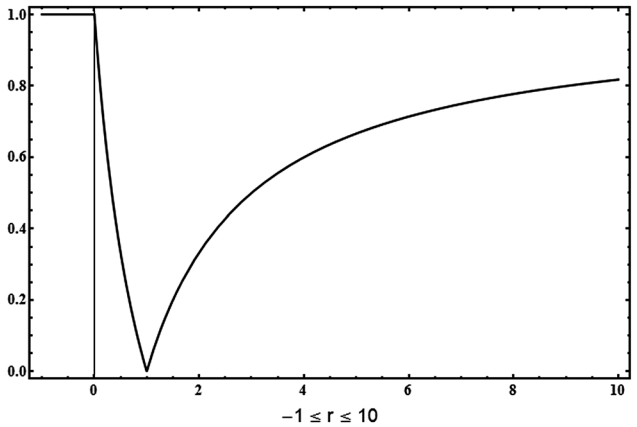

Fig. 5 A plot of the function $f(r)=\frac{|r-1|}{|r|+1}$ for $-1 \le r \le 10$ showing the behaviour of local distance for a particular energy (wavenumber), *i.e.* one term of the dissimilarity function (1). We assume that the second derivative of intensity of the measured mixture spectrum equals one and the second derivative of intensity of the reconstructed spectrum equals *r*.

value. With the above in view we defined a dissimilarity function that is more responsive to the differences in the position between peaks and their differences in widths than the differences in intensity:

Table 2 The first iteration of the method for mixture f (see Table 4). The results are sorted according to the values in the second column. First column, acronym of the amino acid (see Table 1); second column, values of the function (1) for optimum coefficients; third column, difference between the current value and the value of function (1) for the worst match expressed as percent of the value for the worst match; fourth column, the optimum scaling coefficient for the corresponding amino acid. In the present case the best match corresponds to serine

Amino acid	Values of function (1)	Difference in the values of (1) [%]	Optimum coefficient
Ser	865	−22.0	1.76
Thr	872	−21.0	1.14
Arg	910	−18.0	1.95
Gln	999	−10.0	0.76
Pro	1009	−9.1	1.77
Ala	1010	−9.1	0.99
Ile	1016	−8.5	0.58
Lys	1029	−7.3	1.23
Asn	1031	−7.2	0.73
Cys	1033	−7.0	2.27
Trp	1043	−6.0	0.93
Tyr	1046	−5.8	2.96
Asp	1049	−5.5	0.93
Met	1051	−5.4	1.78
Phe	1054	−5.1	2.30
Leu	1070	−3.7	0.67
Val	1070	−3.6	0.67
Glu	1077	−3.0	0.99
Gly	1097	−1.2	2.54
His	1110	0.0	1.84

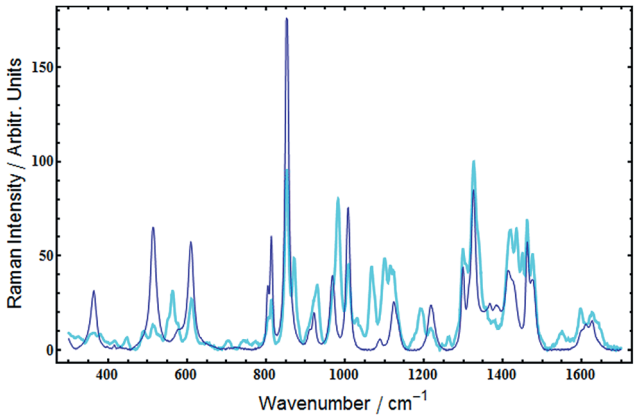

Fig. 6 Spectra of mixture f (cyan thick line) and serine (dark blue thin line) multiplied by the corresponding optimum coefficient (see Tables 1, 2 and 4).

$$\sum_{i=1}^{N} \frac{\left|R_L^{(2)}(w_i)-M_k^{(2)}(w_i)\right|}{\left|R_L^{(2)}(w_i)\right|+\left|M_k^{(2)}(w_i)\right|} \quad (1)$$

where: $R_L^{(2)}=\sum_{l\in L} c_l a_l^{(2)}$, $L\subset\{1,\ 2,\ 3,\ \ldots, M\}$ is the reconstructed second derivative of the mixture spectrum; $a_l^{(2)}$, $l = 1, \ldots, M$ is the second derivative of the template spectrum number l (see Table 1 and Fig. 1 and 2); c_l, $l = 1, \ldots, M$ is the l-th scaling coefficient; $M_k^{(2)}$, $k \in \{1, 2, 3, \ldots, P\}$ is the second derivative of the measured mixture spectrum (see Table 4); w_i, $i = 1, 2, 3, \ldots, N$ is the wavenumber corresponding to the i-th point in the spectrum.

In our analysis $M = 20$, $N = 1401$,

$$w_i = 299 + i\ \mathrm{cm}^{-1},\ i = 1, 2, 3, \ldots, 1401 \quad (2)$$

because we considered the spectral region of 300–1700 cm^{-1}, and the set of measured mixtures $(1, 2, 3, \ldots, P\}$ is replaced by $\{a, b, c, \ldots, t\}$ (see Table 4). The function (1) is the sum of local distances for the wavenumbers w_i. The local distance is the absolute value of the difference between $M_k^{(2)}$ and $R_L^{(2)}$ for a given wavenumber, divided by the sum of their absolute values. The definition of (1) can be viewed as the Canberra distance[15] between objects $R_L^{(2)}$ and $M_k^{(2)}$. In our case this function is minimized with respect to each c_l under restrictions $c_l \ge 0$, $l \in L \subset \{1, 2, 3, \ldots, 20\}$.

Since there is no analytical formula for the spectra, the second derivative has to be evaluated numerically. To this end we first obtained a vector of intensities for the wavenumbers w_i from (2) by interpolating the initial data, and then applied Savitzky–Golay filtering by trying to fit a second order polynomial locally to the data window of 21 points. The order of the polynomial and the width of the window were chosen by trial and error by comparing the original and the fitted data. An example of fitting is presented in Fig. 3 and 4.

One should note first that if there is a difference in the sign of the compared second derivatives of spectra for a given energy (wavenumber) then the corresponding term in (1) reaches its

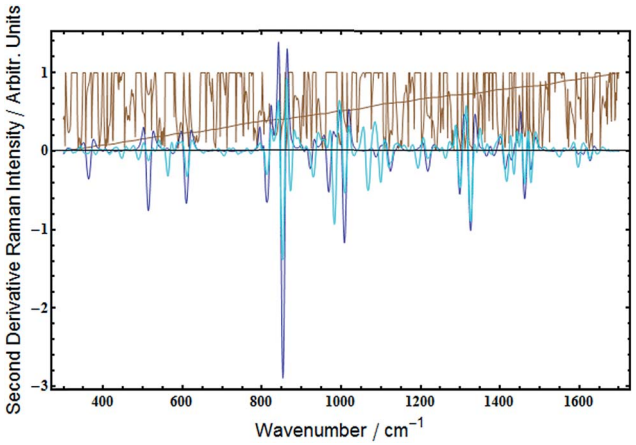

Fig. 7 Second derivatives of the spectra from Fig. 6. Local distances (components of (1)) are shown as a brown line. The line increasing monotonically is the normalized cumulative value of the local distances. Note that this line increases uniformly, which means that intervals of a large value of the second derivative do not dominate the value of the global distance, *i.e.*, the value of (1). We claim that this contributes to the quality of the method.

maximum value of one. On the other hand, if the signs are the same then this term approaches the maximum value only for a big difference between the second derivative spectra (see Fig. 5). Obviously, if the values of the compared second derivatives are the same the term equals zero. Second, if we scale both second derivative spectra by some constant then the value of (1) does not change. This means that those parts of second derivative spectra that are small in value have the same influence on the dissimilarity function as those which are large in value. Third, under assumption that the background varies slowly with respect to the curve representing the spectrum, the use of second derivatives has additional benefit, because it minimizes the error arising from the subtraction of the background, which is not defined precisely.

The expression (1) as a function of the coefficients c_l is non-convex, which means that it might have local minima, contrary to the case of the *least squares* problem, where one has one a global minimum. In practice we find a global minimum for the function (1) by

- choosing a suitable minimization method, *e.g. differential evolution* stochastic optimization method,[18,19]
- trying to avoid solving problems for large L, *i.e.* for many templates in the linear combination, by analyzing templates successively (one at a time), which will be explained in detail later on,
- monitoring the results of successive optimizations.

In what follows we shall be using acronyms for the corresponding amino acids taken into analysis. The correspondence is given in Table 1.

Now we describe the method in detail.

1st iteration

We try to find a single template matching the mixture spectrum the best, *i.e.* the one for which the optimum scaling coefficient gives the smallest value of function (1); see (1) for the case $L = \{m\}$, $1 \leq m \leq 20$ (see also Table 2 and Fig. 6 and 7).

Table 3 The second iteration of the method for mixture f (see Table 4). The results are sorted according to the second column. First column, acronym of the amino acid (see Table 1); second column, value of function (1) for optimum coefficients; third column, difference between the current value and the value of function (1) for the worst match as percent of the value for the worst match; fourth column, optimum scaling coefficient for the spectrum reconstructed in the first iteration (serine); fifth column, optimum scaling coefficient for the corresponding amino acid spectrum. In this case the best match is the combination of spectra of serine and threonine (see Tables 1 and 2)

Amino acid	Value of function (1)	Difference in the values of (1) [%]	Optimum coefficient for the reconstructed spectrum	Optimum coefficient for the corresponding amino acid
Thr	653	−25.0	0.76	0.63
Arg	774	−11.0	0.51	1.00
Ala	810	−6.3	0.84	0.36
Gln	811	−6.2	0.85	0.14
Asn	817	−5.5	0.69	0.27
Ile	822	−4.9	0.73	0.16
Asp	825	−4.6	0.70	0.27
Gly	831	−3.9	0.75	0.51
Cys	841	−2.8	0.92	0.14
Pro	848	−1.9	0.93	0.37
Glu	851	−1.6	0.76	0.33
Phe	852	−1.4	0.97	0.08
Trp	854	−1.3	0.97	0.10
Leu	857	−0.9	0.93	0.10
His	859	−0.7	1.00	0.12
Val	861	−0.5	0.87	0.04
Tyr	862	−0.4	0.83	0.65
Met	863	−0.2	1.06	0.00
Lys	865	0.0	1.00	0.00

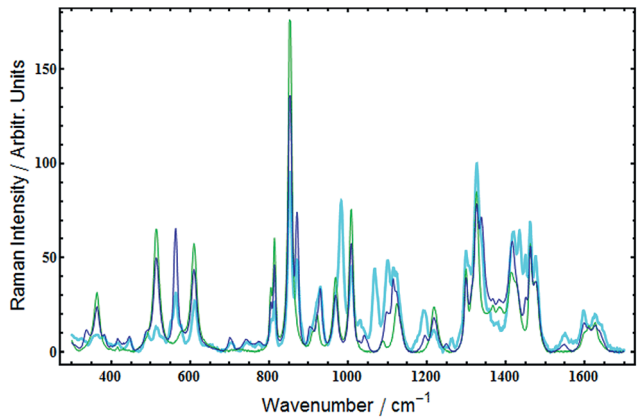

Fig. 8 Measured Raman spectrum of mixture f (cyan thick line), spectrum of serine (dark green thin line) with the optimum coefficient (see Table 2) and the linear combination of serine and threonine (dark blue thin line) with optimum coefficients (see Table 3). It can be seen that the combination fits the mixture spectrum better than the single template of serine if we give more weight to the distribution of peaks than their intensities according to the construction of function (1) (see the description below), see also Tables 1 and 4.

n-th iteration

Let us assume we have already accepted some number of amino acid spectra forming, with their coefficients, a linear combination. The combination defines the reconstructed spectrum at the $(n-1)$-th iteration. Now we want to check if the next template spectrum should be included in it. We try to accept a single template from the collection of templates that have not yet been included in the reconstructed spectrum. This template, together with the reconstructed spectrum, forms a linear combination for the optimization process. It should be underlined that now we try to find two scaling coefficients: one for the template and second for the whole reconstructed spectrum from the $(n-1)$-th iteration. The chosen combination should yield the smallest value for the dissimilarity function (1). If the stop condition is not fulfilled then we accept the best new template and perform additional optimization for all already accepted templates, thus defining the new reconstructed spectrum (see Table 3 and Fig. 8). The stop condition means that the absolute value of the difference in values of the function (1) for the best and the worst match divided by the value for the worst match is less than some threshold value or the value of the scaling coefficient corresponding to the best match falls below another threshold.

3. Results

The method was verified using twenty measured spectra of mixtures of solid powders containing from one to eight amino acids (see Table 4) taken from the set of twenty presented in Table 1. The mixtures contained approximately equal volumes of components, which do not mean that their contributions to the measured spectrum were equal, as different substances yield weaker or stronger Raman signals depending on their polarizability. We tested the power of the method in qualitative analysis of samples, *i.e.* in identifying the components. The results are presented in Table 5. There may be two kinds of errors in the analysis: identification of the substance not present in the mixture (false positive) and failure of detecting a substance present in the mixture (false negative). We assumed

Table 4 Sample identification letters and qualitative composition of measured mixtures

Mixture	Composition
a	Phe Ala
b	His Arg Pro
c	Tyr Asn
d	Gly Thr Ser Gln Leu
e	Trp Glu Gln Ile
f	Ser Thr Arg Ala
g	Met Ala His Gly Leu
h	Glu Leu Ile
i	His Gly Leu
j	Glu Met Lys
k	His Gly Tyr Pro
l	Thr Ala
m	Asn
n	Met Val Leu Ile
o	Phe Gln
p	Tyr Thr Pro Asn Asp
q	Gly Glu Ala
r	Tyr Trp His Arg Ser
s	Trp Asp
t	Tyr Cys Phe His Ser Leu Thr Ile

Table 5 Identification results of measured mixture spectra (denoted by letters; see Table 4). All amino acids shown in the second column were identified correctly using the adjusted mechanical stop condition of the iterative algorithm for all mixtures. The third column shows two false negative cases involving serine and isoleucine from the mixtures r and t of five and eight components, correspondingly. The last column shows zero false positive cases

Mixture	Identified correctly	False negatives	False positives
a	Phe Ala		
b	His Arg Pro		
c	Tyr Asn		
d	Gly Thr Ser Gln Leu		
e	Trp Glu Gln Ile		
f	Ser Thr Arg Ala		
g	Met Ala His Gly Leu		
h	Glu Leu Ile		
i	His Gly Leu		
j	Glu Met Lys		
k	His Gly Tyr Pro		
l	Thr Ala		
m	Asn		
n	Met Val Leu Ile		
o	Phe Gln		
p	Tyr Thr Pro Asn Asp		
q	Gly Glu Ala		
r	Tyr Trp His Arg	Ser	
s	Trp Asp		
t	Tyr Cys Phe His Ser Leu Thr	Ile	

that both errors are equally serious and, accordingly, tried to minimize their sum by adjusting the stop condition (see the description of the n-th iteration of the method), which in the case of the analyzed mixtures (see Table 4) means that the difference in values of function (1) for the best and the worst match divided by the value for the worst match must be less than 2.4% or the value of the scaling coefficient of a potential component corresponding to the best match must fall below 0.04. The number of misclassifications for the optimum stop condition can serve as a measure of the quality of the identification algorithm. There are no false positive and only two false negative cases: they concern the mixtures with a high number of components (five and eight).

4. Comparison with the *non-negative least squares* (NLS) method

The objective function (1) is nonstandard for component identification in mixtures. The standard one is the Euclidean distance corresponding to the *least squares* (LS) method, and consequently LS can serve as a benchmark. The LS method has been used extensively in linear mixture analysis (see the work by Heinz and Chang[24] and references therein). Two options are possible, namely we can use zero or second derivative spectra. We performed calculations for the second derivative spectra and obtained vastly different numbers of misclassifications for spectra normalized as described in Section 2 (normalization to the strongest peak both for templates and mixture spectra) and for spectra normalized as described in Section 5. In contrast, for zero derivative spectra the numbers of misclassifications were quite close for both normalizations and this is the reason why we present here the results for zero derivative spectra with normalization of Section 2. Now the following function is minimized:

$$\sum_{i=1}^{N} |R_L(w_i) - M_k(w_i)|^2 \qquad (3)$$

where R_L and M_k are defined analogously to $R_L^{(2)}$ and $M_k^{(2)}$ in (1), except that we now consider the zero derivative of spectra, and consequently the superscript $^{(2)}$ is omitted in the definition; L, k, and w_i are defined in (1) and (2).

This function is the sum of squared differences of intensities in the mixture spectrum and the linear combination of template spectra. The function is minimized with respect to all the c_l coefficients under restrictions $c_l \geq 0$, $l \in L \subset \{1, 2, 3, \ldots, 20\}$.

We repeated the algorithm of Section 2 by calculating optimum coefficients of linear combinations for the objective function (3). In this case they can be found more efficiently by the Lawson and Hanson algorithm,[4] but it was sufficient to use the *FindMinimum* procedure from the *Mathematica* package,[25] as time here was not a parameter for optimization. The results are presented in Table 6. The stop condition in this case means that the difference in values of function (2) for the best and the worst match divided by the value for the worst match must be less than 11% or the value of the scaling coefficient of a potential component corresponding to the best match must fall below 0.04. As for the Canberra distance case, the parameters in the stop condition were adjusted to obtain the least number of misclassifications. This number equals eleven, and it is much higher than in the case of function (1) (see Table 6).

Table 6 Identification results of measured mixture spectra (see Table 4) in the case of benchmark objective function (3). All amino acids shown in the second column were identified correctly using the mechanical stop condition from the iterative algorithm (see Section 4); the third column shows false negative cases and the fourth column shows false positive cases

Mixture	Identified correctly	False negatives	False positives
a	Phe Ala		
b	His Arg	Pro	
c	Tyr Asn		
d	Gly Thr Ser Gln Leu		
e	Trp Glu Gln	Ile	
f	Ser Thr Arg	Ala	
g	Met Ala His Gly Leu		
h	Glu Leu Ile		
i	His Gly Leu		
j	Glu Met Lys		
k	His Gly Tyr Pro		
l	Thr Ala		
m	Asn		
n	Met Val Leu Ile		
o	Phe	Gln	Trp
p	Tyr Thr Pro Asn	Asp	
q	Gly Glu	Ala	Lys
r	Tyr Trp His Arg	Ser	
s	Trp Asp		
t	Tyr Cys Phe His Ser Ile	Leu Thr	

5. Comparison with the *partial least squares* (PLS) method

Another standard procedure used widely in chemometrics for calibration is the *partial least squares* (PLS) method. It was introduced for the first time several decades ago in econometrics and then it gained popularity in chemistry for modelling the relationship between some explanatory (easily obtainable) variables and the difficult or expensive to obtain response variables.[6,26] This is a method for solving the *least squares* (LS) problem approximately. The matrix of intensities from LS is replaced by a matrix of a much simpler structure, and usually of a lesser rank, that can be represented as a sum of some number of outer products of vectors of scores and loadings. The number is equal to the rank of the matrix and is referred to as a number of factors. The replacement is particularly useful if the columns of the intensity matrix are strongly correlated, which means that the explanatory (independent) variables are correlated, as well as in the case of noisy data.[7,26] The colinearity of variables is unavoidable if the number of explanatory variables is greater than the number of observations, which is usually the case if we seek the signal contributions of substances using spectral intensities. It must be added that for the PLS method the calculated approximate matrix is dependent on the values of the

response variable. In a simpler technique called principal component regression (PCR) the approximate matrix is defined independent of the response variable. This technique is related to the singular value decomposition procedure from the linear algebra and relies on choosing only the singular vectors from this decomposition related to the largest singular values. It was shown that the PLS technique leads often to a faster reduction of the residuals than PCR.[7] The technique has been implemented in many software chemometrics packages, *e.g. Grams*,[27] *Unscrambler®X*,[28] and this section can be treated as a comparison of the identification capabilities of the commercially available software with the identification power of the Canberra distance (1). Generally, there are training and testing steps in the analysis. First, we treated the set of templates as the training set and found the scores and loadings to model signal contributions. We applied the PLS1 version of the algorithm, which means that we independently calibrated the signal contribution of each amino acid. For a given amino acid the contribution related to its corresponding spectrum was set equal to one, and for the rest of the templates the contributions were set to zero. Second, we tested the model on twenty mixtures (see Table 4), predicting the signal contributions with the calculated scores and loadings. The sequential character of the algorithm was maintained by successive spectral subtractions of the identified templates multiplied by the found contribution coefficients (scaling coefficients) from the analyzed mixture spectrum. So first we subtracted the template with the highest calculated contribution in the mixture, and then repeated the procedure to find the next highest contribution for the difference spectrum and the rest of potential component spectra. The procedure stopped when the calculated contribution dropped below a preset threshold level. It should be underlined here that for the procedure from Section 2 we did not perform the subtraction operation, but tried to find the best fit for the combination of the reconstructed spectrum and a potential component spectrum.

Since we decided to model the real signal contribution ratios of constituents in the mixture, the measurement and scaling of templates and mixtures were different than in Sections 2, 3 and 4. The spectra of templates (components) were recorded for the same time so that their intensity ratios would reflect the contribution ratios in a mixture. Then we scaled both templates and mixtures so that the average of the strongest peaks in all templates equaled 100 and for each mixture spectrum the integral over the whole wavenumber range equaled the average of the integrals for templates. The above can be viewed as a simple preprocessing step.

We performed some number of simulations by varying the number of factors in the PLS method and analyzing centered (after the subtraction of the intensity average) or non-centered (original, positive) intensity vectors (spectra). The PLS procedure is essentially quantitative and its quality can be assessed by the *predictive residual sum of squares* (PRESS).[6] Here we use PLS for identification and therefore we must define the contribution threshold for confirmation of the presence of a component in the mixture. If we find the optimum value for this threshold corresponding to the smallest number of misclassifications, *i.e.* the sum of false positives and false negatives for a given test set of mixtures, then this number can serve as a measure of quality of identification corresponding to PRESS in the basic quantitative case.

Table 7 Identification results of measured mixture spectra denoted by letters (see Table 4) in the case of the PLS method. All amino acids shown in the second column were identified correctly using the optimum mechanical stop condition from the iterative algorithm (see Section 5); the third column shows false negative cases and the fourth column shows false positive cases

Mixture	Identified correctly	False negatives	False positives
a	Phe Ala		
b	His Arg Pro		
c	Tyr Asn		
d	Gly Thr Ser Gln Leu		
e	Trp Glu Gln Ile		
f	Ser Thr Arg	Ala	
g	Met Ala His Gly Leu		
h	Glu Leu Ile		
i	His Gly Leu		
j	Glu Met Lys		
k	His Gly Tyr Pro		
l	Thr Ala		
m	Asn		Gly
n	Met Val Leu Ile		
o	Phe Gln		Ser Trp
p	Tyr Thr Pro Asn Asp		
q	Gly Glu Ala		His Lys
r	Tyr Trp His Arg	Ser	
s	Trp Asp		Ile
t	Tyr Cys His Ser Leu Ile	Phe Thr	

The best results were obtained for the case of four factors and non-centered data together with the optimum contribution threshold of 0.04, which yielded ten misclassifications: four false negatives and six false positives (see Table 7). Since the mixtures were prepared by mixing approximately equal volumes of powders the threshold of 0.04 seems to be rather small, which means that if the volumes of components in mixtures were very small this method of identification would probably fail. The number of four factors is relatively small if we compare it with the number of explanatory variables (intensities for 1401 wavenumbers), which means that many variables contain similar information on the signal contributions. We also obtained twelve misclassifications for two factors only in the PLS method. Interestingly, the number of misclassifications increased to sixteen for nine factors, which probably means that more detailed data in spectra were treated as noise. Comparing the results presented in Table 7 with those in Table 6 leads to the conclusion that PLS does not have more identification power than the NLS method.

6. Conclusions

A method for identifying components in a mixture was developed and tested on powder mixtures of amino acids. The procedure is based on the linear model of the mixture and involves searching for scaling coefficients of the linear

combination of template spectra minimizing a function of dissimilarity referred to in the literature as Canberra distance.[15,17] The Canberra distance is related to a non-convex objective function in the optimization process and consequently the method requires a stochastic optimization algorithm in view of the possibility of existence of local minima. The method was tested using twenty measured spectra of mixtures. Each mixture contained approximately equal volumes of powders of several amino acids taken from the collection of twenty. The number of amino acids varied between one and eight. The method does not attempt to find coefficients of the combination of all twenty template spectra simultaneously, but accepts them into the reconstructed spectrum of the mixture successively, starting from those most similar to the measured spectrum of the mixture. Most components were identified correctly: there were only two false negative cases for mixtures of five and eight components and zero false positives (see Table 5). These results were achieved for the optimum values of two threshold parameters (see Section 3) defined for all considered mixtures. The results compare favorably with those obtained using the *non-negative least squares* (NLS) method, which, for the two optimum parameters, gave eleven misclassifications (see Table 6), and those provided by the *partial least squares* (PLS) method, which, for one optimum parameter, gave ten misclassifications (see Table 7).

It should be mentioned that PLS is much faster than the remaining two methods, especially the method based on the Canberra distance, which, however, is superior in identification power. The speed of the PLS method is due to the two-step process mentioned in Section 5. The model parameters calculated in the training step serve for prediction of amino acid signal contributions in all analyzed mixtures, which includes few vector multiplications only, without solving any equations. In contrast, for the two previous methods we performed a series of optimizations for each mixture to find scaling coefficients, though in the NLS method these optimizations were relatively fast, because they involved minimizing simple quadratic functions. Moreover, one should be aware that only one contribution threshold parameter is required for PLS, whereas for the other two methods there are two of them, which may somewhat bias the results.

The total time for the analysis of all 20 mixtures and 20 templates for the Canberra distance method amounted to 210 minutes (approximately 5–14 minutes for two component mixtures up to 17–26 minutes for more than five components in a mixture), whereas for the PLS method it took only 3 seconds if we do not count the training step. Of course the present method can be made faster if we consider less data points (wavenumbers) in spectra. Moreover, the optimization method (*differential evolution*) is time-consuming, as it requires many calls to the objective function. Therefore, substituting it with a simple gradient method could greatly accelerate the identification process, but at the expense of the possibility of falling into a local (not global) minimum of the objective function and, consequently, increasing the number of misclassifications. Both ways of acceleration, however, were not verified in practice. We think that by the already obtained results we could combine the two methods together. First, we could use the PLS method with parameters defined so as to make the false negative cases (almost) absent, and then use the present method to additionally verify the already identified components treated as a limited set of templates.

In practice we often face the situation where there are some unknown components in the mixture spectrum, *i.e.* the components that cannot be spanned by the templates, and the applicability of the method in such cases is important. Obviously, the stronger the unknown components with respect to templates the less the identification power of the method. The method works sequentially, *i.e.* the first identified components are more dominant in the Raman signal, at least with respect to the chosen Canberra metric. Since the method performed well in the case of 5–8 components in the mixture, this suggests that if the sought components are reasonably strong they should be identified. However there is a problem of the threshold condition in this case, *i.e.* how close the template spectrum should be to the mixture spectrum to be considered as its part. In general, the problem can hardly be solved. In practice, one is interested in 1–3 components. They can be chosen from the set of templates by the devised algorithm, and then it can be checked visually how many peaks in the compared spectra coincide, taking also into account how strong they are in the templates.

Acknowledgements

The research was supported by the European Union within the European Regional Development Fund, through grant Innovative Economy (POIG.01.01.02-00-008/08).

References

1 K. Tanabe and S. Saëki, *Anal. Chem.*, 1975, **47**, 118.
2 M. Mallick, B. Drake, H. Park, A. D. Register, P. West, R. Palkki, A. Lanterman and D. Emge, in *12th International Conference on Information Fusion*, Seattle, WA, USA, 6–9 July, 2009.
3 B. Drake, J. Kim, M. Mallick and H. Park, in *Proceedings of the Thirteenth International Conference on Information Fusion*, Edinburgh, UK, 2010.
4 C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, 1974.
5 M. L. O'Connell, T. Howley, A. G. Ryder, M. N. Leger and M. G. Madden, in *Proceeding of: Opto-Ireland 2005: Optical Sensing and Spectroscopy*, Ireland, 2005, p. 340.
6 P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, 1986, **185**, 1.
7 L. Elden, *Comput. Stat. Data Anal.*, 2004, **46**, 11.
8 N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.
9 B. S. Everitt, S. Landau, M. Leese and D. Stahl, Miscellaneous Clustering Methods, in *Cluster Analysis*, 5th edn, John Wiley & Sons, Ltd., Chichester, UK, 2011.
10 L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1984.

11 M. Barile, *Taxicab Metric*, From MathWorld–A Wolfram Web Resource, created by Eric W. Weisstein, http://mathworld.wolfram.com/TaxicabMetric.html.
12 A. H. Lipkus, *J. Math. Chem.*, 1999, **26**, 263.
13 J. Li, D. B. Hibbert, S. Fuller and G. Vaughn, *Chemom. Intell. Lab. Syst.*, 2006, **82**(1–2), 50.
14 K. Varmuza, M. Karlovits and W. Demuth, *Anal. Chim. Acta*, 2003, **490**, 313.
15 G. N. Lance and W. T. Williams, *Comput. J.*, 1966, **9**, 60.
16 G. Jurman, S. Riccadonna, R. Visintainer and C. Furlanello, in *Advances in Ranking, NIPS 09 Workshop*, ed. S. Agrawal, C. Burges and K. Crammer, 2009, p. 22.
17 J. Wu, M. Gan, W. Zhang and R. Jiang, *Int. J. Biosci., Biochem. Bioinf.*, 2011, **1**, 102.
18 R. Storn and K. J. Price, *Global Optim.*, 1997, **11**, 341.
19 V. Plagianakos and E. W. Weisstein, *Differential Evolution*, From MathWorld—A Wolfram Web Resource, http://mathworld.wolfram.com/differentialevolution.html.
20 J. Forshed, I. Schuppe-Koistinen and S. P. Jacobsson, *Anal. Chim. Acta*, 2003, **487**(2), 189.
21 Y. Liu, Q. Meng, R. Chen, J. Wang, S. Jiang and Y. Hu, *J. Chromatogr. Sci.*, 2004, **42**(10), 545.
22 T. Roliński, S. Gawinkowski, A. Kamińska and J. Waluk, in *Optical Spectroscopy and Computational Methods in Biology and Medicine*, ed. M. Baranska, Springer, 2014, vol. 14, p. 329.
23 R. de Gelder, R. Wehrens and J. A. Hageman, *J. Comput. Chem.*, 2001, **22**(3), 273.
24 D. C. Heinz and C.-I. Chang, *IEEE Trans. Geosci. Rem. Sens.*, 2001, **39**(3), 529.
25 Wolfram Mathematica 8, http://www.wolfram.com/.
26 S. Wold, M. Sjostrom and L. Eriksson, *Chemometr. Intell. Lab. Syst.*, 2001, **58**, 109.
27 GRAMS Suite 9.1, http://gramssuite.com/.
28 The Unscrambler®X 10.3, http://www.camo.com/.