Research on single live cells using a Raman microscope (hereafter denoted as bio-Raman research) was first reported in 1990.1 Biomolecules such as proteins, nucleic acids, and lipids, that is, atoms combined with electrons, vibrate with different frequencies. These oscillators interact with light and are detected as Raman scattering with different wavenumbers. Using Raman signals, we are capable of analyzing the molecular structure and distribution of biomolecules within a cell. Based on noteworthy studies,2–5 today, bio-Raman research is applicable to the fields of cellular biology and biophysics, and thus attracting many researchers.6–13

Multivariable analysis, such as principal component analysis (PCA) or a PCA-based approach, called chemometrics in chemistry, is frequently applied to bio-Raman research for visualizing and extracting intrinsic information.12–21 This is because bio-Raman data are generally complicated and large, making it almost impossible to interpret the data in an intuitive manner. For example, the number of spectra can easily reach a thousand when the time lapse of cellular differentiation is monitored, which typically takes several days.

Before proceeding to multivariable analysis, it is necessary to subtract background components from observed spectra; however, the background removal of bio-Raman data is usually manual and time consuming, preventing effective bio-Raman data analysis. Figure 1 exemplifies 15 Raman spectra of single HeLa cells on a quartz substrate, as well as a background spectrum of quartz where cells are not present. Owing to the uncertainty of the height of the confocal point and the difference in the refractive index between a cell and water medium, the Raman intensities of the quartz with and without a cell are not perfectly identical. Thus, after the proper correction of baselines such as autofluorescence, the quartz spectrum is subtracted from the measured Raman spectra, one by one, so that the generated Raman spectra have no Raman signal from the quartz.

The purpose of this study is to propose and develop an automated method for removing background information from bio-Raman data. The proposal of this method enables more effective and stress-free bio-Raman data analysis.

Research on single live cells using a Raman microscope (hereafter denoted as bio-Raman research) was first reported in 1990.1 Biomolecules such as proteins, nucleic acids, and lipids, that is, atoms combined with electrons, vibrate with different frequencies. These oscillators interact with light and are detected as Raman scattering with different wavenumbers. Using Raman signals, we are capable of analyzing the molecular structure and distribution of biomolecules within a cell. Based on noteworthy studies,2–5 today, bio-Raman research is applicable to the fields of cellular biology and biophysics, and thus attracting many researchers.6–13

Multivariable analysis, such as principal component analysis (PCA) or a PCA-based approach, called chemometrics in chemistry, is frequently applied to bio-Raman research for visualizing and extracting intrinsic information.12–21 This is because bio-Raman data are generally complicated and large, making it almost impossible to interpret the data in an intuitive manner. For example, the number of spectra can easily reach a thousand when the time lapse of cellular differentiation is monitored, which typically takes several days.

Before proceeding to multivariable analysis, it is necessary to subtract background components from observed spectra; however, the background removal of bio-Raman data is usually manual and time consuming, preventing effective bio-Raman data analysis. Figure 1 exemplifies 15 Raman spectra of single HeLa cells on a quartz substrate, as well as a background spectrum of quartz where cells are not present. Owing to the uncertainty of the height of the confocal point and the difference in the refractive index between a cell and water medium, the Raman intensities of the quartz with and without a cell are not perfectly identical. Thus, after the proper correction of baselines such as autofluorescence, the quartz spectrum is subtracted from the measured Raman spectra, one by one, so that the generated Raman spectra have no Raman signal from the quartz.

The purpose of this study is to propose and develop an automated method for removing background information from bio-Raman data. The proposal of this method enables more effective and stress-free bio-Raman data analysis.

HeLa cells were incubated using Dulbecco’s modified Eagle’s medium (abbreviated as DMEM, Sigma D5796), adding inactivated fetal bovine serum (10% of DMEM in volume, Gibco 10437-028) and penicillin-streptomycin solution (1% of DMEM in volume, Gibco 15070-063). A measurement of the Raman spectrum of a single HeLa cell was initiated after replacing of DMEM with Hank’s buffered salt solution (Sigma H8264) and incubation for about 30 min. A Raman microscope system (Renishaw, inVia Basic) equipped with a Leica upright microscope with a diode-pumped continuous-wave laser (532 nm, 180 mW, Excelsior) and a water immersion objective lens with a long focal length (×63 and numerical aperture: 0.90). The power of the excitation light was ca. 2 mW at the sample. In the Raman measurement, the center of the confocal volume was set to 2 μm above the quartz plane. The exposure time was set at 250 s.
Our proposal is expressed by simple relations. Based on an experimentally obtained spectrum, \( p(w_j) \), and an experimentally measured background spectrum, \( q(w_j) \), both as functions of Raman shift, \( w_j \), where \( j = 1, 2, 3 \ldots n \), a spectrum with no background signal, \( r(w_j) \), is given as

\[
r(w_j) = p(w_j) - kq(w_j).
\]  

(1)

Here \( k \) is a scaling factor, such that \( r(w_j) \) becomes independent of \( q(w_j) \). Figure 2 gives a schematic interpretation of \( r(w_j) \) varied with the scaling factor \( k \) around the optimized \( r(w_j) \). We assume here that \( \Delta w \equiv w_{j+1} - w_j \) is constant (evenly spaced) for simplicity. If \( \Delta w \) is unevenly spaced, \( \Delta w \) can be transformed to be evenly spaced using an interpolation technique. The important point is that the factor \( k \) can be approximately determined, without manually tuning \( k \) to obtain a suitable spectrum. The relation is given by

\[
k = \frac{\sum_{j=1}^{n} p'(w_j)q'(w_j)}{\sum_{j=1}^{n} q'(w_j)^2}
\]  

(2)

where \( p'(w_j) \) and \( q'(w_j) \) are the first derivative spectra of \( p(w_j) \) and \( q(w_j) \), respectively.

Our proposal is based on finding the “least length” of a spectrum, \( r(w_j) \), by changing the scaling factor, \( k \). Using this concept, we can obtain the simple relation of the parameter \( k \) of Eq. (2). The summation of the squared lengths of each segment, \( s_k \), is given by \( s_k = \sum_{j=1}^{n} (\Delta w^2 + \Delta w^2 r'(w_j)^2) \), where \( r'(w_j) \) is the first derivative spectrum of \( r(w_j) \). Since \( \Delta w \) is set to be evenly spaced, \( s_k \) can be simplified as

\[
s = \sum_{j=1}^{n} s'(w_j)^2,
\]  

(3)

where \( s = (s_k - n\Delta w^3)/\Delta w^2 \). \( s \) is thought to be another expression for the length of the spectrum \( r(w_j) \). Substituting Eq. (1) into Eq. (3), we obtain

\[
s = \sum_{j=1}^{n} [p'(w_j)^2 - 2kp'(w_j)q'(w_j) + kq'(w_j)^2].
\]  

(4)

Equation (4) is the equation of \( s \) with \( k \) serving as a variable; and the minimum value of \( s \) is obtained when \( ds/dk = 0 \). Thus, the relation \(-2p'(w_j)q'(w_j) + 2kq'(w_j)^2 = 0\) provides the minimum \( s \) value, and \( k \) is eventually obtained from Eq. (2).

As for the spectrum \( r(w_j) \) that we try to find, the “intrinsic” length of the spectrum is invariant to changes in \( k \). Thus, under- or over-estimations of the scaling factor \( k \) yield solely increments in the lengths of spectra \( r(w_j) \), compared to the minimum length. In this way, it is possible to approximately determine the scaling factor \( k \) that corresponds to the spectrum \( r(w_j) \) independent of the background signal \( q(w_j) \).

Users of the proposed method should be careful concerning the following points, since this approach is not always perfect, accompanied by a few weak points. For instance, the method does not serve properly (i) when the desired \( r(w_j) \) and the background \( q(w_j) \) are identical or very similar to each other; (ii) when the signal to noise ratios of spectra \( p(w_j) \) and \( q(w_j) \) are low (that is, when the spectra are noisy).

In application of the proposed method, shown in Fig. 3(a), three arbitrarily selected spectra and the quartz Raman spectrum of Fig. 1 were exemplified. The set of the three spectra and the quartz spectrum were named as \( p(w_j) \) and \( q(w_j) \), respectively. Using the linear interpolation technique, the data were transformed to be evenly spaced, with \( \Delta w = 1.5 \). Figure 3(b) shows the three spectra, \( r(w_j) \), calculated by Eqs. (1) and (2).

A Savitzky–Golay differentiation filter was used to calculate the first derivative spectrum (21 data points and fourth degree polynomials were chosen). The spectra, \( r(w_j) \), obtained by the proposed method, were shown to have no contribution from the background spectrum \( q(w_j) \) (Fig. 3(b)). As the calculation for removing the background information is automatic, the proposed method will provide a smooth linkage to PCA based bio-Raman research.

In conclusion, we have developed an automated method for subtracting a background spectrum from bio-Raman spectra via scaling, resulting in the Raman spectra without contribution of the background spectrum. The scaling factor was automatically determined based on the concept of least length for subtracted
spectra. The proposed method was applied to three spectra of HeLa cells on quartz and the background spectrum of quartz without a cell. Based on the development of the method, we are further interested to (i) study whether the proposed method depends on the signal to noise ratio of spectra, (ii) extend the method to the use of other substrates such as glass and polystyrene bottoms for the background removal, and (iii) perform quantitative analysis of a pure chemical component in a multi-component system by applying the \( k \)-determine method to general Raman measurements. These will be reported elsewhere in the near future. In acknowledgment, the study was supported by KAKENHI 15H03821.

References