Hierarchical Cluster Analysis (HCA) of Microorganisms: An Assessment of Algorithms for Resonance Raman Spectra

ANN-KATHRIN KNIGGENDORF,* TOBIAS WILLIAM GAUL, and MERVE MEINHARDT-WOLLWEBER

Gottfried Wilhelm Leibniz University of Hannover, Institute of Biophysics, Herrenhaeuser Str. 2, 30419 Hannover, Germany (A.-K.K.); and Gottfried Wilhelm Leibniz University of Hannover, HOT–Hannoversches Zentrum für Optische Technologien, Nienburger Str. 17, 30167 Hannover, Germany (T.W.G., M.M.-W.)

Resonance Raman microspectroscopy in combination with hierarchical cluster analysis (HCA) is one of the most promising tools for the rapid examination of complex biological and medical samples. HCA is a ready, computerized tool for examining large sets of data for common characteristics, and a multitude of algorithms for this purpose have been developed over the years. However, resonance Raman spectra obtained from complex biological samples may originate from different chromophores as well as from a common chromophore found in different host environments, i.e., bacteria. Therefore, algorithms applied to resonance Raman spectra must handle data of high intrinsic similarity, i.e., spectra originating from a common chromophore, and data with highly dissimilar features, i.e., spectra from different chromophores, in the same unsupervised analysis. We examined the performance of eight widely used algorithms for hierarchical cluster analysis in clustering resonance Raman spectra of bacteria: Single-Linkage (Nearest-Neighbor), Complete-Linkage (Farthest-Neighbor), Average-Linkage, Weighted-Average-Linkage, Centroid, Median, and the Ward algorithm. Algorithm performance was evaluated by comparing the results of clustering a set of high-quality reference spectra with the results obtained when clustering a set of spectra recorded from single cells. References were formed by averaging 100 spectra of individual cells. While all algorithms returned highly similar results when clustering the reference spectra, their performance differed significantly when applied to single spectra. The best-performing algorithm, Weighted-Average-Linkage, correctly grouped single spectra with a reliability of above 95% while the spectral distances between the clusters deviated less than 10% from the results obtained with reference spectra. In contrast, the algorithm performing worst showed no similarity to the reference clustering at all. The widely used Ward algorithm deviated up to 30% from the reference in the spectral distances and returned a different spectral relation between bacteria expressing the same chromophore.

Index Headings: **Resonance Raman spectroscopy; Hierarchical cluster analysis; HCA; Cluster algorithm; Chromophores; Purple non-sulfur bacteria; *Nitrosomonas*.**

INTRODUCTION

Confocal resonance Raman microscopy combines the specificity of Raman fingerprints and a highly increased sensitivity under resonant excitation conditions with the spatial resolution of confocal laser microscopy. Its ability to analyze and identify microorganisms, based even on a single cell if necessary, without further preparation or previous cultivation is of great interest in a wide field of applications requiring fast and reliable methods for the identification of microbial organisms.[1,2] In contrast to most common identification methods, resonance Raman microscopy can be used on native as well as fixed samples, even though fixation may have a significant effect on the Raman signals.[3]

In combination with hierarchical cluster analysis (HCA), resonance Raman microspectroscopy is one of the most promising tools for the rapid in vivo identification of biological and medical samples, especially in the case of complex samples.[4–6] However, comparatively little thought has been given to the choice of HCA algorithms best suited for reliable clustering based on resonance Raman spectra. HCA is a means of structuring a complex set of observations into unique, mutually exclusive groups (clusters) of subjects similar to each other with respect to certain characteristics. It has become a widely used tool for the unsupervised structuring of complex experimental data, because—in contrast to other clustering techniques such as K-means clustering or support vector machines—it does not require a priori knowledge of the data and is not limited to the starting conditions.[7]

The most widely used algorithms for HCA are Single-Linkage, Complete-Linkage, Average-Linkage, Weighted-Average-Linkage, Centroid, Median, and—most notably—the Ward algorithm, which is especially popular for clustering biological data.[8,9] All of these algorithms choose or calculate a representative for each found cluster from the data clustered together. Subsequently, the distance between the clusters is calculated with a specified distance (or similarity) measure such as Euclidian distance or factorization.[7]

Raman spectra, as well as many other types of optical spectra, are composed of the desired signal, background intensity of different origins (sometimes this is called "determinate noise"), and random noise such as shot noise and detector noise. Resonance Raman spectra are primarily Raman spectra of resonantly excited molecules, i.e., chromophores, and resonance Raman spectroscopy owes its broad range of accessible samples to the ubiquity of these chromophores. In addition, different host environments affect the resonance Raman spectra, resulting in host-specific resonance Raman spectra originating even from the same chromophore.[10–12] Therefore, hierarchical clustering of biological samples based on resonance Raman spectra requires algorithms sensitive enough to handle spectra with a high intrinsic similarity due to a common resonant chromophore in the presence of highly dissimilar spectra originating from other chromophores.

In order to test HCA algorithms for this ability, we used a set of six different bacterial strains expressing three different chromophores: the carotenoids spheroidene and spirilloxanthin as similar, but not identical chromophores, and heme C as a chromophore distinctly different from the carotenoids in molecular structure, spectral characteristics, and metabolic function within the host organisms. The Raman spectra of all

Received 13 July 2010; accepted 29 October 2010.

* Author to whom correspondence should be sent. E-mail: kniggendorf@biophysik.uni-hannover.de.

DOI: 10.1366/10-06064

0003-7028/11/6502-0165$2.00/0

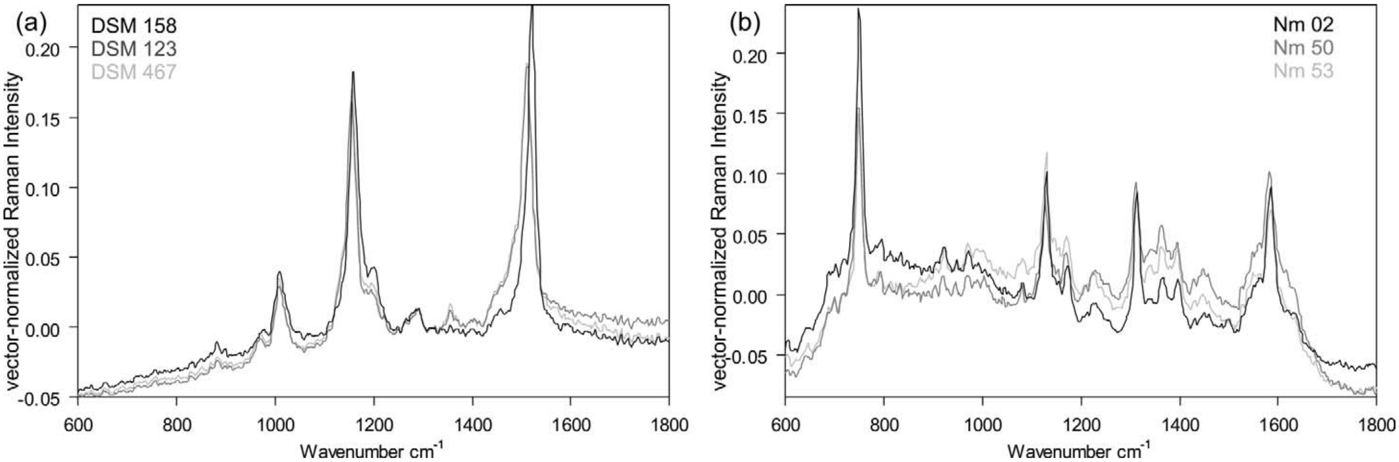

FIG. 1. Reference spectra (averages of 100 single spectra of S/N $\geq$ 10 each) of (**a**) *Rhodobacter sphaeroides* DSM 158^T, *Rhodopseudomonas palustris* DSM 123^T, and *Rhodospirillum rubrum* DSM 467^T, and (**b**) *Nitrosomonas communis* Nm-02 and two strands of *Nitrosomonas europaea* (Nm-50 and Nm-53). Note the slightly different Raman lines of *Rhodobacter* in comparison to the spectra of *Rhodopseudomonas* and *Rhodospirillum* (similar for *Nitrosomonas communis* in comparison to *Nitrosomonas europaea*). Spectral values below zero are due to the vector normalization of the spectra to this region.

three chromophores can be resonantly (pre-resonantly in the case of the carotenoids) enhanced with a frequency-doubled Nd : YAG laser at 532 nm as the excitation source.

MATERIALS AND METHODS

Bacterial Cultures and Chromophores. Three bacterial strains of photosynthetic Alphaproteobacteria—*Rhodobacter sphaeroides* DSM 158^T, *Rhodopseudomonas palustris* DSM 123^T, and *Rhodospirillum rubrum* DSM 467^T—were cultured using media and conditions described previously by Kniggendorf et al.[3] *Rhodobacter* contains spheroidene as the main chromophore for resonance Raman spectroscopy with 532 nm excitation, whereas *Rhodopseudomonas* and *Rhodospirillum* both form spirilloxanthin.[13]

In addition, three strains of *Nitrosomonae—Nitrosomonas communis* Nm-02, *Nitrosomonas europaea* Nm-50, and *Nitrosomonas europaea* Nm-53—holding heme C as part of cytochrome c were cultured using media and conditions described previously by Koops et al.[14]

Spheroidene and spirilloxanthin were chosen as similar but not identical chromophores because these carotenoids have the same function in the photoreaction centers and are specific to the respective bacteria.[15] Their Raman spectra as recorded in cultures of the aforementioned purple non-sulfur bacteria are given in Fig. 1a.

While heme C is also present in purple non-sulfur bacteria,[13] it requires significantly longer excitation times to record resonant Raman spectra of similar intensity than do the aforementioned carotenoids and thus the heme C does not affect the resonant Raman spectra of carotenoids recorded from purple non-sulfur bacteria. The cytochrome c based Raman spectra of *Nitrosomonas communis* and *Nitrosomonas europaea* are given in Fig. 1b.

Resonance Raman Measurements. Purple non-sulfur bacteria samples were prepared as follows: A volume of 1 mL of cell suspension was sampled from the respective actively growing culture at a cell density of 10^6 to 2×10^6 cells per milliliter (as estimated from OD), centrifuged for 5 min at 14 500 g at 4 °C, washed with phosphate buffered saline (PBS; pH 7.2), and pelletized again. For the *Nitrosomonas* cultures, a volume of 15 mL of cell suspension was sampled, which was centrifuged for 30 minutes at 8000 g, washed with PBS, and pelletized.

Washed cell pellets were mounted on standard microscope slides (ISO Norm 8037/l microscope slides by Menzel-Gläser, Braunschweig, Germany), covered with a 0.17 mm cover slip, and measured immediately.

Resonance Raman measurements were performed at room temperature with a confocal Raman microscope (CRM200, by WITec GmbH, Ulm, Germany), equipped with an oil-immersion objective (Nikon CFI Achromat) with a magnification of 100×, a numerical aperture of 1.25, and corrected for cover slips of 0.17 mm thickness. A stabilized, frequency-doubled continuous-wave Nd : YAG laser at 531.9 nm was used for excitation. The system had an ellipsoid measurement volume of approximately 0.5 μm^3 with a spatial resolution of 300 nm in the horizontal plane and 1.2 μm perpendicular to it. Slit width was 50 μm, realized by a multimode fiber connecting the Raman microscope with the spectrometer (UHTS 300, by WITec). The used grating had 600 lines per millimeter. Spectra were recorded with an electron multiplying charge-coupled device (emCCD) camera (ANDOR DU970N-BV-353), electrically cooled to −69 °C. The spectral resolution of the setup was 2 cm^{-1}, with a spectral accuracy of 1 cm^{-1}. Recorded spectra covered the range between −80 and 3710 rel. cm^{-1}.

Laser intensity was adjusted to 25 mW, giving 2.8 MW/cm^2 on the sample within the measurement volume. Measurement time per spectrum was set to 0.1 s for purple non-sulfur bacteria and 0.5 s for *Nitrosomonas* samples. Spectra of 100 different bacteria cells were recorded from each sample. Measured cells had a minimal distance of 3 μm from one another to avoid effects of photo-bleaching and thermal damage.

Spectral Analysis and Data Preparation. Preparation and analysis of the Raman spectra were done with commercial spectral analysis software (OPUS version 5.5 by Bruker).

Signal intensity was measured as intensity exceeding the underlying fluorescent background. In order to reliably quantify a noise level for each spectrum, the notch-filtered part of the data set was analyzed. Radiation below 115 cm^{-1} is blocked. The random noise was determined as the maximal amplitude about the mean intensity between 40 and 90 cm^{-1}. To verify the appropriateness of this definition, the random noise in the respective dark spectra was determined in intervals

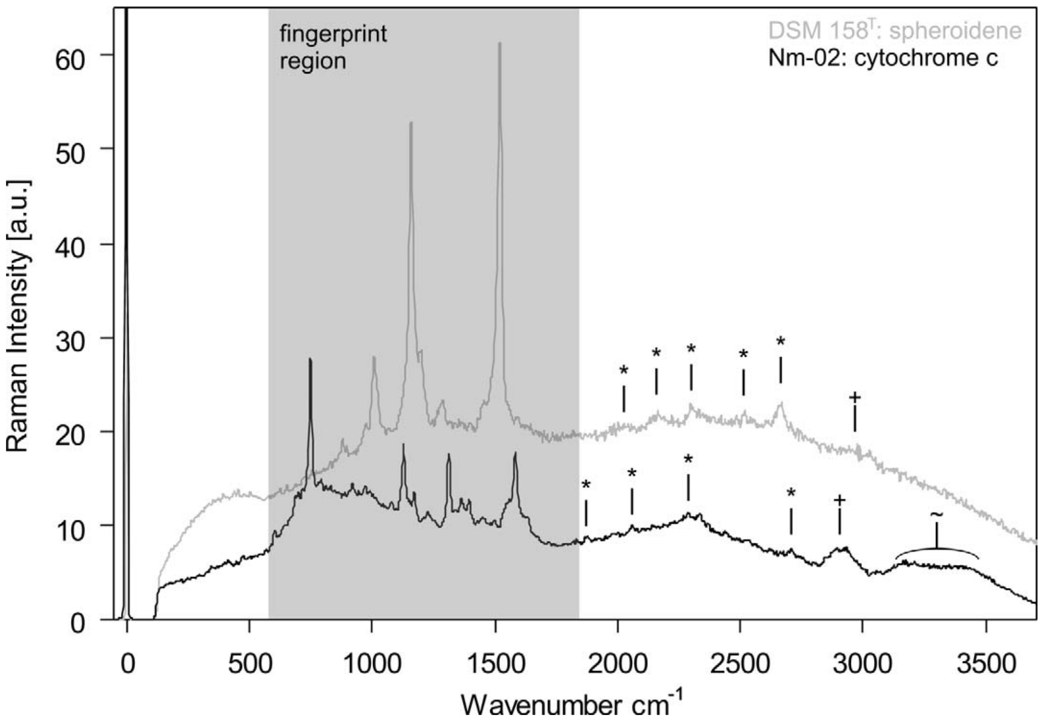

FIG. 2. Resonant Raman spectra of *Rhodobacter sphaeroides* DSM 158^T and *Nitrosomonas communis* Nm-02 as recorded. * denotes higher orders of the Raman bands seen in the fingerprint area (shaded gray). + denotes non-resolved CH stretching modes typical for bacterial substance. ~ denotes non-resolved OH stretching modes of water from the bacterial growth medium.

of 50 cm^{-1} around the positions of the four most prominent signal peaks of each chromophore. The results differed less than 5% from the noise level determined in the filtered area of the actual spectra records.

In-spectrum components of noise were deliberately not taken into account because the determination between tiny Raman lines and high random noise amplitudes is inherently unreliable due to the complexity and delicacy of the resonant chromophore spectra of cytochrome c and carotenoids, especially given different bacterial samples with possibly unknown Raman features.

The signal-to-noise ratio was determined for the four most prominent peaks of each spectrum and averaged to give a single signal-to-noise ratio (S/N) reflecting the quality of the whole spectrum within the interval of interest. Resonance Raman spectra of single bacterial cells (single spectra) with an S/N between 10 and 20 were used for hierarchical cluster analysis. Single spectra of lower or higher S/N were excluded to prevent spectrum quality from becoming a clustering criterion superior to the spectral differences caused by the chromophores (see the subsection Effects of Spectrum Quality in Hierarchical Cluster Analysis below).

Reference spectra for each bacterial strain type were formed by averaging 100 single spectra of S/N $\geq$ 10 of the recorded single spectrum from the same bacterial culture. Reference spectra had an S/N of approximately 100. A spectral comparison carried out using OPUS IDENT returned an agreement of 99.5% or better for different reference spectra of the same bacterial strain type.

All spectra were prepared for HCA as follows: The spectral region below 600 cm^{-1} was excluded, because it does not hold any of the prominent Raman peaks of the chromophores of interest. The spectral region above 1800 cm^{-1} was excluded because it does not add to the information about the investigated chromophores. It primarily holds higher-order Raman lines of the carotenoids spheroidene and spirilloxanthin as well as—in the case of heme C spectra—a set of Raman bands unrelated to the chromophore but very common in bacterial Raman spectra, such as C–H stretching bands around 2840 cm^{-1} and the broad Raman band associated with the O–H stretching modes of water around 3200 cm^{-1} (compare Fig. 2 outside the shaded area).

Spectra were then vector normalized to the spectral region of 600–1800 cm^{-1} according to

$$y'_k = (y_k - Y)/[\Sigma_k(y_k - Y)^2]^{1/2} \qquad (1)$$

where k is the number of pixels in the interval of interest (600–1800 cm^{-1}), y_k is the recorded intensity at pixel k, and Y is the mean value of y in the interval of interest, resulting in the vector norm of the spectrum being 1 in the interval of interest. This procedure is implemented in the OPUS software.

Hierarchical Cluster Analysis. In typical measurements of biological samples, only a single spectrum can be obtained from an individual cell; therefore, each algorithm was used on a set of single spectra contributing to the respective reference spectrum. In addition, a set of six high-quality spectra were clustered with each algorithm for reference.

All dendrograms were based on the spectral region 600–1800 cm^{-1}, which holds the most prominent and defining resonant Raman lines of the characteristic chromophores found in the samples.

Spectral distances were calculated using the Euclidian distance:

$$D_{ij} = [\Sigma_k(y_{jk} - y_{ik})^2]^{1/2} \qquad (2)$$

with the spectral distance D_{ij} between the spectra y_i and y_j and k going over all recorded data points within the spectral region 600–1800 cm^{-1} used for the clustering.

The following algorithms were examined for the hierarchical cluster analysis of resonance Raman spectra:

(a) Single-Linkage (Nearest-Neighbor-Clustering): $D(r,i) = \min[D(p,i), D(q,i)]$
(b) Complete-Linkage (Farthest-Neighbor-Clustering): $D(r,i) = \max[D(p,i), D(q,i)]$
(c) Average-Linkage: $D(r,i) = [D(p,i) + D(q,i)]/2$
(d) Weighted-Average-Linkage: $D(r,i) = [n(p)D(p,i) + n(q)D(q,i)]/[n(p) + n(q)]$
(e) Centroid: $D(r,i) = \{[n(\mathrm{p})D(p,i) + n(q)D(q,i)]/n\} + \{[n(p) + n(q)D(q,p)]/n^2\}$
(f) Median: $D(r,i) = \{[D(p,i) + D(q,i)]/2\} - [D(p,q)/4]$
(g) Ward: $H(r,i) = \{[n(p) + n(i)]H(p,i) + [n(i) + n(q)]H(q,i) - n(i)H(q,i)\}/[n + n(i)]$

where r is the cluster formed out of the clusters p and q, $D(r,i)$ is the spectral distance between the clusters r and i, $H(r,i)$ is the heterogeneity between the clusters r and i, $n(i)$ is the number of spectra included in cluster i, and n is the number of spectra included in the clustering. The variables $H(p,i)$, $H(q,i)$, $D(p,i)$, $D(q,i)$, $D(q,p)$, $n(p)$, and $n(q)$ are defined analogously.

In contrast to algorithms (a) through (f), the Ward algorithm (g) does not compute the spectral distance between clusters. Instead it calculates the within-cluster sum of squares (also called the error sum of squares) of every possible cluster.[8] Clusters are chosen so that this sum of squares within all clusters is minimized.[7] Therefore, the so-called heterogeneity H relates to a spectral distance D as used by algorithms (a) through (f) as

$$H(r,i) \sim D(r,i)^2 n(r+i) \tag{3}$$

where $n(r+i)$ is the number of spectra included in the clusters r and i. Therefore, the dendrogram calculated by the Ward algorithm is proportional to the weighted square of the spectral distances rather than the spectral distances themselves. In order to achieve quantitative comparability between the results of the Ward algorithm (g) and the clustering algorithms (a) through (f), the spectral distance D was calculated from the heterogeneity H.

Hierarchical cluster analyses were performed with the additional software module OPUS IDENT for OPUS version 5.5 by Bruker incorporating algorithms (a) through (g). The clustering results were analyzed on the level of four clusters with respect to the spectral distance between clusters and the spectral relation of the examined bacteria species grouped into the clusters. For example, *Rhodobacter* holding spheroidene forming cluster (1) and *Rhodopseudomonas* and *Rhodospirillum*, both with spirilloxanthin, together being grouped into cluster (2). In addition, the number of spectra assigned to a wrong cluster was taken into account, independent of spectral distance or spectral relation of the clusters.

RESULTS AND DISCUSSION

Reference Spectra Clustering. Reference spectra were obtained from six different bacterial strains: *Rhodobacter sphaeroides* DSM 158$^{\mathrm{T}}$ (main Raman chromophore: spheroidene), *Rhodopseudomonas palustris* DSM 123$^{\mathrm{T}}$ (spirilloxanthin), *Rhodospirillum rubrum* DSM 467$^{\mathrm{T}}$ (spirilloxanthin), *Nitrosomonas communis* Nm-02 (cytochrome c), *Nitrosomonas europaea* Nm-50 (cytochrome c), and *Nitrosomonas europaea* Nm-53 (cytochrome c). All reference spectra used in this study had a S/N ratio of approximately 100. Reference spectra taken from the same sample but from different cells showed less than 0.5% variance when compared with the OPUS IDENT routine for spectral identity, giving a spectral distance of 0.11 or less when subjected to HCA.

Figures 1a and 1b show, respectively, the carotenoid and cytochrome c based reference spectra as they were used for HCA. The dendrogram resulting from the hierarchical cluster analysis of the six reference spectra is given in Fig. 3.

All applied HCA algorithms returned the same spectral relation between the tested bacteria: one branch of purple non-sulfur bacteria with *Rhodobacter* DSM 158$^{\mathrm{T}}$ in one cluster (1), and *Rhodopseudomonas* DSM 123$^{\mathrm{T}}$ and *Rhodospirillum* DSM 467$^{\mathrm{T}}$ grouped together in a second cluster (2), while the *Nitrosomonae* were collected in a second branch with *Nitrosomonas communis* Nm-02 separated in cluster (3) and the two strains of *Nitrosomonas europaea* Nm-50 and Nm-53 collected together in another cluster (4). The cluster labels (1) through (4) were assigned solely with respect to the bacteria primarily allocated to the clusters. The spectral distances found by the investigated algorithms are given in Table I.

As can be seen, the resonant Raman spectra from different chromophores can easily be separated. Clusters (1) and (2) have a spectral distance of 0.45 ± 0.02 depending on the applied algorithm. The Ward algorithm calculated a spectral heterogeneity of 0.58, which gives a spectral distance of 0.44. The chromophore of all spectra sorted into clusters (3) and (4) is cytochrome c. The spectral distance found between these clusters is given as 0.50 ± 0.09 by all tested algorithms.

Comparison of the spectra in Fig. 1 provides an impression of how spectral distance relates to visible spectral differences. Please note that the largest spectral distance between reference spectra of the same bacterial culture was found to be 0.11 or less by all algorithms.

The spectral distance found between the clusters holding purple non-sulfur bacteria (1 and 2) and the clusters holding *Nitrosomonae* (3 and 4) is given as 1.4 by all algorithms save the Ward algorithm. The heterogeneity calculated by the Ward algorithm was 22.3, corresponding to a spectral distance of 1.9.

It is noteworthy, that in comparison to the algorithms (a) through (f) directly using the spectral distance of clusters, the Ward algorithm (g) computes larger spectral distances between clusters of spectra originating from the same (cytochrome c) or from distinctly different chromophores (cytochrome c versus carotenoids), but the spectral distances computed between clusters of spectra originating from similar—but not identical—chromophores (spheroidene, spirilloxanthin) are smaller than those determined by the algorithms (a) through (f).

Single Spectra Clustering. A total of 44 single spectra previously contributing to the reference spectra were chosen for the clustering: 8 spectra of *Rhodobacter sphaeroides*, 9 spectra of *Rhodopseudomonas palustris*, 10 spectra of *Rhodospirillum rubrum*, 4 spectra of *Nitrosomonas communis*, and 13 spectra of *Nitrosomonas europaea* (7 spectra of Nm-50 and 6 spectra of Nm-53). The size of the individual sample batches was varied to simulate realistic conditions when clustering data of unknown mixed samples. Figure 4 gives a vector-normalized single spectrum of *Nitrosomonas communis* with an S/N of 11.5 in comparison to the corresponding reference spectrum to

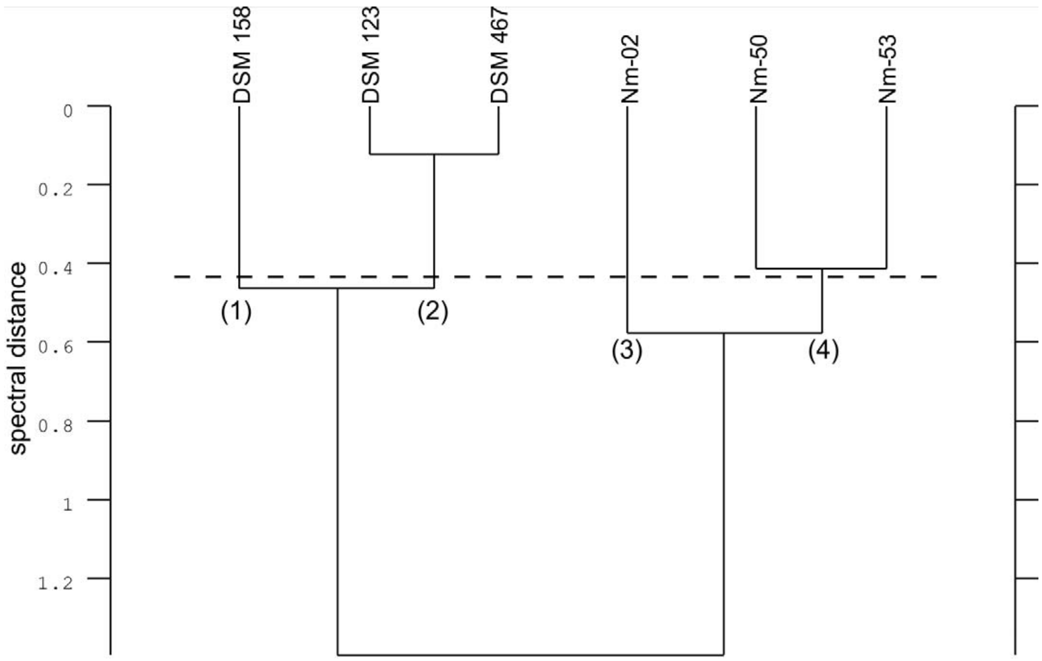

FIG. 3. Reference dendrogram. All examined algorithms clustered the reference spectra as shown here. For the exact distances between clusters (1) and (2) and between (3) and (4), see Table I.

illustrate the differences in spectrum quality between single and reference spectra.

As can be seen from the data given in Table II, the clustering algorithms performed very differently when used on non-averaged resonance Raman spectra (single spectra) with an S/N between 10 and 20. Distinctly different chromophores (carotenoids and cytochrome c) were also separated by all seven of the tested algorithms when applied to single spectra. Slightly different chromophores (the carotenoids spheroidene and spirilloxanthin) were separated by all algorithms except Average-Linkage (c), which placed all carotenoid spectra in one group. The spectral relation between the three types of purple non-sulfur bacteria, i.e. *Rhodopseudomonas* and *Rhodospirillum* being grouped into one cluster separate from *Rhodobacter*, as established in the clustering of the reference spectra, was also maintained by all algorithms except Average-Linkage.

Reliability in the clustering of individual single spectra ranged from above 97% (Weighted-Average-Linkage (d)), over 95% (Single-Linkage (a) and Ward (g)) and 90% (Median (f) and Centroid (e)), down to 85% (Complete-Linkage (b)).

Spectra originating from the same chromophore (cytochrome c) were grouped with a reliability of 90% or better by all algorithms, excluding Average-Linkage. However, only two algorithms—Weighted-Average-Linkage (d) (the dendrogram is given in Fig. 5) and Centroid (e)—maintained the spectral relation of *Nitrosomonas communis* Nm-02 being separated (cluster 3) from the two strands of *Nitrosomonas europaea* (Nm-50 and Nm-53 grouped in cluster 4) found when clustering the reference spectra.

These differences in performance when clustering resonance Raman spectra of single cells can be understood from the perspective of the respective clustering procedures used by the algorithms. The best algorithm, Weighted-Average-Linkage (d), determines the spectral distance between clusters by forming the weighted average of their components. In terms of Raman spectra this amounts to calculating the averaged spectrum of the respective cluster. Thus, the S/N ratio of the spectra representing the clusters improves with each clustering step, resulting in almost perfect agreement (1% deviation from

TABLE I. Reference spectra clustering.

Algorithm	Cluster	Spectral distance	Species	Chromophore
Single-Linkage	1	0.451	DSM 158	spheroidene
	2		DSM 123, DSM 467	spirilloxanthin
	3	0.518	Nm 02	cytochrome c
	4		Nm 50, Nm 53	cytochrome c
Complete-Linkage	1	0.474	DSM 158	spheroidene
	2		DSM 123, DSM 467	spirilloxanthin
	3	0.631	Nm 02	cytochrome c
	4		Nm 50, Nm 53	cytochrome c
Average-Linkage	1	0.463	DSM 158	spheroidene
	2		DSM 123, DSM 467	spirilloxanthin
	3	0.575	Nm 02	cytochrome c
	4		Nm 50, Nm 53	cytochrome c
Weighted-Average-Linkage	1	0.463	DSM 158	spheroidene
	2		DSM 123, DSM 467	spirilloxanthin
	3	0.575	Nm 02	cytochrome c
	4		Nm 50, Nm 53	cytochrome c
Centroid	1	0.431	DSM 158	spheroidene
	2		DSM 123, DSM 467	spirilloxanthin
	3	0.471	Nm 02	cytochrome c
	4		Nm 50, Nm 53	cytochrome c
Median	1	0.431	DSM 158	spheroidene
	2		DSM 123, DSM 467	spirilloxanthin
	3	0.471	Nm 02	cytochrome c
	4		Nm 50, Nm 53	cytochrome c
Ward[a]	1	0.438 [0.575]	DSM 158	spheroidene
	2		DSM 123, DSM 467	spirilloxanthin
	3	0.458 [0.629]	Nm 02	cytochrome c
	4		Nm 50, Nm 53	cytochrome c

[a] The spectral distance D given for the Ward algorithm was calculated from the heterogeneity H (given in square brackets) based on Eq. 3 with $H(r,i) = n(r + i)\, D(r,i)^2$.

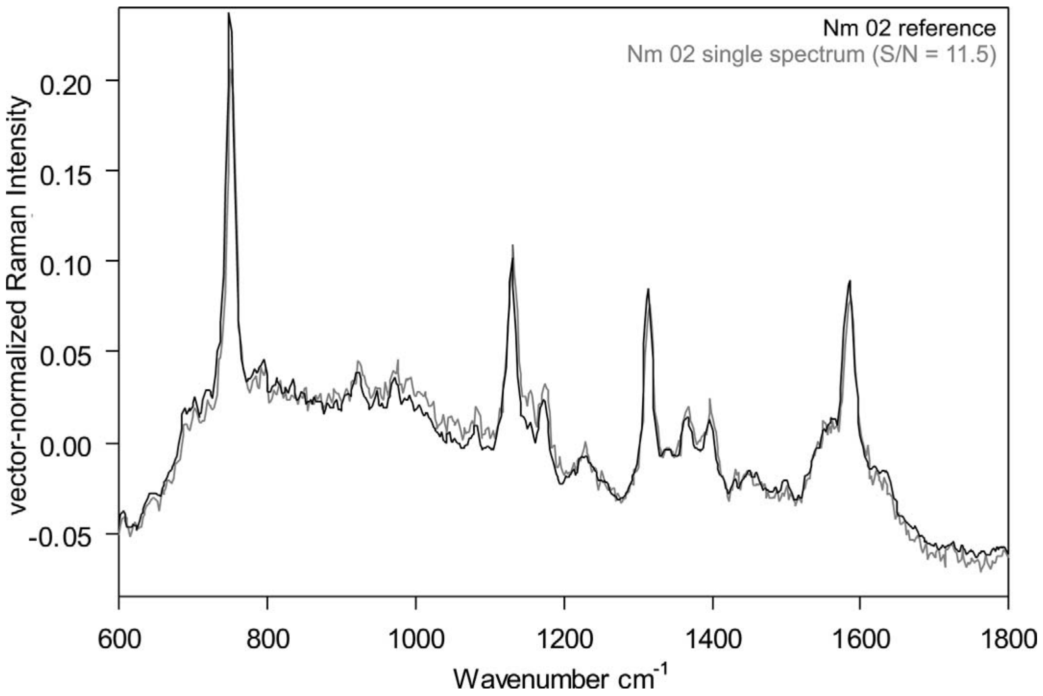

FIG. 4. Reference spectrum (average of 100 single spectra of S/N $\geq$ 10) and single spectrum (S/N = 11.5) of *Nitrosomonas communis* Nm-02. The spectra are presented as subjected to HCA: in vector-normalization. Spectral values below zero are due to the vector normalization of the spectra to this region.

TABLE II. Single spectra clustering.

Algorithm	Cluster	Spectral distance	Deviation from ref. (%)	Spectral relation maintained?	Assignment errors
Single-Linkage	1	0.377	16	+	
	2				2
	3	0.222	57	−	
	4				
Complete-Linkage	1	0.666	41	+	
	2				6
	3	0.931	48	−	
	4				2
Average-Linkage	1	n. a.	n. a.	−	19
	3				2
	4		n. a.	−	
	4’				
Weighted-Average-Linkage	1	0.493	7	+	
	2				1
	3	0.581	1	+	
	4				2
Centroid	1	0.363	25	+	
	2				7
	3	0.353	25	+	
	4				2
Median	1	0.334	23	+	
	2				6
	3	0.389	17	−	
	4				2
Ward[a]	1	0.389 [4.090]	13	+	
	2				2
	3	0.356 [2.152]	29	−	
	4				2

[a] The spectral distance D given for the Ward algorithm was calculated from the heterogeneity H (given in square brackets) based on Eq. 3 with $H(r,i) = n(r+i)\,D(r,i)^2$.

the reference clustering) in the case of spectra originating from the same chromophore (cytochrome c in *Nitrosomonas*) and still a very good agreement (7% deviation) in the case of slightly different chromophores (carotenoids in purple non-sulfur bacteria).

In contrast, the second algorithm maintaining the spectral relation between the bacteria, Centroid (e), shows better agreement only in later stages of the clustering when the clusters hold a larger number of spectra, resulting in the centroid being more representative of the weighted spectral average. Thus, the spectral relation determined with reference spectra is reproduced, but a comparatively high number of single spectra are assigned to the wrong sub-clusters.

In comparison, the spectral distances calculated by algorithms failing to reproduce the spectral relation between the bacteria deviated from the results obtained by clustering reference spectra between 20% (Median algorithm (f) with eight erroneously assigned single spectra) and 45% (Complete-Linkage (b), also with eight erroneously assigned single spectra). Median, choosing the spectrum closest to the centroid of the cluster as representative of the cluster, shows an amount of erroneously assigned single spectra comparable to that of the Centroid algorithm. Average-Linkage (c) was not considered in this part of the analysis, because its result showed no similarity to that of the reference clustering at all.

The widely used Ward algorithm (g) as well as Single-Linkage (a) reproduced the spectral distance between clusters of spectra from slightly different chromophores (carotenoids in purple non-sulfur bacteria) with only 13% (16% in the case of Single-Linkage) deviation in the spectral distance between clusters (1) and (2) (and two erroneously assigned spectra), but failed to maintain the spectral relation between the *Nitrosomonae* (common chromophore: cytochrome c) with a deviation of 29% (57% respectively) in the spectral distance found between clusters (3) and (4). Figure 6 shows the

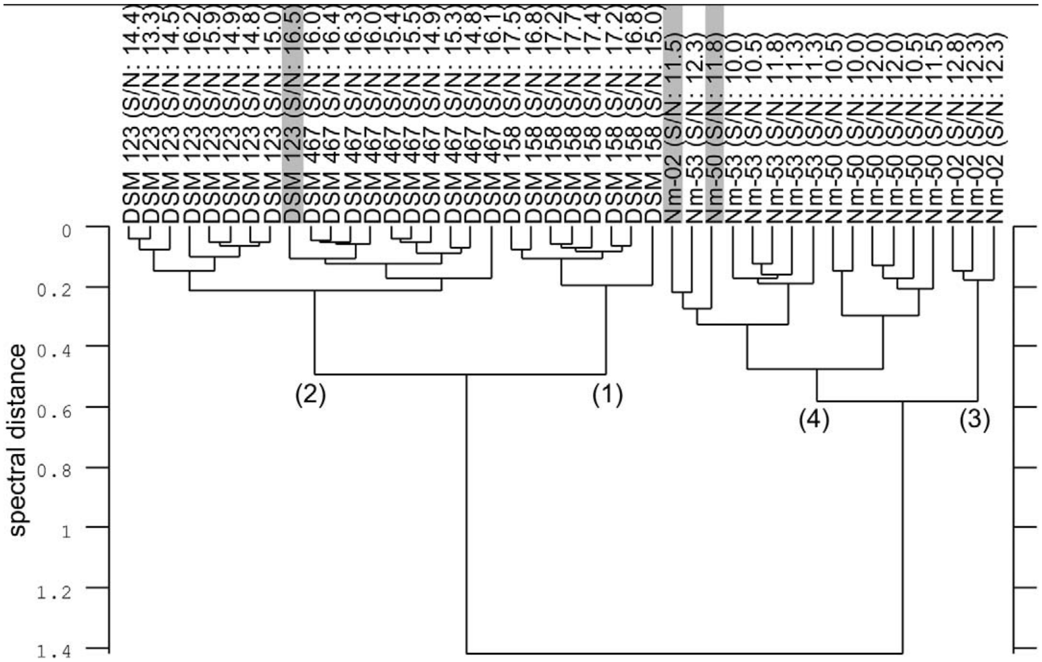

FIG. 5. Resonance Raman spectra of single cells (S/N $\geq$ 10) clustered with the Weighted-Average-Linkage algorithm. Grayed spectra were assigned to the wrong cluster. For the exact distances found between clusters (1) and (2) and between (3) and (4), see Table II.

heterogeneity dendrogram as calculated by the Ward algorithm. This similarity in results is unsurprising when considering that the Single-Linkage algorithm (a) determines the spectral distance between clusters as the minimum of the spectral distances between the components of said clusters, while the Ward algorithm (g) does essentially the same with a weighted square of the spectral distances, giving it a slight advantage when handling spectra of the same chromophore (cytochrome c in *Nitrosomonas*).

To the best of our knowledge, there has not been a performance evaluation of HCA algorithms for (resonant) Raman spectra so far, but several studies have been conducted for other types of data, such as the concentrations of chemical compounds used as markers for eutrophic levels in coastal

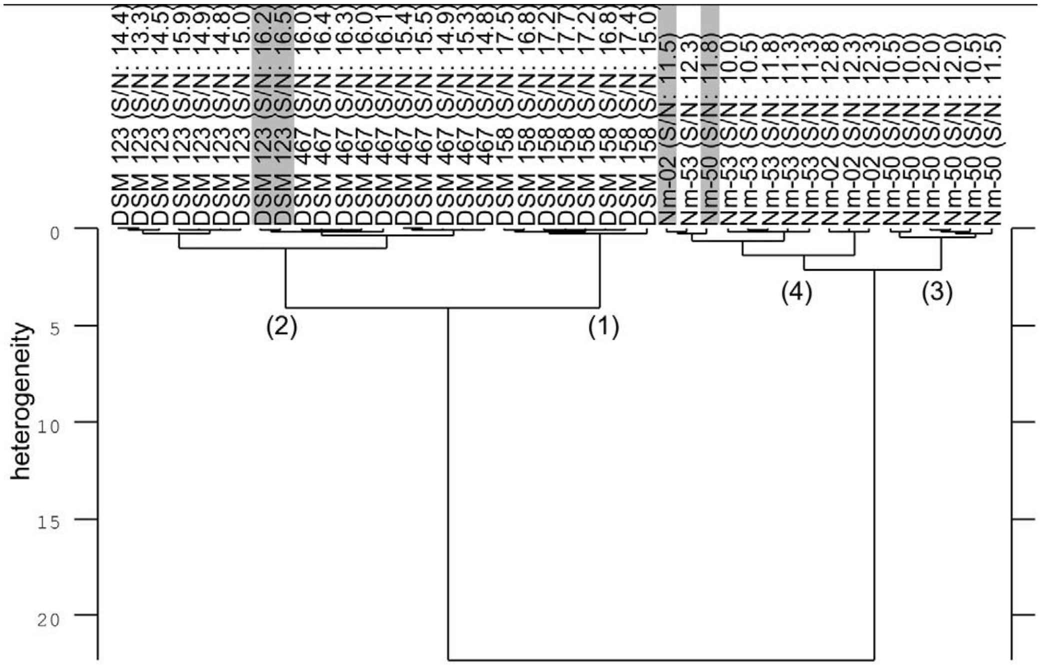

FIG. 6. Single spectra of S/N $\geq$ 10 clustered using the Ward algorithm. Grayed spectra were assigned to the wrong cluster. Please note that the Ward algorithm computes the heterogeneity H, which is proportional to the square of the spectral distance rather than the spectral distance itself. For the exact Heterogeneity found between clusters (1) and (2) and between (3) and (4), see Table II.

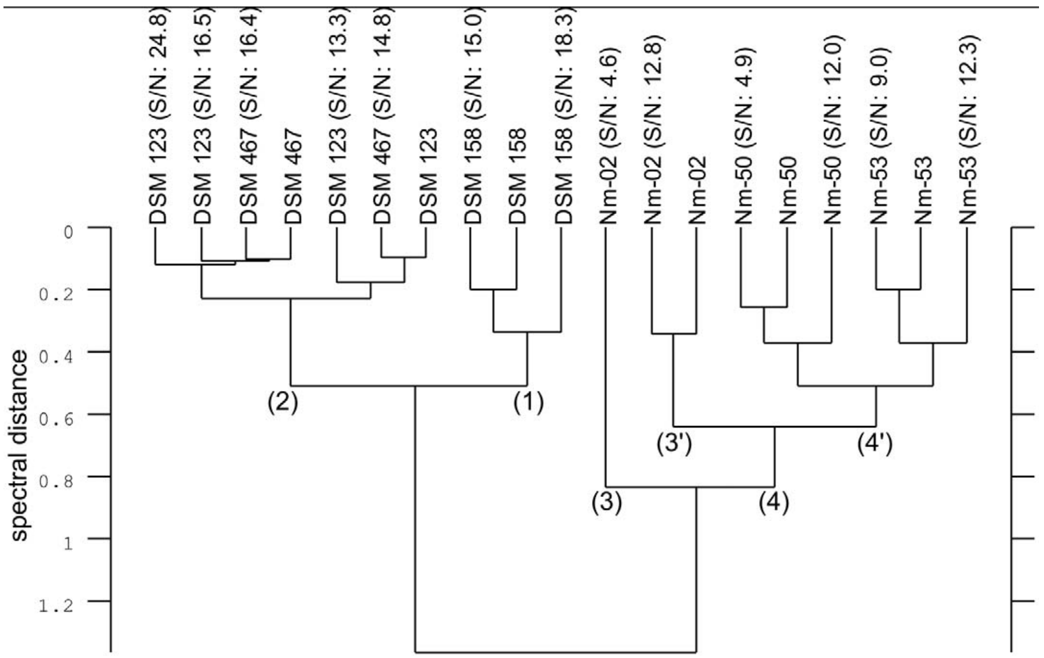

FIG. 7. Effects of low signal-to-noise ratios in hierarchical cluster analysis of resonance Raman spectra. The dendrogram was calculated with the Weighted-Average-Linkage algorithm, which performed best among the tested algorithms. Single spectra are indicated by their S/N ratio given in brackets.

waters,[9] most of which show a clear advantage for the Ward algorithm. However, the data subjected to clustering in these studies are scalars, whereas Raman spectra are vectors. These vectors describe intensity over pixel, which is composed of signal intensity, unavoidable background intensity (determinate noise), and random noise. Weighted-Average-Linkage (and to a lesser degree, Centroid) reduces random noise when calculating the representative spectra of clusters, whereas the procedures of other algorithms do not affect the random noise component of the spectra.

In addition, our findings for the Ward algorithm are in good agreement with the results of Harz et al., analyzing Raman spectra of single bacterial cells from cerebrospinal fluid during bacterial meningitis.[5] They reported that the Ward algorithm correctly grouped an average of 94% of the spectra (spectral relation was not considered).

Effects of Spectrum Quality in Hierarchical Cluster Analysis. Each algorithm handles the differences between spectra obtained from single cells differently (see Materials and Methods: Hierarchical Cluster Analysis, above). These differences in the spectra are due to the inherent heterogeneity of bacterial spectra per se, as well as the variant spectrum quality, quantified by the S/N ratio.

Large differences in the S/N ratio may have severe effects on the results of clustering Raman spectra obtained from single cells. While spectra of distinctly different chromophores (carotenoids versus cytochrome c) are still separated properly, the differences in the S/N ratios may well exceed signal variations caused by different host organisms in spectra originating from the same chromophore (cytochrome c or spirilloxanthin), as can be seen in Fig. 7 with the grouping of DSM 123 and DSM 467, both of which carry spirilloxanthin as the main chromophore. This agrees with the findings of Bonifacio et al.,[4] who observed that HCA (Ward algorithm)—for them, unexpectedly—also distinguished between intense and weak resonant Raman spectra of intra-cellular heme molecules.

In addition, spectra of comparatively low quality may result in early separation from the main bulk and the creation of additional clusters. An automated analysis of the dendrogram given in Fig. 7 returns the marked clusters (3) and (4), whereas the important species separation happens in the marked clusters (3′) and (4′), both grouped together in (4). This also agrees with Bonifacio et al.[4]

To avoid these unwanted effects, it is imperative to subject only spectra within a certain range of spectrum quality to HCA, thus excluding spectra of the lowest quality as well as spectra well above average quality. The range of spectrum quality is best chosen so that the majority of the recorded data may be included in the HCA. This corresponds with the requirement of excluding extremes from the data intended for clustering,[7] which—in the case of scalar data (see Primpas et al.[9] for example)—is typically done by filtering extremes during the matrix formation.

CONCLUSION

The choice of a suitable algorithm for the hierarchical cluster analysis of bacterial Raman spectra proved to be non-trivial. The algorithms tested showed surprisingly disparate results. Of all algorithms, only Weighted-Average-Linkage can be fully recommended for clustering resonant Raman spectra of single cells independent of chromophore, because its clustering procedure optimizes spectrum quality with each clustering step through the reduction of random noise. In clustering non-averaged spectra of single cells, it reproduced results obtained by clustering high-quality reference spectra with an accuracy of 99% (in the case of spectra originating from the same or distinctly different chromophores) and 93% in the case of slightly different chromophores. In contrast, the widely used and often recommended Ward algorithm only reproduced the reference results with an accuracy of 87% for spectra originating from slightly different chromophores and less than 70% for spectra from the same chromophore, failing to maintain their spectral relation.

Because Raman spectra are vector rather than scalar data, algorithms performing weighting and averaging (Weighted-Average-Linkage, Centroid), thus reducing the effect of random noise in the spectra representing clusters, perform superior to algorithms choosing or calculating the representatives based on a fixed relation. Nevertheless, algorithms calculating the spectral distances based on minimal distances between their components (Single-Linkage, Ward) still produce acceptable results in grouping individual spectra, but they may not return the spectral relation of bacteria expressing the same chromophore.

The signal-to-noise ratio range is an important parameter for the successful clustering of Raman spectra, which needs to be monitored closely in order to prevent spectrum quality from becoming a clustering criterion superior to the spectral differences caused by different chromophores and/or host organisms.

The algorithms Complete-Linkage and Average-Linkage cannot be recommended for the hierarchical cluster analysis of resonant Raman spectra.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Andreas Pommerening-Röser of the Biozentrum Klein Flottbek, Mikrobiologie und Biotechnologie, in Hamburg for providing us with the *Nitrosomonas* strains and invaluable information about their cultivation. The technical assistance of Heidemarie Bliedung with culturing the investigated strains is gratefully acknowledged. This work was kindly supported by the German Research Foundation (DFG, grant no. AN 712/1-5).

1. M. Harz, M. Kiehntopf, S. Stöckel, P. Rösch, E. Straube, T. Deufel, and J. Popp, J. Biophoton. **2,** 70 (2009).
2. A. Kudelski, Talanta **76,** 1 (2008).
3. A.-K. Kniggendorf, T. W. Gaul, and M. Meinhardt-Wollweber, Microsc. Res. Tech. (2010), DOI: 10.1002/jemt.20889.
4. A. Bonifacio, S. Finaurini, C. Krafft, S. Parapini, D. Taramelli, and V. Sergo, Anal. Bioanal. Chem. **392,** 1277 (2008).
5. M. Harz, P. Rösch, and J. Popp, Cytometry A **75,** 104 (2009).
6. K. M. Tan, C. S. Herrington, and C. T. A. Brown, J. Biophoton. (2010), DOI: 10.1002/jbio.201000083.
7. S. Sharma, *Applied Multivariate Techniques* (John Wiley & Sons, New York, 1996), Chap. 7, pp. 185–217.
8. J. H. Ward, J. Am. Statist. Assoc. **58,** 236 (1963).
9. I. Primpas, M. Karydis, and G. Tsirtis, Global NEST J. **10,** 359 (2008).
10. A. Desbois, Biochimie **76,** 693 (1994).
11. P. Qian, K. Saiki, T. Mizoguchi, K. Hara, T. Sashima, R. Fujii, and A. Koyama, Photochem. Photobiol. **74,** 444 (2001).
12. R. Schweitzer-Stenner, Q. Huang, A. Hagarmann, M. Laberge, and J. A. Carmichael, J. Phys. Chem. B **111,** 6527 (2007).
13. J. F. Imhoff, H. G. Trüper, and N. Pfennig, Int. J. Syst. Bacteriol. **34,** 340 (1984).
14. H.-P. Koops, B. Böttcher, U. C. Möller, A. Pommerening-Röser, and G. Stehr, J. Gen. Microbiol. **137,** 1689 (1991).
15. G. Britton, "Overview of Carotenoid Biosynthesis", in *Carotenoids Volume 3: Biosynthesis and Metabolism,* G. Britton, S. Liaaen-Jensen, and H. Pfander, Eds. (Birkhäuser, Basel, 1998), pp. 42–54.