Multivariate Analysis of Single-Molecule Spectra: Surpassing Spectral Diffusion

Clemens Hofmann,[1] Hartmut Michel,[2] Marin van Heel,[3] and Jürgen Köhler[1]

[1]*Experimental Physics IV and BIMF, University of Bayreuth, 95440 Bayreuth, Germany*
[2]*Department of Molecular Membrane Biology, Max-Planck Institute of Biophysics, Frankfurt, Germany*
[3]*Department of Biological Sciences, Imperial College London, London SW7 2AY, United Kingdom*

(Received 4 October 2004; published 16 May 2005)

The full exploitation of single-molecule spectroscopy in disordered systems is often hampered by spectral diffusion processes of the optical transitions due to structural fluctuations in the local environment of the probe molecule which leads to temporal averaging of the signal. Multivariate statistical pattern recognition techniques, originally developed for single-molecule cryoelectron microscopy, allow us to retrieve detailed information from optical single-molecule spectra. As an example, we present the phonon side band of the B800 excitations of the light-harvesting 2 (LH2) complex from *Rhodospirillum molischianum,* revealing the electron-phonon coupling strength for these transitions. The measured Debye-Waller factors, ranging from 0.4 to 0.9, fall in the regime of weak electron-phonon coupling.

DOI: 10.1103/PhysRevLett.94.195501 PACS numbers: 63.20.Kr, 02.50.Sk, 33.50.Dq, 73.22.−f

Single-molecule spectroscopy provides the unique opportunity to study chromophores inherently free from ensemble averaging which is of particular interest for the investigation of disordered host systems like polymers or proteins [1–4]. However, any variation in the environment of the probe molecule leads to changes of the local interactions and concomitantly to fluctuations of the electronic levels. The single-molecule signal itself is thus free from averaging only if each emission act takes place under exactly the same conditions. As a consequence, the spectral profile obtained from a single molecule depends crucially on the relationships between the temporal resolution of the experiment, δt_{exp}, the time scale of the spectral fluctuations, δt_{fluc}, and the duration of the experiment, δT_{exp} [1]. Therefore full exploitation of the advantages of single-molecule spectroscopy in a disordered host is often limited by spectral diffusion which results in temporal averaging of the signal.

Typically, the spectrum of the first electronically excited state of an organic molecule embedded in a matrix gives rise to a homogeneously broadened zero-phonon line (ZPL) accompanied by a relatively broad phonon side band (PSB) [5]. The ZPL results from the pure electronic transition, whereas the PSB corresponds to an electronic transition in combination with the simultaneous excitation of a vibration of the host. For convenience we refer to these vibrations as phonons, independent of the crystallinity of the host system. The distribution of the intensity ratio between the ZPL and the PSB, i.e., the profile of the electronic spectrum, is determined by the electron-phonon coupling strength, whereas the width of the ZPL reflects the total dephasing time of the electronically excited state. The ZPL approaches its lifetime limited linewidth at low temperatures [6] and provides an extremely sensitive tool to monitor interactions between the molecule and the host matrix [7,8]. Often single-molecule spectroscopy experiments are performed by repeatedly scanning the excitation laser through the absorption spectrum of the single molecule resulting in a stack of consecutively recorded individual spectra [9,10]. δt_{exp} then corresponds to the time needed to scan the laser across the molecular absorption spectrum once, and δT_{exp} corresponds to the duration of the sequence of scans. For molecules embedded in a highly ordered environment, for instance, a single crystal, $\delta t_{\mathrm{fluc}} \gg \delta T_{\mathrm{exp}}$, and the ZPL of an individual molecule can therefore be detected [6]. In contrast, if $\delta t_{\mathrm{exp}} \gg \delta t_{\mathrm{fluc}}$ the spectral diffusion remains unresolved and the profile of the electronic spectrum is washed out [11]. For single molecules in polymers and proteins, an intermediate situation is often encountered, namely, that $\delta t_{\mathrm{exp}} \ll \delta t_{\mathrm{fluc}} \ll \delta T_{\mathrm{exp}}$. Under these conditions, the ZPL of a single molecule can be identified in a single scan and its ''spectral trail'' can be followed throughout the stack of spectra, which thus allows one to obtain valuable information about the dynamics of the nanoenvironment of the probe molecule [1,6,10,12]. However, such a single spectral scan is typically extremely noisy and does not allow for a quantitative analysis of the profile of the electronic spectrum.

Here we illustrate how the underlying signal can nevertheless be retrieved from the extremely noisy data using the example of the B800 absorptions from an individual light-harvesting 2 (LH2) complex from *Rhodospirillum molischianum*. In this example we were able to recover the (high signal-to-noise ratio) profile of the electronic spectrum and the electron-phonon coupling strength from the data by our statistical analysis. The LH2 complex is a pigment-protein complex that plays an important role in bacterial photosynthesis. *Rhodospirillum molischianum* LH2 comprises 24 BChl *a* molecules arranged in two concentric rings [13]. One ring features 18 closely interacting BChl *a* molecules (B850) with an absorption band at about 850 nm. The other ring consists of eight weakly coupled BChl *a* molecules (B800) absorbing at around 800 nm. The B800 band shows several relatively narrow spectral lines resulting from the absorptions of individual

BChl *a* molecules. These spectral lines feature a strong dependence on the polarization of the incident radiation consistent with the circular arrangement of the molecular transition-dipole moments [14].

The LH2 complexes of *Rhodospirillum molischianum* were isolated as described previously [15]. The samples were prepared by spin coating a highly diluted LH2-detergent solution on a Li-fluoride substrate, and then cooled in a liquid He bath cryostat to 1.4 K. The Ti:sapphire laser-excited fluorescence (spectral window: 20 nm; bandwidth of excitation: 1 cm^{-1}) around 880 nm was detected with an avalanche photodiode (SPCM-AQR-16, EG&G). Further details can be found in [16].

In Fig. 1(a) we show a stack of 7700 fluorescence-excitation spectra of the B800 band from an individual LH2 complex, recorded consecutively at 1.4 K. In Fig. 1(b) we compare five spectra extracted from the data in Fig. 1(a). The top trace shows the total average of all 7700 scans, the three traces in the center of the illustration show the spectra for three successive individual sweeps, and the lowest trace shows the average of a few individual sweeps around scan no. 2040. The individual spectra feature narrow absorptions that are attributed to zero-phonon lines from individual BChl *a* molecules of the B800 assembly. However, the signal-to-noise ratio does not permit one to resolve the profile of the PSB. From the lower trace in Fig. 1(b) it is evident that simple averaging of successive individual traces to improve the signal-to-noise ratio is not an option.

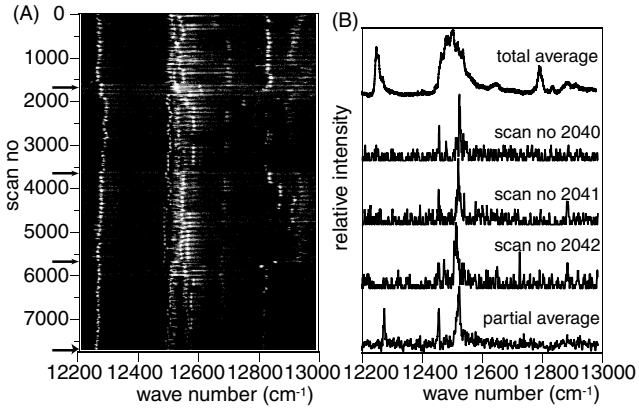

FIG. 1. (a) Two-dimensional representation of 7700 fluorescence-excitation spectra from the B800 band of an individual LH2 complex. The spectra have been recorded consecutively at a scan speed of 50 cm^{-1}/s, an excitation intensity of 15 W/cm^2, and a temperature of 1.4 K. Between two successive scans, the polarization of the excitation light has been rotated by 1.8°. The spectral accuracy, reproducibility, and resolution amounts to 1 cm^{-1}. The horizontal axis corresponds to the laser-excitation energy, the vertical axis to the polarization and scan number, respectively, and the fluorescence intensity is given by the gray scale. The total experiment covered 35 h distributed over four days. The arrows indicate interruptions of the experiment to refill the cryostat. (b) Top trace: Average of all 7700 individual fluorescence-excitation spectra. Center traces: Fluorescence-excitation spectra corresponding to the single scans 2040, 2041, and 2042, respectively. Bottom trace: Average of the individual scans from 2032–2042. For better comparison all traces have been normalized.

However, from spectral hole burning it is known that spectral dynamics in those systems occurs on time scales longer than about 100–1000 ms at low temperatures [17]. Moreover, we could show by single-molecule spectroscopy that spectral shifts on the order of a few wave numbers take place on a 10 s time scale [12]. Therefore it is reasonable to assume that significant spectral changes of the profile of the electronic spectrum are not observed during the scan across the ZPL ($\leq$ 150 ms) or the PSB ($\approx$ 1 s), i.e., $\delta t_{\mathrm{exp}} \ll \delta t_{\mathrm{fluc}}$, and that those fluctuations occur only on the time scale of the full sweep ($\delta t_{\mathrm{fluc}} \ll \delta T_{\mathrm{exp}}$). Consequently, an improvement of the signal-to-noise ratio of the fluorescence-excitation spectra can be achieved if those individual scans that yield similar spectra are determined and averaged separately.

In order to extract this information from the stack of spectra we employed a multivariate statistical analysis (MSA) pattern recognition approach. Such algorithms have been used for the comparison of amino acid sequences of proteins from different species or for the reconstruction of the three-dimensional structure of large biological macromolecules from two-dimensional projections obtained by cryoelectron microscopy [18]. For this type of analysis we consider the raw spectra of individual sweeps as a linear combination of their n data points (here $n = 1552$). The data points span an n-dimensional vector space and each single-scan spectrum can be represented by a point in this space. This data cloud of 7700 points features certain directions in which it exhibits a pronounced elongation. These are the directions in which the individual spectra differ most from each other. The basic idea of the MSA approach is first to find these main directions within the data set (i.e., an eigenvector-eigenvalue or principal components analysis) and to describe the raw spectra of the individual sweeps as linear combinations of the 50–100 most significant eigenvectors—in this case ''eigenspectra''—of the set. (It should be noted that the eigenvectors do not need to have any resemblance with the real spectra as they simply form an orthogonal basis for the data space. This reduction of dimensionality decreases the amount of data significantly, thereby facilitating its interpretation.)

The next step of the MSA analysis is the pattern recognition part, i.e., finding similar looking spectra. The goal of this classification is to group close-lying elements into compact classes. Such a so-called optimal partition can always be defined for a given number of classes; it is the partition in which the interclass variance (between classes) is maximal and the total intraclass variance (within classes) is minimal (in a mathematical least-squares sense). For this, an automatic hierarchical classification algorithm

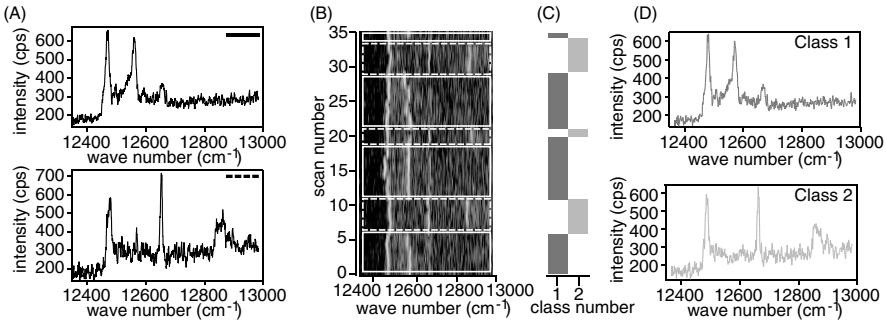

FIG. 2. (a) Application of multivariate statistical analysis to single-molecule spectroscopic data. Fluorescence-excitation spectra that result when the individual sweeps within the solid (top) and dashed (bottom) time windows boxed in (b) are averaged, respectively. (b) Two-dimensional representation of 35 fluorescence-excitation spectra from the B800 band of an individual LH2 complex. The spectra have been recorded consecutively at a scan speed of 50 cm^{-1}/s at an excitation intensity of 60 W/cm^2. The horizontal axis corresponds to the laser-excitation energy and the vertical axis to the scan number. The intensity is indicated by the gray value. The solid and dashed boxes have been positioned by visual inspection and distinguish two realizations of the spectrum. (c) Assignment of the individual sweeps into two distinct classes by the MSA algorithm. (d) Averages of the sweeps assigned to class 1 (top) and class 2 (bottom), respectively.

was used in which similar spectra (classes) are merged to form larger classes until a predetermined number of classes has been reached. In simple words, the classes can be envisaged as subclouds of points in the n-dimensional vector space, whereby the data points within each subcloud should be as close together as possible, while the distance between the subclouds should be as large as possible. More details about this algorithm can be found in [18,19].

Before applying the MSA algorithm to the data shown in Fig. 1 we verified the operation of the software and performed a simple test. For some LH2 complexes we had observed that the transition energies of the B800 pigments showed temporal variations on a time scale of minutes that resulted in the observation of two distinct spectral profiles [16]. In Fig. 2(b) the stack of consecutively recorded spectra is shown for such a complex. Visual inspection then led to the conjecture that the spectrum switches between two realizations, which we have indicated by the solid and dashed boxes. This becomes more apparent upon averaging all individual sweeps which are grouped either in a solid (upper panel) or in a dashed (lower panel) box, respectively, in Fig. 2(a). We then applied the MSA algorithm to the stack of spectra and restricted the number of classes to two [Fig. 2(c)]. Comparison of the class-averaged spectra, Fig. 2(d), with those obtained by visual inspection, Fig. 2(a), shows an excellent agreement of the two classification schemes demonstrating the suitability of this algorithm for spectroscopic tasks.

Subsequently we applied the MSA-classification algorithm to the full data set presented in Fig. 1(a). Running an analysis with N classes yields N class-averaged spectra where the number of individual sweeps that contribute to a distinct class varies according to the result of the statistical analysis of the data set. Of course, the sum of all class averages is identical to the sum of all 7700 sweeps and yields the total average shown in the top trace in Fig. 1(b). In order to study the influence of the choice of the number of classes N on the resulting spectra, we ran the analysis with 30, 50, and 100 classes, respectively. The variations between the spectra from the three separate evaluations is

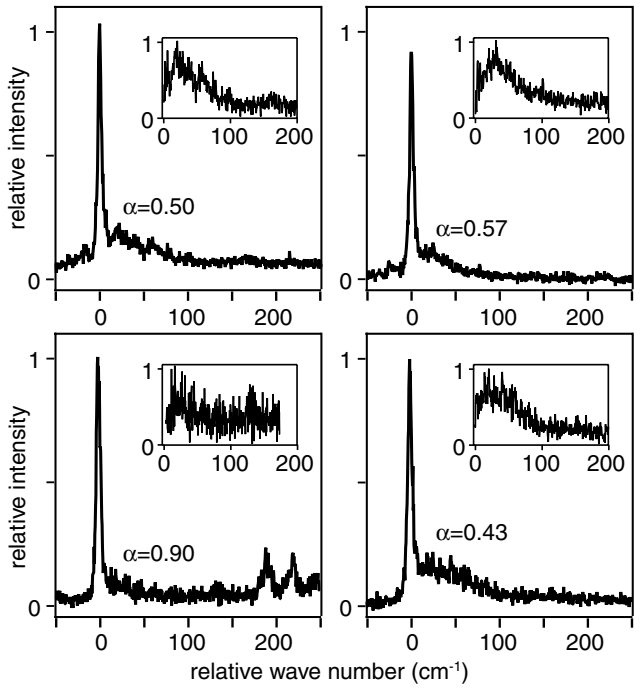

FIG. 3. Recurrent motifs in class-averaged excitation spectra. The spectra result from the data set presented in Fig. 1 for an analysis based on 50 classes. For better comparison the peak positions of the sharp spectral features have been set arbitrarily to zero and the vertical scale has been normalized. The insets show an expanded view of the broad shoulders on the high-energy side of the sharp peak after subtraction of a Lorentzian that has been fitted to the low-energy wing of the sharp line. The obtained Debye-Waller factor α is given in each panel.

TABLE I. Parameters from the spectral profiles for an analysis with 30, 50, and 100 classes, respectively. Γ_{ZPL} denotes the full width half maximum of a Lorentzian curve fitted to the ZPL, Γ_{PSB} and ω_m denote the full width half maximum and the center frequency of the PSB, respectively, and α is the Debye-Waller factor. For each parameter its average and standard deviation are given. The last two rows refer to ensemble averaged data from the literature.

	Γ_{ZPL} (cm^{-1})	Γ_{PSB} (cm^{-1})	ω_m (cm^{-1})	α
30 classes	5.6 ± 0.8	48 ± 10	31 ± 5	0.62 ± 0.12
50 classes	5.6 ± 1.4	44 ± 12	34 ± 8	0.64 ± 0.15
100 classes	5.5 ± 1.2	45 ± 15	33 ± 12	0.62 ± 0.11
[20]	5		20–30	0.74
[22]		30–40	20	0.6

not significantly different from the variations among the spectra within one evaluation. This shows that the actual choice of the number of classes is only of minor influence as long as it is chosen within a reasonable range. As could have been expected, some of the class-averaged spectra show a superposition of absorptions from individual BChl *a* molecules. However, a recurrent motif featured is a narrow peak accompanied by a weak broad shoulder on its high-energy side. Examples are shown in Fig. 3 for the evaluation with 50 classes.

We assign the sharp feature in the class-averaged spectra to the ZPL from the transition of an individual BChl *a* molecule and the broad feature to the associated PSB. In order to analyze the spectral profile in more detail, we fitted the low-energy wing of the ZPL by a Lorentzian which was subtracted from the data to uncover the PSB as shown in the insets of Fig. 3. The values obtained for the line width of the ZPL cover the range between 3–10 cm^{-1} (FWHM) and are in agreement with the results of other experiments [20,21]. The different PSBs show clear variations with respect to integrated intensity, shape, width, and center frequency. The width (FWHM) ranges from 20–80 cm^{-1}, and the center frequency is distributed between 20–60 cm^{-1}. Within the Born-Oppenheimer approximation the electron-phonon coupling can be described by the Debye-Waller factor α defined as $\alpha = \frac{I_{ZPL}}{I_{ZPL}+I_{PSB}}$ where I_{ZPL} (I_{PSB}) refers to the integrated intensity of the ZPL (PSB), respectively [7]. From our data we obtain for the Debye-Waller factor values between 0.4 and 0.9 reflecting only a weak electron-phonon coupling. We have summarized our findings in Table I together with the data obtained from the evaluations based on 30 and 100 classes which yield essentially the same result. Generally, the values obtained for the averages are in agreement with those available in the literature [20,22].

We have demonstrated that pattern recognition software can be applied successfully to single-molecule spectroscopy. A prerequisite is that spectral fluctuations occur on an intermediate time scale with respect to the temporal resolution and the duration of the experiment. Under these conditions the method opens the possibility to take full advantage of single-molecule spectroscopy in disordered host systems by avoiding spatial averaging and suppressing temporal averaging. The procedure is obviously not restricted to the model systems studied here.

We thank Cornelia Münke (Frankfurt) for her excellent technical assistance and Michael Schatz of Image Science (Berlin) for help with the IMAGIC software. This work is financially supported by the Volkswagen Foundation (Hannover), which is gratefully acknowledged.

[1] P. Tamarat, A. Maali, B. Lounis, and M. Orrit, J. Phys. Chem. A **104**, 1 (2000).
[2] W. E. Moerner, J. Phys. Chem. B **106**, 910 (2002).
[3] Y. Jung, E. Barkai, and R. J. Silbey, J. Chem. Phys. **117**, 10 980 (2002).
[4] A. Kiraz, M. Ehrl, C. Bräuchle, and A. Zumbusch, J. Chem. Phys. **118**, 10 821 (2003).
[5] K. K. Rebane, *Impurity Spectra in Solids* (Plenum Press, New York, 1970).
[6] W. P. Ambrose, T. Basché, and W. E. Moerner, J. Chem. Phys. **95**, 7150 (1991).
[7] M. Orrit, J. Bernard, and R. I. Personov, J. Phys. Chem. **97**, 10 256 (1993).
[8] W. E. Moerner, Science **265**, 46 (1994); W. E. Moerner and M. Orrit, Science **283**, 1670 (1999).
[9] W. P. Ambrose and W. E. Moerner, Nature (London) **349**, 225 (1991).
[10] A.-M. Boiron, P. Tamarat, B. Lounis, R. Brown, and M. Orrit, Chem. Phys. **247**, 119 (1999).
[11] L. Fleury, R. Brown, J. Bernard, and M. Orrit, Laser Phys. **5**, 648 (1995); R. Kettner, J. Tittel, T. Basché, and C. Bräuchle, J. Phys. Chem. **98**, 6671 (1994); M. Vacha, Y. Liu, H. Nakatsuka, and T. Tani, J. Chem. Phys. **106**, 8324 (1997).
[12] C. Hofmann, T. J. Aartsma, H. Michel, and J. Köhler, Proc. Natl. Acad. Sci. U.S.A. **100**, 15 534 (2003).
[13] J. Koepke, X. Hu, C. Muenke, K. Schulten, and H. Michel, Structure **4**, 581 (1996).
[14] C. Hofmann *et al.*, Phys. Rev. Lett. **90**, 013004 (2003).
[15] L. Germeroth, F. Lottspeich, B. Robert, and H. Michel, Biochemistry **32**, 5615 (1993).
[16] C. Hofmann, T. J. Aartsma, H. Michel, and J. Köhler, New J. Phys. **6**, 8 (2004).
[17] U. Störkel, T. M. Creemers, F. T. H. den Hartog, and S. Völker, J. Lumin. **76&77**, 327 (1998).
[18] M. van Heel *et al.*, Q. Rev. Biophys. **33**, 307 (2000).
[19] M. van Heel, Optik **82**, 114 (1989); L. Borland and M. van Heel, J. Opt. Soc. Am. A **7**, 601 (1990).
[20] H.-M. Wu *et al.*, J. Phys. Chem. **100**, 12 022 (1996).
[21] C. A. de Caro, R. W. Visschers, R. van Grondelle, and S. Völker, J. Phys. Chem. **98**, 10 584 (1994).
[22] G. J. Small, Chem. Phys. **197**, 239 (1995).