Analytica Chimica Acta 910 (2016) 1–11

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry

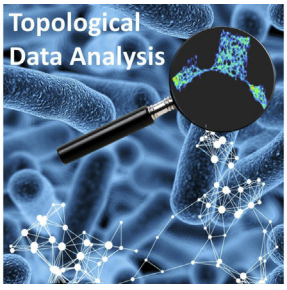
Marc Offroy, Ludovic Duponchel*

Laboratoire de Spectrochimie Infrarouge et Raman, LASIR, CNRS UMR 8516, Bât. C5, Université Lille 1, Sciences et Technologies, 59655, Villeneuve d'Ascq, Cedex, France

HIGHLIGHTS

- First use of Topological Data Analysis in spectroscopy.
- Detection of sub-populations with TDA which are not observed with PCA or HCA.
- Topological data analysis less sensitive to noise, spectral resolution and spectral shift.
- Topological data analysis is a highly scalable method (can handle very big data sets).

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 15 October 2015
Received in revised form
12 December 2015
Accepted 15 December 2015
Available online 5 January 2016

Keywords:
Topological data analysis
Big data exploration
Chemometrics
Bacteria
Raman spectroscopy
Analytical chemistry
Physical chemistry
Biology

ABSTRACT

An important feature of experimental science is that data of various kinds is being produced at an unprecedented rate. This is mainly due to the development of new instrumental concepts and experimental methodologies. It is also clear that the nature of acquired data is significantly different. Indeed in every areas of science, data take the form of always bigger tables, where all but a few of the columns (i.e. variables) turn out to be irrelevant to the questions of interest, and further that we do not necessary know which coordinates are the interesting ones. Big data in our lab of biology, analytical chemistry or physical chemistry is a future that might be closer than any of us suppose. It is in this sense that new tools have to be developed in order to explore and valorize such data sets. Topological data analysis (TDA) is one of these. It was developed recently by topologists who discovered that topological concept could be useful for data analysis. The main objective of this paper is to answer the question why topology is well suited for the analysis of big data set in many areas and even more efficient than conventional data analysis methods. Raman analysis of single bacteria should be providing a good opportunity to demonstrate the potential of TDA for the exploration of various spectroscopic data sets considering different experimental conditions (with high noise level, with/without spectral preprocessing, with wavelength shift, with different spectral resolution, with missing data).

1. Introduction

Topology, a sub-field of pure mathematics, is the mathematical study of shape. Although topologists usually study abstract objects, they have developed recently what they call Topological Data Analysis (TDA) [1]. The idea is here to use topology in order to

* Corresponding author.
E-mail address: ludovic.duponchel@univ-lille1.fr (L. Duponchel).

http://dx.doi.org/10.1016/j.aca.2015.12.037
0003-2670/

visualize and explore high dimensional and complex real-world data sets. This concept has been successfully used in different topics like gene expression profiling on breast tumors [2,3], T-cell reactivity to antigens for different type of diabetes [4], viral evolution [5], population activity in visual cortex [6] but also on unexpected topic as 22 years of voting behavior of the members of the US House of Representatives [7], characteristics of NBA basketball players via their performance statistics [7].

The two main tasks of TDA is the measurement of shape and its representation. One fundamental idea of TDA is to consider a data set to be a sample or point cloud taken from a manifold in some high-dimensional space (Fig. 1a). The sample data are used to construct simplices, generalizations of intervals, which are, in turn, glued together to form a kind of wireframe approximation of the manifold. This manifold and the wireframe represent the shape of the data. It is clear that many data analysis methods i.e.chemometric tools are available in order to explore data sets. However there are not yet ready for the analysis of future big data set which will be generated in many areas as biology, analytical chemistry or physical chemistry.

The main question is now, why topology is well suited for such data analysis? In general, TDA is considered to have three key properties. The first one is called *coordinate invariance*. Topology studies shapes in a coordinate free-way. Indeed topological constructions do not depend on the coordinate system chosen, but only on the distance function that specifies the shape. In Fig. 1b, the two A letters (constituted of millions of points) could represent a data set of samples analyzed with two different analytical platforms (different coordinate systems) while the topological construction extracts the main features of it. The second key property is *deformation invariance*. Topological properties are unchanged when a geometric shape is stretched or deformed. In Fig. 1c, the letter A deform, but the key features, the two legs and the closed triangle remain what are retrieved in the topological representation. It is because our brain works in a topological way that one can recognize A letters regardless of the font used [8]. In general, topologists consider TDA as a method which is less sensitive to noise. Indeed it possesses the ability to pick out the shape of a data set despite countless variations or deformations. The third property is *compression*. If we are willing to sacrifice a little bit of detail, a simple representation of the fundamental properties of A letter i.e. a close triangle and two legs can obtained (Fig. 1d). Considering this A letter as a big data set with millions of points, TDA can generate in this case a topological network with five nodes and five edges. Thus

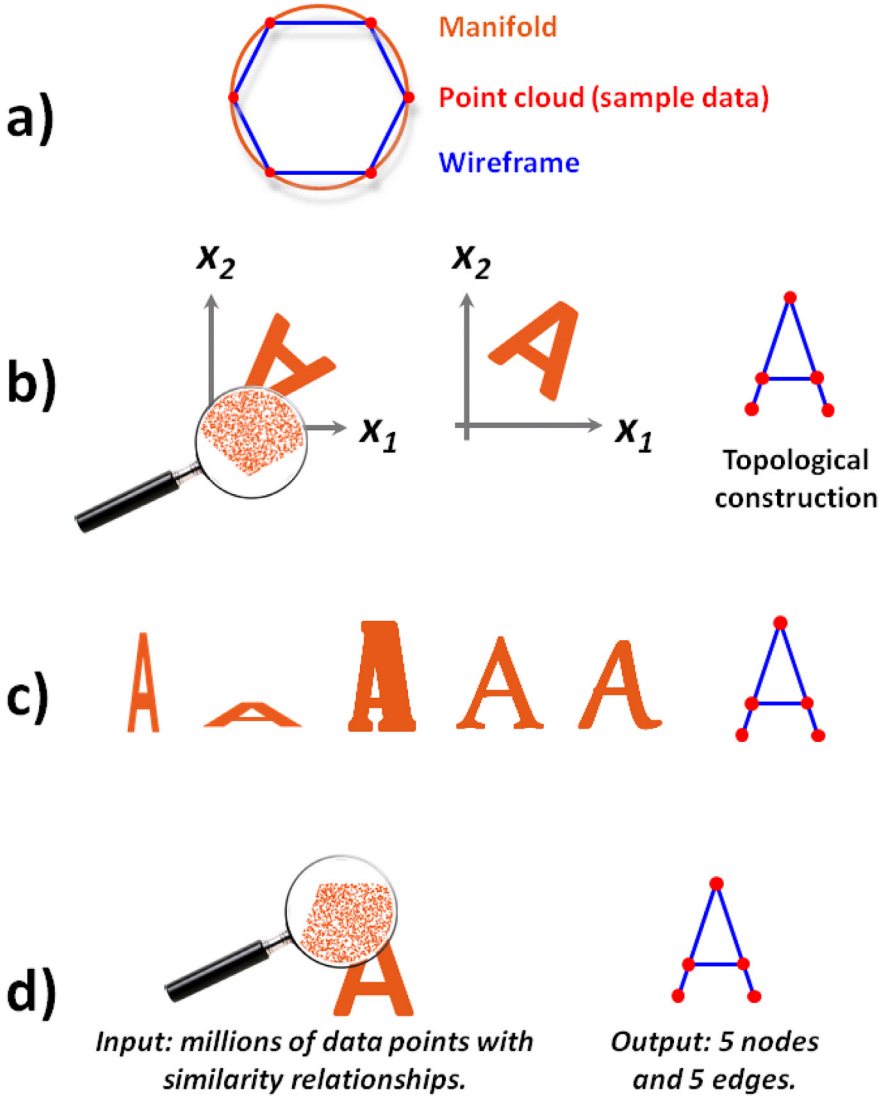


Fig. 1. (a) Fundamental idea of Topological Data Analysis. Its three key properties: (b) coordinate invariance, (c) deformation invariance and (d) compression.

this compressed representation encodes all these relationships in a very simple form. For all these reasons, this highly scalable method gives us a good opportunity to analyze very big data sets which will be generated in biology, analytical chemistry or physical chemistry. The main purpose of this work is to introduce TDA and also highlighted these nice properties which will be necessary to manage our future data structures.

2. Topological data analysis

Before going into details, a general framework of TDA will be first introduced in order to know how it can be used to generate a topological network from our data and how to interpret it. The final network will represent the shape of our data. The shape will have meaning for data exploration. TDA uses mathematical functions as lenses on data similar to using an objective of a microscope to bring focus to your sample (Fig. 2). Different lenses highlight different aspects of a data set. Due to this, networks generated with different lenses can look very different.

The first step of TDA is a partition. In fact the lens drives the division of data points into overlapping bins i.e. sub-populations. In a second step, this partition is analyzed. Within these bins data are clustered such that a cluster contains rows that are similar to each other. From a spectroscopic point of view a row can be a spectrum characterizing a sample. Because the data set is divided into bins in an overlapping way, each row is oversampled and falls into more than one cluster. In a final step data are reassembled in order to generate the final network. Then if two clusters in different bins share one or more rows, an edge is used between the two clusters generating the final network.

In order to really understand how TDA works, it is proposed to analyze a very simple synthetic data set containing 7300 samples i.e. rows defined by only two variables x_1 and x_2. Moreover this example will allow us to understand how TDA can be used on a very high amount of variables usually observed in spectroscopy. This 2-variables example is selected here just to assist in understanding the TDA concept. Fig. 3a represents a scatterplot of the data set. It is possible to observe four groups with varying density and a significant number of points in the middle that have similarity with each of these groups. The first task for TDA network construction is to choose a metric and a lens. In fact the metric is the measure of similarity or distance between any two data points while the lens is the mathematical function through which data are observed. What is probably most important to point out here is that anything producing a number from a data point can be a lens. It gives us the freedom to choose different functions which provide different point of view on the data set. As a non-exhaustive list, lenses could come from statistics (mean, max, min …), from geometry (centrality, curvature …), from chemometrics (PCA scores, SVM Distance from hyperplane, MDS scores …) and much more. A data set can even be observed simultaneously through different lenses by simply multiplying their effects. Considering the given data set, the chosen metric is Euclidean distance between any two points and the chosen lens is Gaussian density. This lens applies a row-wise Gaussian kernel estimator over the data as described in eq. (1).

$$Gaussian\ Density(x_i) = \sum_{\mathbf{x_j} \in\ data\ set} e^{-d^2(\mathbf{x_i},\mathbf{x_j})} \quad (1)$$

Where $\mathbf{x}_i$ is a row vector defining sample i for which a lens value is calculated, $\mathbf{x}_j$ are all the other samples in the data set and d the Euclidian distance between any two points. In Fig. 3b, original data are now colored by the lens value (i.e. Gaussian density in this case) where red points have the highest density and dark blue points have the lowest one. In the next step, the whole lens range (i.e. the colorbar) is divided up into overlapping subset. All the data points that have lens values within each subset range are grouped together in a bin (represented by white boxes in Fig. 3c). In this way, a partition is obtained from all the data points based on the lens value. Because lens value subsets overlapping, it is possible for data points to exist in multiple bins. In a third step we consider each bin separately. Given a bin, theses points are clustered based on the metric initially selected. A conventional cluster algorithm as single linkage [9] is usually used but other techniques can be implemented for this task. For example, if all data points in a bin are very similar under the Euclidian metric, thus these points form a single cluster represented as a single node in the final network (Fig. 3d). In general the default node color is the average lens value for all the

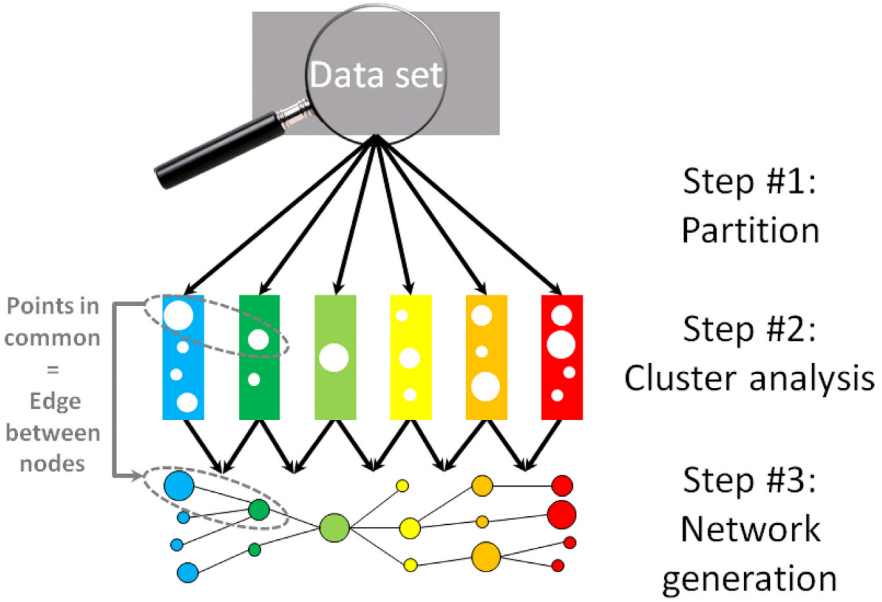


Fig. 2. General framework of Topological Data Analysis.

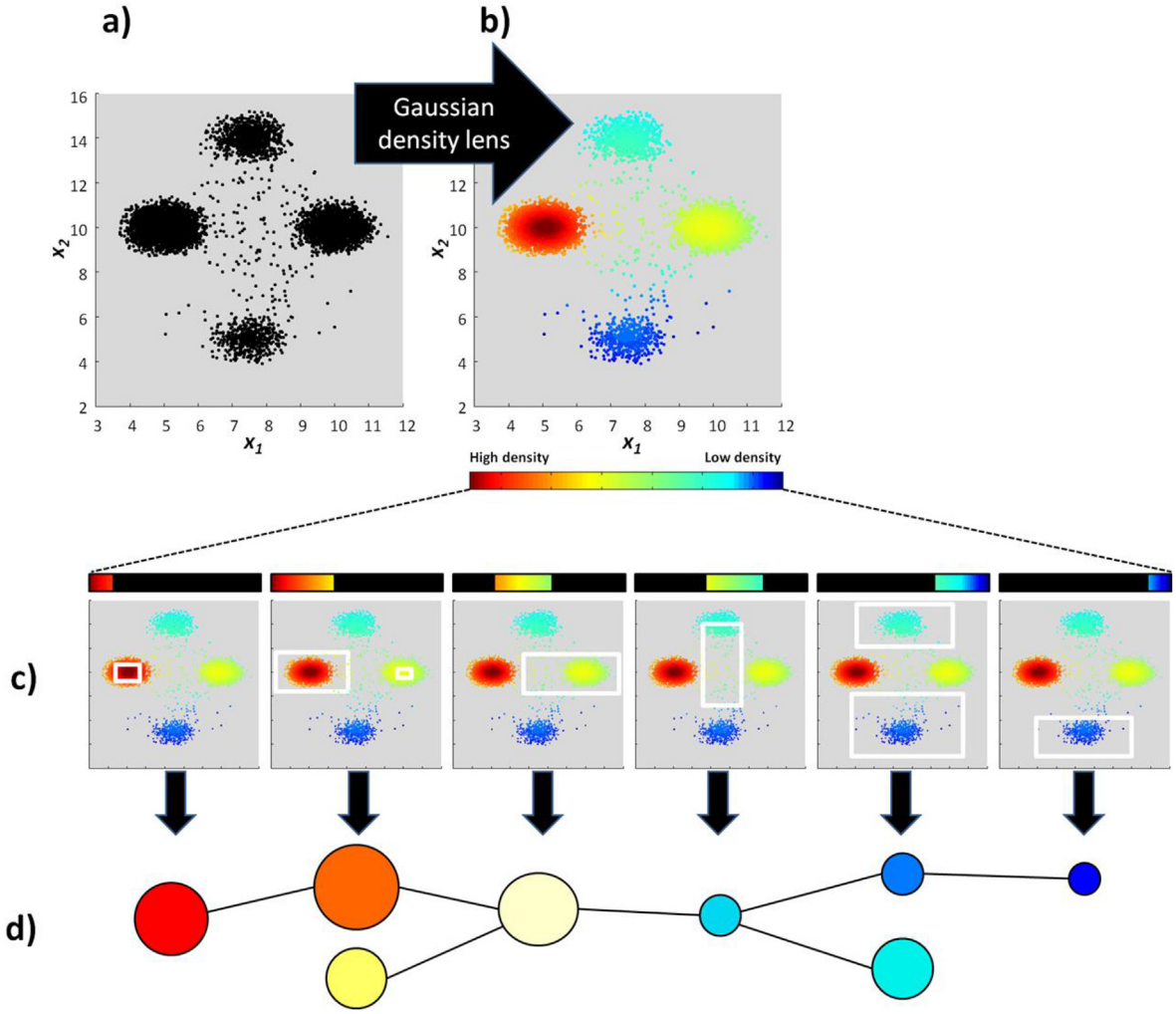


Fig. 3. TDA in action: a simple 2-variables example.

data points in the cluster while the node size is proportional to the number of data points in the cluster. Doing this task for all the bins, all the nodes are generated for the considered data set. Finally, nodes are connected when corresponding clusters have data points in common generating the topological network.

Two tuning parameters are attached to the lens. The first one is called *resolution* which is the number of bins the data is split into. Fig. 4a shows topological network of the considered data set at different resolutions. At low resolution a rather coarse view of data is obtained that could be a good way to summarize it. At high resolution, a detailed view is generated where the four groups can be observed. However the fragmentation induces the loss of similarity between the groups. In this case, singletons are also observed which correspond to nodes without edges and thus absence of common rows. They are grouped in a circular area in Fig. 4a but do not represent a cluster on their own. The second parameter is called *gain* which controls overlaps between bins (eq. (2)).

$$Percentage\ of\ overlap = (Gain - 1)/Gain \quad (2)$$

As presented in Fig. 4b, the larger the gain the larger the number of edges within your network. This parameter can thus highlights relationships within your data. However, there is no optimal value for these two tuning parameters. Similar to using different objectives of a microscope to zoom in or out on a sample, parameters are selected depending on what we want to observe from the data set. As a general observation, the highest resolution (with many details) is not always the best way to extract knowledge from a multidimensional data set. Thus, TDA can give us a multiscale analysis which is rather rare in data exploration. The next experimental part will show how Topological Data Analysis, with its three key properties, can better handle noisy data, data from different instruments or platforms and even big data.

3. Material and methods

3.1. Samples

Analysis of single bacteria is a hot topic particularly in the framework of air biomonitoring. This is largely due to the need to develop new spectroscopic instrumentations capable of detecting agents in real time for civil and military applications. In this context, four bacteria strains were prepared in this work i.e. *Staphylococcus epidermidis* (a Gram-positive bacterium), *Pseudomonas fluorescens* (a Gram-negative bacterium), *Pseudomonas syringae* and *Escherichia coli* (a Gram-negative bacterium). Bacteria were first aerosolized and deposited on a CaF_2 window for Raman analysis. This method was necessary to have a good dispersion and insure single bacteria Raman analysis described below.

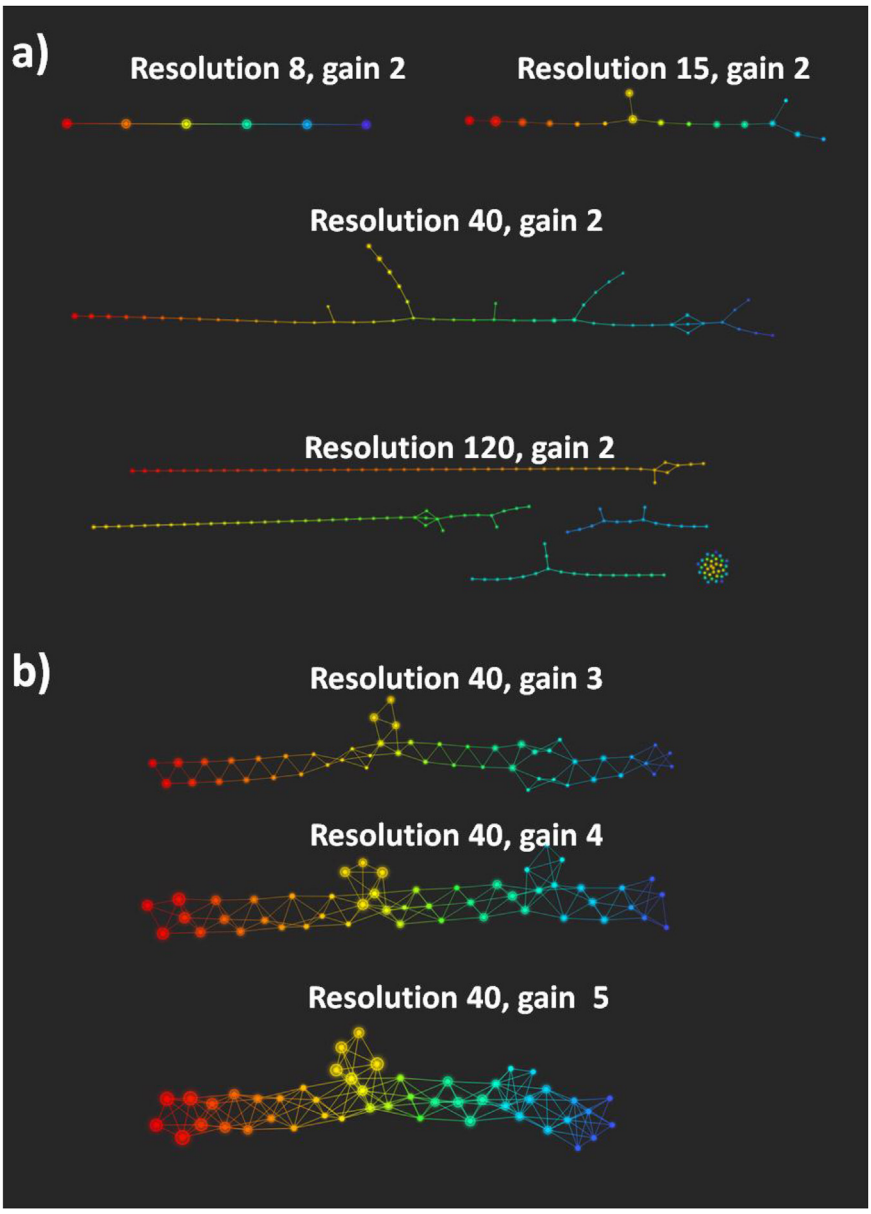


Fig. 4. TDA tuning parameters: (a) Resolution and (b) Gain.

3.2. Spectroscopy

Raman spectra of single bacteria were acquired with the LabRAM HR confocal scanning spectrometer manufactured by Horiba Jobin Yvon Scientific Company. The spectrometer is coupled confocally with an Olympus BX 40 high-stability microscope equipped with × 100 objective (NA = 0.9). This instrument is equipped with a 300 grooves/mm holographic grating that enables a spectral resolution of 2 cm^{-1}. A liquid nitrogen-cooled CCD detector is used in the spectrometer. Raman backscattering is obtained with a 633 nm excitation wavelength (10 mW) supplied by a solid-state laser. A total of 1000 single bacteria were analyzed for each strain thus generating a 4000 spectra data set on the 500–3500 cm^{-1} spectral domain. Each bacteria were analyzed with two different acquisition times (i.e. 1 min and 60 ms) in order observe two different signal to noise ratios.

3.3. Data analysis

Principal component analysis [10] and Hierarchical Cluster Analysis (Ward's method) [11] were performed using the Eigenvector PLS toolbox (Eigenvector Research Inc., Wenatchee, WA, USA) in the MATLAB environment, version 8.0 (The MathWorks, Natick, MA, USA). Satvitsky-Golay derivative [12] or SNV (Standard Normal Variate) normalization [13], when needed, were also done with the PLS toolbox. Topological Data Analysis was performed with the Ayasdi 3.0 software platform (ayasdi.com, Ayasdi Inc., Menlo Park CA).

4. Results and discussion

The main aim of this part is to observe the behavior of common data analysis tools vs TDA when exposed to different data structures induced by different experimental conditions. First when

working with spectroscopic data sets, it is almost compulsory to apply a spectral pretreatment in order to suppress artifacts or unwanted variances. Because finding a good preprocessing algorithm or a combination of several ones is not always a trivial task, it is interesting to see if raw data can be analyzed directly with TDA. Fig. 5a shows Raman spectra of the 4000 single bacteria acquired in 1 min each. An important baseline shift is observed due to fluorescence. PCA analysis of the data set is shown in Fig. 5b where the fours strains are retrieved with a strong overlap (*S. epidermidis* in red, *P. fluorescens* in dark green, *P. syringae* in dark blue and *Escherichia coli* in light blue). As observed PCA is unable to retrieve four distinct clusters. It is mainly due to the fact that directions of main variances (i.e. fluorescence) are not correlated to strain type for this data set. Moreover this overlap is observed whatever the selected PCA hyperplane. Unfortunately chemical variance, the more interesting information, is really smaller than fluorescence in this case. HCA results in Fig. 5c reveal the same trend. Fig. 5d were generated by performing Topological Data Analysis on the same data set. Nodes in the network represent clusters of bacteria and edges connect nodes that contain samples (i.e. single bacteria) in common. In the first representation nodes are colored by the total number of bacteria while the four others are colored in order to indicate the presence of a particular bacteria strain. On its part, TDA extracts distinct groups for the four strains. Sub-groups or subpopulations of bacteria are even observed particularly for *S. epidermidis* and *E. coli* which are not present in PCA nor HCA. Considering now the same data set corrected by a first derivative and SNV normalization, PCA scatterplot (Fig. 6b) shows a better

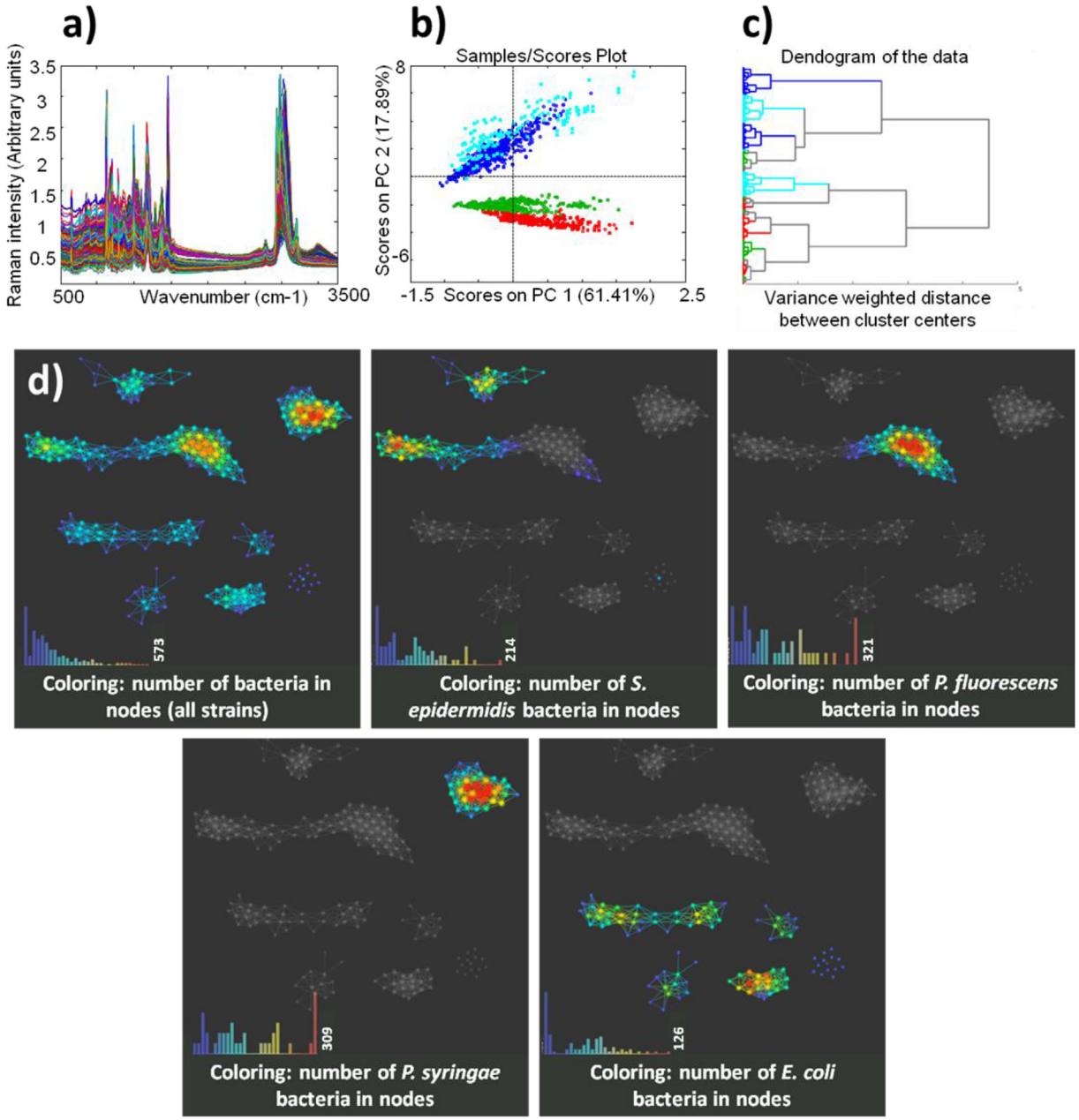


Fig. 5. Discriminating single bacteria with Raman spectroscopy. (a) Raw data, (b) PCA score plot, (c) HCA dendogram. PCA and HCA coloring: *Staphylococcus epidermidis* in red, *Pseudomonas fluorescens* in dark green, *Pseudomonas syringae* in dark blue and *Escherichia coli* in light blue. (d) TDA network (metric: Euclidian distance, Neighborhood lens, resolution = 30 and gain = 3). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

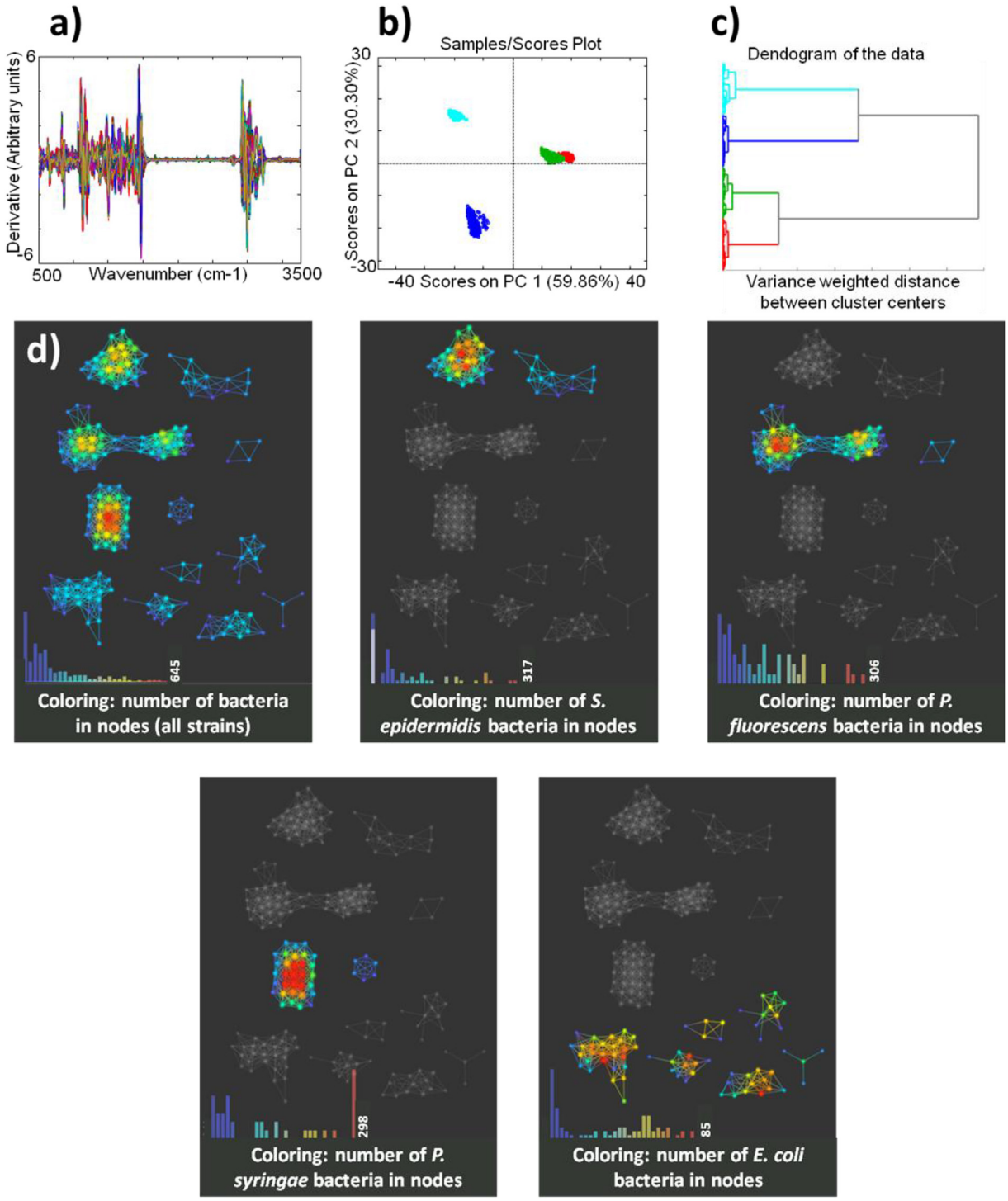


Fig. 6. Influence of spectral preprocessing effect. (a) First derivative and SNV normalized spectra, (b) PCA score plot, (c) HCA dendogram. PCA and HCA coloring: *Staphylococcus epidermidis* in red, *Pseudomonas fluorescens* in dark green, *Pseudomonas syringae* in dark blue and *Escherichia coli* in light blue. (d) TDA network network (metric: Euclidian distance, Neighborhood lens, resolution = 30 and gain = 3). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

clustering with a persistent overlap between *S. epidermidis* (in red) and *P. fluorescens* (in green). HCA result in Fig. 6c is good, but what is the more interesting is that strains and sub-populations are always observed with TDA. With this first part, it is observed that TDA can be invariant to deformation which is observed in spectroscopy when a preprocessing method is not applied. '

Another problem we often face in spectroscopy is the weakness of the signal to noise ratio. It happens when the observed physical effect (scattering, absorption …) is weak or when acquisition time is low. Thus it is proposed here to study the same bacteria with an acquisition of 60 ms instead of 1 min Fig. 7a shows preprocessed spectra (first derivative, SNV) of the 4000 single bacteria considering this new acquisition time. One can observe an extremely low signal to noise ratio. Indeed it is very difficult to retrieve spectral

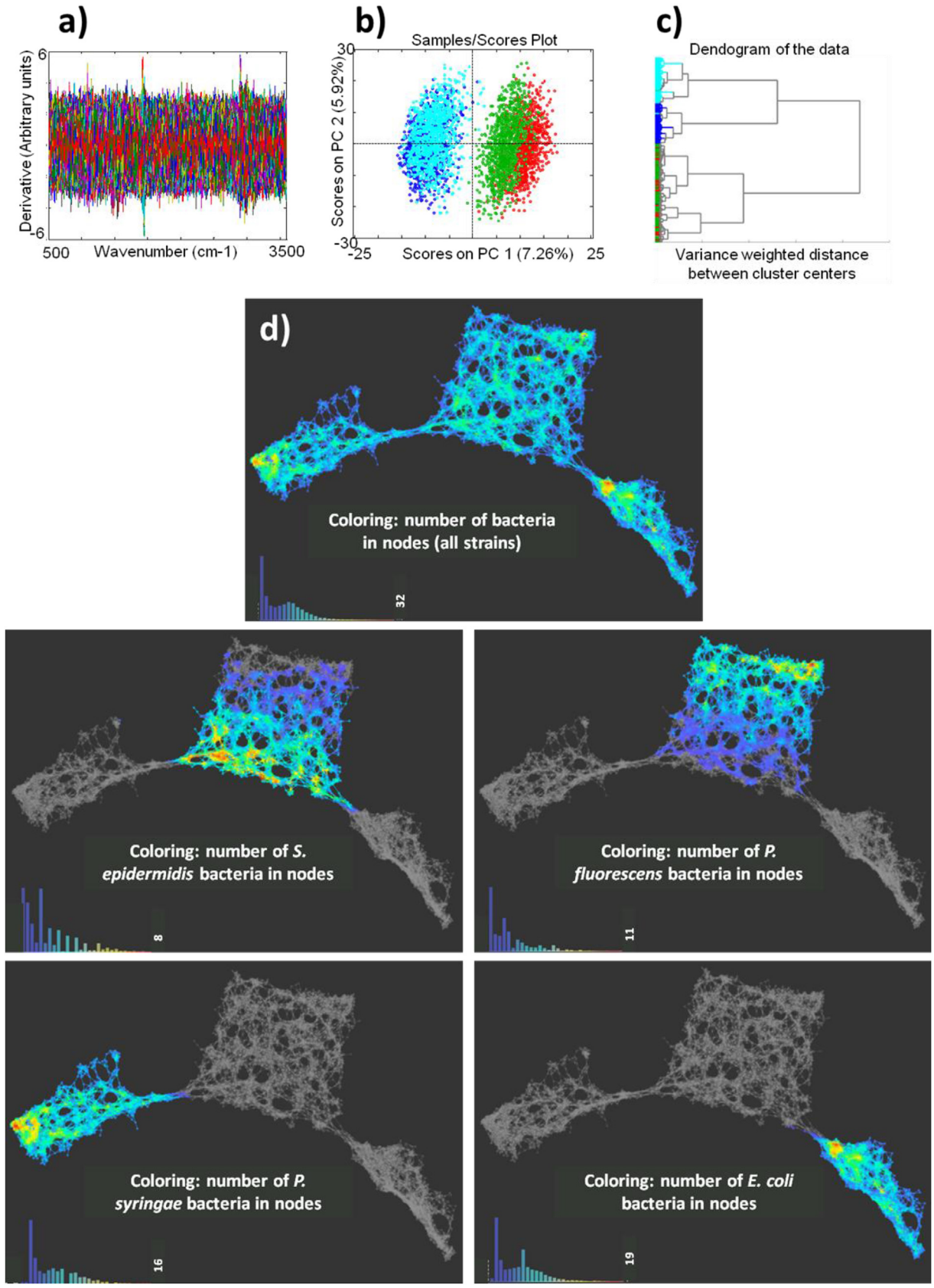


Fig. 7. Influence of noise level. (a) Noisy first derivative and SNV normalized spectra, (b) PCA score plot, (c) HCA dendogram. . PCA and HCA coloring: *Staphylococcus epidermidis* in red, *Pseudomonas fluorescens* in dark green, *Pseudomonas syringae* in dark blue and *Escherichia coli* in light blue. (d) TDA network (metric: Norm correlation, lens: MDS scores, resolution = 30 and gain = 3). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

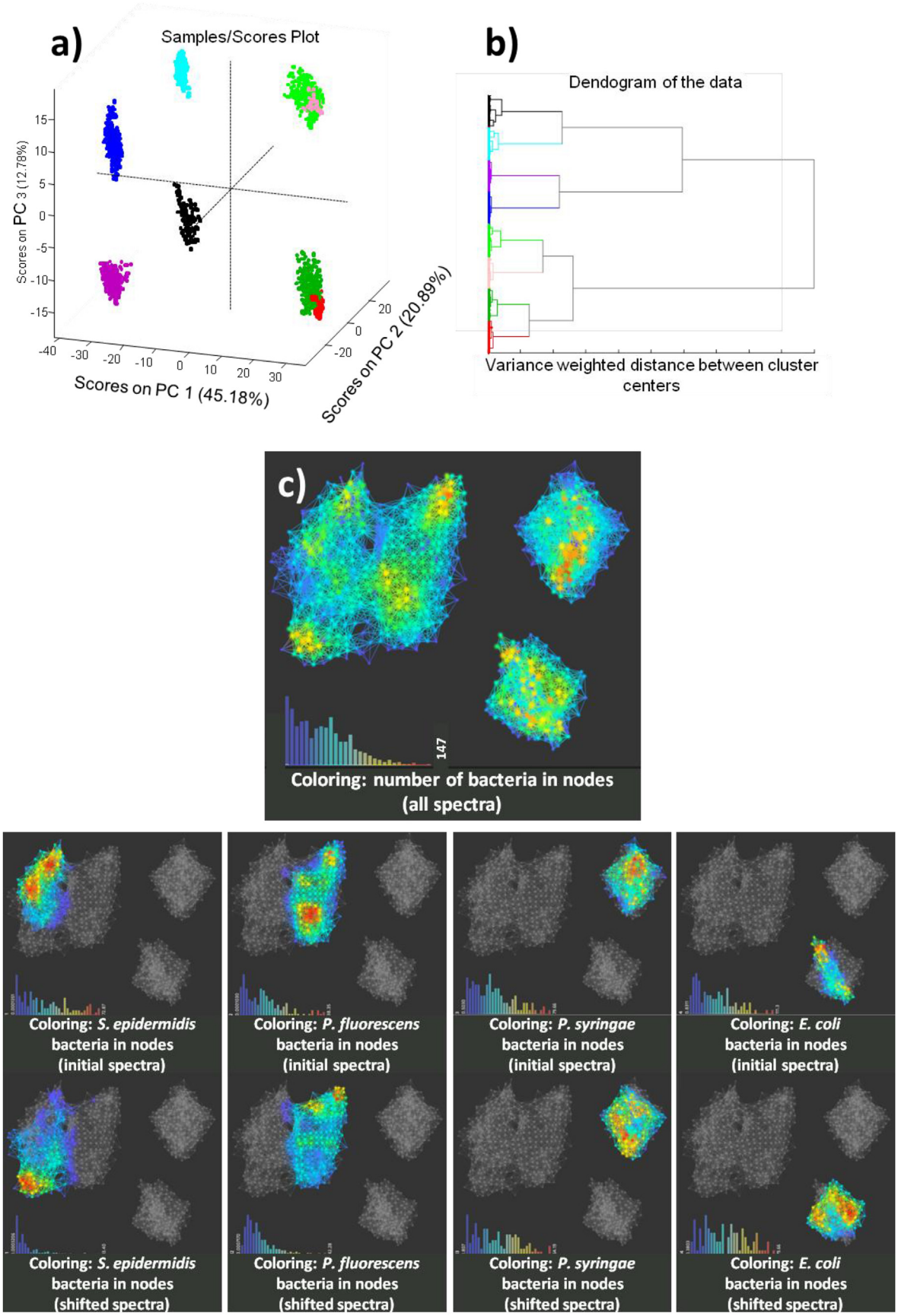


Fig. 8. Influence of spectral shift. (a) PCA score plot, (b) HCA dendogram. PCA and HCA coloring: *S. epidermidis* in red, *P. fluorescens* in dark green, *P. syringae* in dark blue, *E. coli* in light blue, "shifted" *S. epidermidis* in pink, "shifted" *P. fluorescens* in light green, "shifted" *P. syringae* in purple and "shifted" *E. coli* in black. (c) TDA network (metric: Variance normalized Euclidian distance, lens: PCA scores, resolution = 40 and gain = 3). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

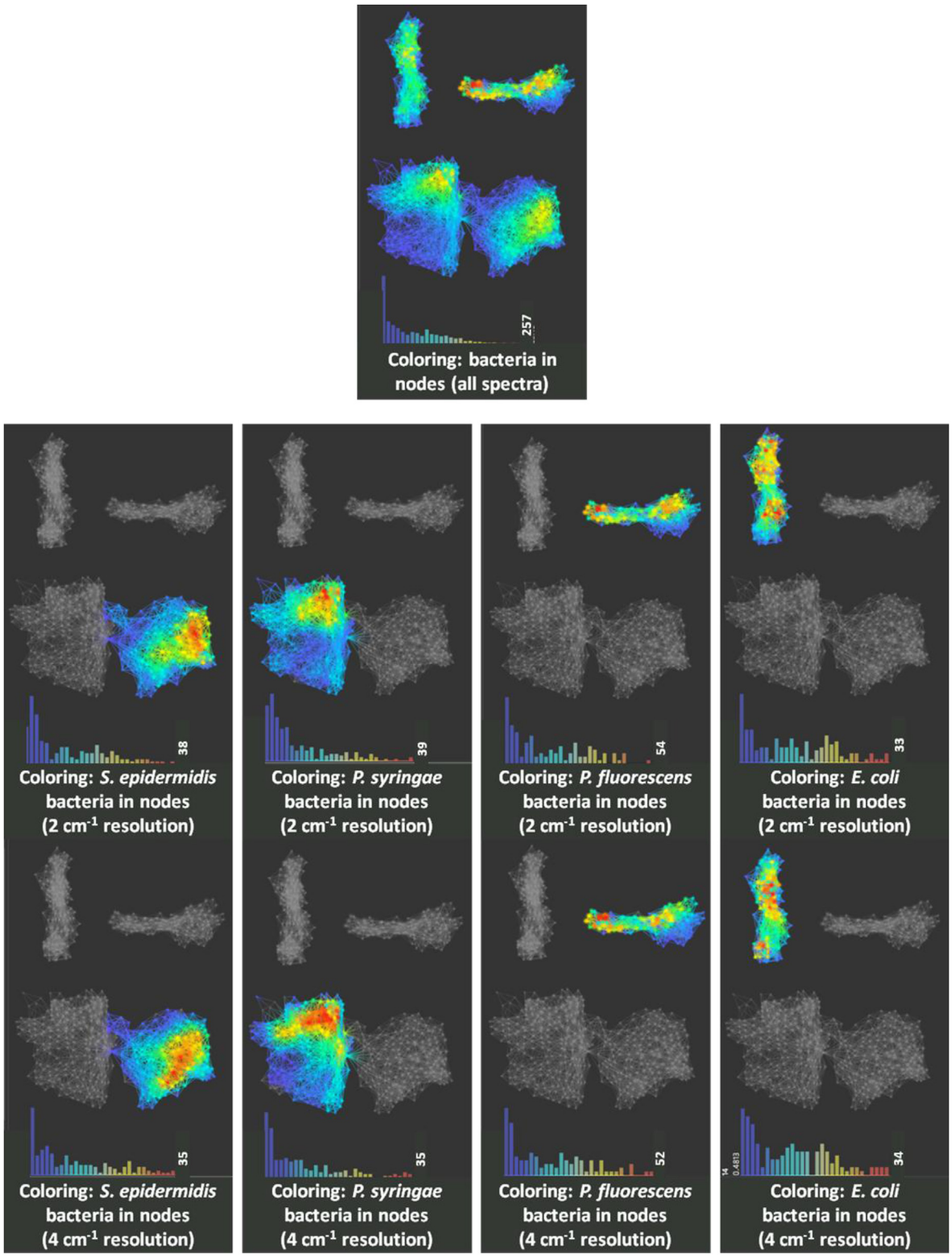


Fig. 9. Influence of spectral resolution. The TDA network with missing data (metric: Norm correlation, lens: MDS scores, resolution = 30 and gain = 3).

features previously observed on Fig. 6a. The noise level is so high in these conditions that a maximum overlap is observed of the four strains in the PCA scatterplot. Indeed the high variance coming from noise is now expressed by the first principal components. It should be noted that HCA cannot retrieve the fours strains for the same reason (Fig. 7c). Topological data analysis provides interesting results as specific parts of the network are dedicated to specific strains despite the very low signal to noise ratio (Fig. 7d). In particular, *P. syringae* and *E. coli* bacteria are strictly separated whereas they completely overlap with PCA. Topological data

analysis shows here the ability to handle noisy data which will be an important issue for the exploration of our next big data set.

From another point of view, when we look at big data sets a little closer, they often consist of merged experiments and even sometimes acquired with different methodologies or platforms. Thus the idea of the next part is to try to analyze such data structures. Because modern Raman spectrometers have very good wavelength reproducibility, a wavelength shift is artificially introduced on the data set in order to simulate two different instruments. It is important to look at this issue because it is apparent for other spectroscopic techniques or even sometimes when matrix effect is observed. Given the initial data set with 1 min acquisition time (corrected with first derivative and SNV normalization), an offset of 8 cm^{-1} is introduced (i.e. 4 spectral variables) on half of the spectra for each strain. However both the total number of spectral variables (i.e. 1501) and the spectral resolution (i.e. 2 cm^{-1}) are kept. Fig. 8a and b present PCA and HCA results respectively. One can observe the shift effect introducing new variances in the data set. Consequently, in addition to the four strains (*S. epidermidis* in red, *P. fluorescens* in dark green, *P. syringae* in dark blue and *E. coli* in light blue) new clusters corresponding to their shifted spectra are observed ("shifted" *S. epidermidis* in pink, "shifted" *P. fluorescens* in light green, "shifted" *P. syringae* in purple and "shifted" *E. coli* in black). TDA does not fall into the trap because a dedicated part of the network is observed for each strain whatever the spectral shift applied or not (Fig. 8c). This feature is particularly important because developing a spectral alignment procedure is never a trivial task. Moreover considering big data sets, it is never easy to know if such spectral shifts are present.

Because data sets can be acquired with different spectrometers with potentially different spectral resolutions, it is interesting in the last part of this work to see if TDA is able to manage this situation for a simultaneous exploration of all spectra. In order to reproduce this conditions, given the initial data set with 1 min acquisition time (corrected with first derivative and SNV normalization), one in two wavenumbers have been deleted on half of the spectra for each strain. In this way, it is possible to observe 500 bacteria for each strain with two different spectral resolutions i.e. 2 cm^{-1} and 4 cm^{-1}. However the total number of spectral variables (i.e. 1501) is kept leading to a missing data structure on half of the spectra. Fig. 9 shows impressive results since four specific parts of the TDA network are dedicated to the different strains whatever the spectral resolution. However selected metric and lens must be able to manage such missing data structure. In this example norm correlation and MDS scores were respectively used. They handle null values by first projecting the pair of rows to the intersection of their non-null columns. This TDA feature is very important because it is often difficult and even sometimes impossible to manage missing values with conventional data analysis tools. Nevertheless, even if TDA calculations are rather fast, we must spend time in finding a good metric/lens combination for the observation of clusters or subpopulations.

5. Conclusion

The main objective of the work was to introduce the new concept of topological data analysis with a comparison to conventional chemometric tools. This allowed us to highlight nice properties of the method. Indeed TDA was able retrieve valuable information from different data structures with very low signal to noise ratio, variable shifts and missing data. As a consequence, it might be regarded as a very robust and promising method to cope with such situations. From a general point of view, it is difficult to say that this 4000 spectra example is huge for the big data community but for us and at the present time it can be. However we are convinced that Topological Data Analysis which is very scalable will be one of the best exploration tools for our future big data set with several hundred thousand or even millions of rows (not only spectra). Deep learning in biology, analytical chemistry and physical chemistry is not so far.

Acknowledgment

L.D is grateful for the scientific and technical support from Devi Ramanan and Alan Lehman at Ayasdi Inc., Menlo Park CA.

References

[1] G. Carlsson, Bull. Amer. Math. Soc. 46 (2009) 255–308.
[2] Y. Yao, J. Sun, X. Huang, G.R. Bowman, G. Singh, M. Lesnick, L.J. Guibas, V.S. Pande, G. Carlsson, J. Chem. Phys. 130 (14) (2009) 144115.
[3] M. Nicolau, A.J. Levine, G. Carlsson, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 7265–7270.
[4] G. Sarikonda, J. Pettus, S. Phatak, S. Sachithanantham, J.F. Miller, J.D. Wesley, E. Cadag, J. Chae, L. Ganesan, R. Mallios, S. Edelman, B. Peters, M. Von Herrath, J. Autoimmun. 50 (2014) 77–82.
[5] J.M. Chan, G. Carlsson, R. Rabadan, Proc. Natl. Acad. Sci. U. S. A. 110 (46) (2013) 18566–18571.
[6] G. Singh, F. Memoli, T. Ishkhanov, G. Sapiro, G. Carlsson, D.L. Ringach, J. Vis. 8 (8) (2008) 1–18.
[7] P.Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, G. Carlsson, Sci. Rep. 3 (2012) 1236.
[8] L. Shen, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 6884–6889.
[9] R. Sibson, Comput. J. 16 (1) (1972) 30–34.
[10] K. Pearson, Philos. Mag. 2 (11) (1901) 559–572.
[11] J.H. Ward Jr., J. Am. Stat. Assoc. 58 (1963) 236–244.
[12] A. Savitzky, M.J.E. Golay, Anal. Chem. 8 (36) (1964) 1627–1639.
[13] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (1989) 772–777.