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

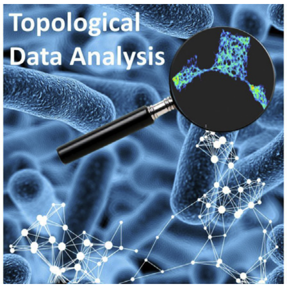
Marc Offroy, Ludovic Duponchel*

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/© 2016 Elsevier B.V. All rights reserved.

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

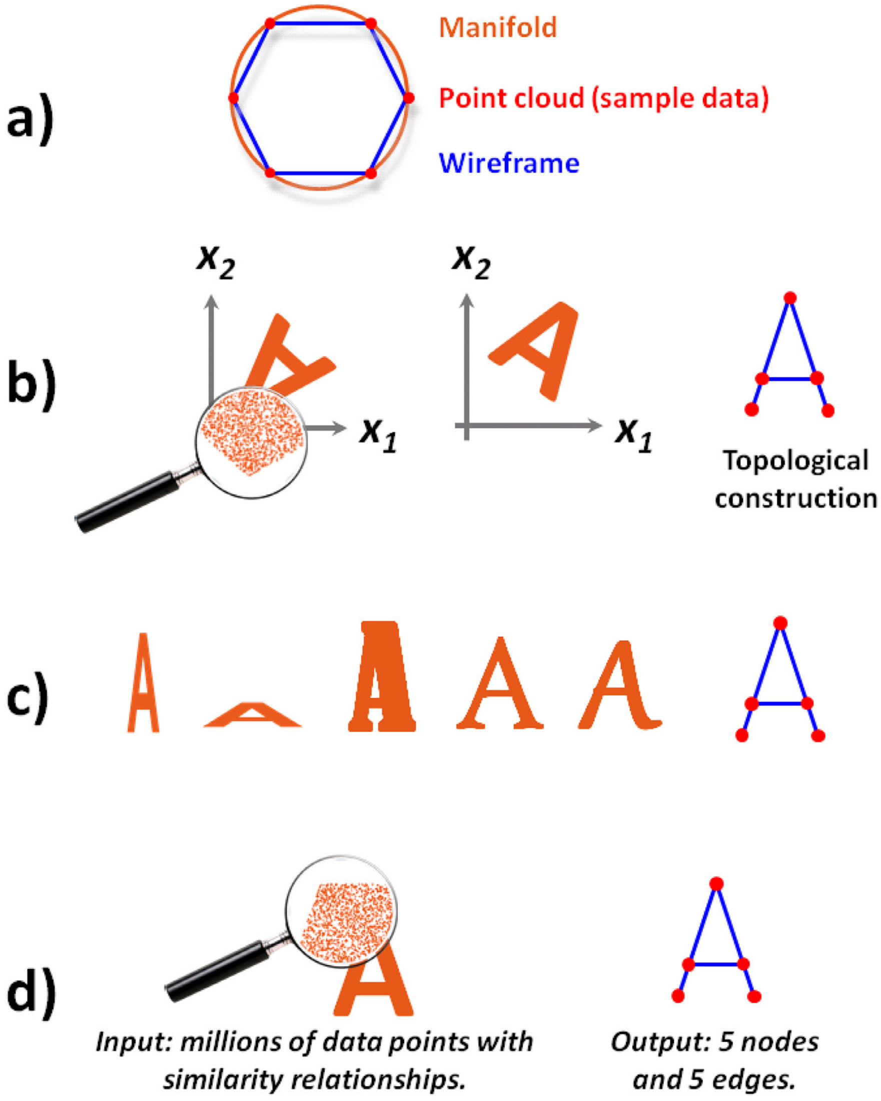

a)
Manifold
Point cloud (sample data)
Wireframe
b)
x_2
x_1
x_2
x_1
Topological construction
c)
d)
Input: millions of data points with similarity relationships.
Output: 5 nodes and 5 edges.

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