Analytical chemistry: from data (pre)-treatment optimization to data mining, data fusion and big data.

Ali Tfayli and Sana Tfaili

Presentations will focus on the researches in analytical chemistry field and our interest in big data; mainly lipids research. At first, examples of analytical applications will be developed. We will try through the exposition of each example to illustrate the nature of the registered data; to describe the steps of different data pretreatment and data treatment.

Then, we will highlight some perspectives:

- using other approaches to explore data,
- possibility of data fusion (data from complementary analysis),
- structuring data storage (database): a database from which data could be used by different programs, by different users, and that will have the major advantage of giving the opportunity to be accessed by multiple users simultaneously,
- improving the power of calculation: possibility to acquire cores and this would be possible via a project under construction,
- development of specific statistical packages.

All these points will be detailed through a discussion at the end of the presentations.

"GCAPS - EA4041", Groupe de Chimie Analytique de Paris Sud, has a primary objective to promote basic and methodological researches in the field of lipids. Our studies focus on the development of analysis tools and on the data processing in the field of lipids in particular. GCAPS main research areas are detailed in page 32 of the CDS project, https://www.lri.fr/~kegl/isd/AAP_Recherche_Idex_2014_Data_Science.pdf

The tools we develop include: separation techniques , mainly chromatographic , with particular attention to the study of stationary phases and detection systems; coupled mass spectrometry techniques (LC , GC and GCxGC/MS); vibrational spectroscopy (IR, NIR and Raman) and electronic spectroscopy (fluorescence) techniques; chemometric techniques for optimization and data processing.

Our research covers four themes: cell membranes lipidomics, lipids in skin barrier, lipids: from natural substances to heritage objects, lipid analogues for diagnostic and therapeutic aims.

Data provided by the different analytical methods could be mainly resumed by tables of intensity versus retention time, intensity versus the mass-to-charge ratio or intensity versus frequency. The profile of a sample obtained by LC/MS contains the relative distribution of the species and the molecular ion for each. Spectroscopic data are represented by point to point spectra, Z-profile spectra or hyperspectral images. Access to databases and further analysis permit to formally identify the compounds of interest. Data analysis requests the use of multivariate statistics and chemometrics. The big quantity of data has to be considered (usually the magnitude of several thousands variables). For this, the development of multiparametric approaches, of algorithms and the use of multivariate statistical analyzes is necessary.

The nature of information provided by chromatographic and spectroscopic techniques are different, the treated subjects and themes too. The first (chromatographic) is destructive and provides structural information, the sample has to be in solution; and the second (spectroscopic) is non-destructive and gives information on the systems organization and their environment. Both analytical techniques are complementary and are used herein for lipid studies. Another important feature is that vibrational spectroscopic (Raman and IR) data are multidimensional in space (x,y,z) and each point of such 3D matrix contains an intensity versus wavenumber spectrum. For the LC/MS the current trend is to associate orthogonal separation techniques before the sample enters the MS interface. The mass spectrometer itself is able to perform simultaneously MS, MS^2 and MS^3 fragmentation of ions. High dimensionalities are then also encountered with this technique. To date, statistical approaches were the key to manage data and to build results for each analytical tool and subject. However, the quantity of data will continue to grow faster, making some latent, potential information not really extracted from the registered data.