On the path forward to molecular systematics of non-living organic matter

Irina V. PERMINOVA
[a] Department of Chemistry, Lomonosov Moscow State University, Leninskie Gory 1-3, 119991 Moscow, Russia
* Tel. & Fax. No. +7-495-939-5546; E-mail: iperm@org.chem.msu.ru
Keywords Non-living organic matter, humic substances; Fourier transform mass spectrometry, molecular space, classification, systematics

Abstract A method for quantitative treatment of Van Krevelen diagrams plotted from the data of Fourier transform mass spectrometry is proposed, which enables generation of numeric descriptors of chemical space of non-living organic matter (e.g., humic substances). The method implies a use of cell-based partitioning technique for quantifying chemical space of complex materials occupied in Van Krevelen diagrams. Its application for data treatment of a broad variety of non-living organic matter from different sources lays ground for deriving comprehensive data set on molecular compositions of HS and other complex organic matrices. Such a data base enables classification analysis of humic substances by source and fractional composition. Its suitability for developing molecular systematics of non-living organic matter is discussed.

Introduction

High resolution Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FTICR MS) opened a new era in molecular understanding of complex non-living organic matter which occurs throughout the entire Earth's environment in the form of natural organic matter (NOM) and humic substances (HS) (Hertkorn et al. 2007). FTICR mass spectra of NOM and HS show several thousands of resolved molecular peaks in one sample. To translate exact molecular masses into molecular formulas, a linear Diophantine equation for a given mass and a given set of atoms (e.g., C, H, N, O, S) have to be solved (Koch et al. 2007). Resulting data are usually plotted as Van Krevelen diagrams which project elemental ratios (e.g., H/C, O/C, N/C) determined from calculated molecular formulas along two or three axes (Kim et al. 2003). This allows for translation of initial data sets into meaningful pictorial images. Still, it is not suitable for quantitative comparison and classification analysis of the samples of non-living organic matter.

In this work we propose application of cell-based partitioning for generating molecular descriptors which enable quantitative description of chemical space of NOM and HS samples and demonstrate its suitability for classification and cluster analysis.

Materials and methods

Mass lists for the humic materials studied were obtained using FTICR MS facilities located at the Institute of Biochemical Physics of RAS, Moscow, Russia. For simplicity of presentation, the molecular formulas were assigned on the basis of CHO compounds only. The obtained formulas were used for plotting the corresponding Van Krevelen diagrams.

CACTUS database (cactus.nci.nih.gov/download/nci/) was used for data mining of the structural analogues of humic substances.

Discriminant analysis was used for the spectral data classification. It was conducted in forward stepwise mode. The discriminant analysis was conducted with a use of "Statistica" software (StatSoft). To assess classification ability of classification functions calculated, cross-validation procedure was conducted.

Results and Discussion

Figure 1 shows how Van Krevelen plot generated from FTICR MS data on any HS or NOM sample can be converted into a numerical set of 20 parameters. A number of parameters can be other than twenty – it depends on the grid which will be used for cell-partitioning of the Van-Krevelen diagram space.

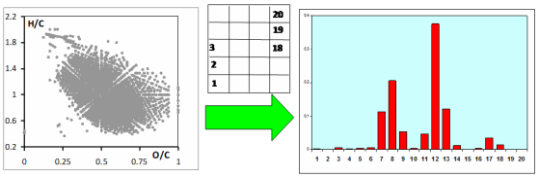

Fig. 1 Translation of the pictorial image of molecular space of HS sample - Van Krevelen diagram – into a set of twenty numerical parameters using cell-based partitioning.

To perform this translation, the area of Van Krevelen diagram was binned into n cells (we have set n equal to 20, but different values can be used) and the cell-based distribution of experimental points was calculated by quantifying simple (or intensity-weighted) population density of each cell (D_i) using equation below:

$$D_i = \frac{N_i}{N}, i = 1,2,\ldots n$$

where D_i is the population density of the i^{th} cell (i = 1 – 20); N_i is the number of data points (molecular formulas) within the i^{th} cell, N is the total number of data points plotted in Van Krevelen diagram.

The proposed descriptors can be used to describe molecular composition of humic ensemble in terms of weighted amounts of main chemical classes, which are considered as major precursors of HS. This information can be revealed by demonstrating the relationship between each descriptor (Di) and the type of structures it is characteristic of.

For this purpose, we have constructed synthetic VK diagram by plotting CHO compositions of the individual organic molecules representing major precursors of HS and NOM precursors including lipids (fatty acids), terpenoids, lignin-derived phenolisopropanoids, condensed tannins (flavonoids),

proteins, carbohydrates, and hydrolysable tannins. These classes were consistent with the proposed by Kujawinski and Behn,9 and by Hockaday et al.10, except for hydrolysable tannins which were missing in the both studies. We have assigned to hydrolysable tannins an extended region with O/C from 0.5 to 1, and H/C from 0.2 to 1.4 based on the trivial structures published for hydrolysable tannins20,21. The obtained model Van Krevelen diagram is shown in Figure 2. To apply our approach, it was binned into the same 20 cells as in case of humic samples.

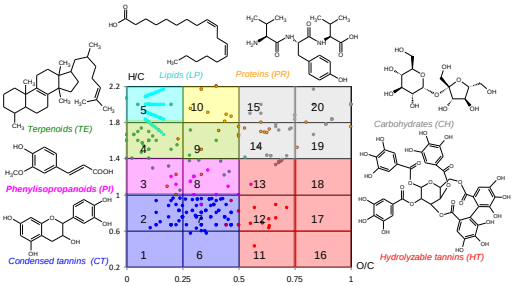

Fig. 2 The model van Krevelen diagram plotted from the CHO compositions of the selected organic compounds collected in the CACTUS database belonging to the major classes of chemical precursors of HS and NOM.

Based on the partitioning of the selected compounds over the cells, the following assignments of predominant chemical compartments were made to the descriptors proposed in this study:

Condensed tannins [CT]	Phenyliso-propanoids [PI]	Terpenoids [TE]	Lipids [LP]
D1, D2, D6, D7	D3, D8	D4, D9	D5
Proteins [PR]	Hydrolyzable tannins [HT]	Carbohydrates [CH]	
D10	D11-13, D16-18	D14-15, D19-20	

Given supramolecular nature of molecular ensemble of HS (Nebbioso and Piccolo 2011), the proposed assignments might allow for representing composition of the humic material as a combination of the selected structural units. A partial contribution of each structural unit into the specific ensemble in this case will be numerically equal to a value of the corresponding D_i calculated as a population density of the assigned cell in the VK diagram. For example, for aquatic materials – IHSS standards from the Suwannee River, composition of their supramolecular ensembles can be written as:

SRFA = 0.595[HT] + 0.212[PI] + 0.119[CT] + 0.055[TE] + 0.012[CH] + 0.004[LP] + 0.003[PR]

SRHA = 0.545[HT]+ 0.238[CT] + 0.169[PI] + 0.022[TE] + 0.016[LP] + 0.006[PR] + 0.004[CH]

It can be seen that the humic acid fraction (SRHA) is characterized with twofold content of condensed tannins which provides for its more hydrophobic nature as opposed to SRFA.

For classification analysis, we have used 37 NOM and HS samples from five different sources. All materials were analyzed using ESI FTICR MS. We also included samples of mumiyo as representatives of other than HS classes of non-living organic matter. Twenty numerical descriptors were calculated for each sample studied as described above. Classification according to source of organic matter yielded 100% of true classifications. For fractional composition it varied from 86% (for HA) to 89-90% (for FA and non-fractionated NOM samples, correspondingly).

Conclusions

The simple approach is proposed to quantitative description of chemical space occupied by different HS and NOM samples. Assignment of each numerical descriptor with the dominating chemical moieties allows representation of a supramolecular ensemble as a combination of the selected set of structural compartments. The good prospects of using these descriptors for classification analysis are demonstrated. They can be also used for quantitative structure activity relationship (QSAR) analysis. The proposed approach provides a long-sought tool for efficient data reduction technique, which is a prerequisite for deriving comprehensive database on molecular constituents of non-living organic matter. This database may lay grounds for deriving molecular systematics of non-living organic matter.

Acknowledgements.

This study was partially supported by the Russian Foundation for Basic Research (grant 16-04-01753A) and International Union of Pure and Applied Chemistry (project #2016-015-1).

References

Hertkorn N, Ruecker C, Meringer M, Gugisch R, Frommberger M, Perdue EM, Witt M and Schmitt-Kopplin P 2007 High-precision frequency measurements: indispensable tools at the core of the molecular-level analysis of complex systems. Anal Bioanal Chem. 389(5), 1311-1327.

Kim S, Kramer RW, Hatcher PG 2003 Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the van Krevelen diagram. Anal. Chem. 75, 5336–5344.

Koch BP, Dittmar T, Witt M, Kattner G 2007 Fundamentals of molecular formula assignment to ultrahigh resolution mass data of natural organic matter. Anal. Chem. 79, 1758–1763

Nebbioso A, Piccolo A 2011 Basis of a humeomics science: chemical fractionation and molecular characterization of humic biosuprastructures. Biomacromolecules 12, 1187–1199.