CHEMINFORMATICS METHODS FOR RAPID PREDICTION OF PHYSICOCHEMICAL PROPERTIES

MONICA MIRANDA RODRIGUES

New University of Lisbon Portugal
E-mail: monikimica@gmail.com

Abstract— In this project was developed cheminformatics methods based on machine learning for a quick preview correlation of energy with the size of organic compounds, the enthalpy of combustion of hydrocarbons and statistical separation of compounds. The data set of organic substances calculated by quantum chemistry was built. However, despite the theoretical chemical calculations based on Density Functional Theory (DFT) to provide estimates of various properties with increasing accuracy, its computational cost is relatively high for many situations. The correlation with the enthalpy of combustion allows the calculations of engineering and unit operations are faster and more accurate. In this work the correlation obtained was R = 1 for combustion enthalpy. Application of Principal Component Analysis and Dendrogram were used to separate alcohol of hydrocarbons by recognizing patterns using vibrational properties calculated by DFT, being a useful tool for physical-chemical analysis.

Keywords— Density Function Theory, Enthalpy, Organic Compound, Principal Component Analysis, Dedrogram

I. INTRODUCTION

The cheminformatics methods is an interdisciplinary field that uses the concepts of chemistry at the computer in the form of computational techniques applied in chemistry. The first challenges in cheminformatics was write molecular structure in computer form, nowadays, there are already several software programs that are possible to draw a molecule and writes it in the linear format for computations[1]. In this study the machine learning techniques and pattern recognition was used and built data set computed by Density Functional Theory (DFT) using the results about the optimization of energy from organic compounds.

The direct relationship between the physicochemical property and obtained linear regression with maximum correlation was found. The molecular weight increase with increasing physical property and direct correlation with the enthalpy of combustion. It is possible to easily calculate the enthalpy of combustion knowing the molecular weight and compound class.

Density Functional Theory (DFT) has emerged as an alternative to traditional methods of quantum chemistry, your B3LYP functional can calculate the properties of quickly and effectively with a gain of computational speed and memory space, the most known and used functional. The electronic density describes distribution of charge in a molecule.

The concepts of this theory were first formulated by Hohenberg and Kohn[2] for inhomogeneous electron gas. The theory was further developed by the researchers and then many researchers[3,4] have contributed to generalize the systems containing generates and degenerates states.

The learning of machine is an area of artificial intelligence aimed at the development of computational techniques in order to build systems capable of acquiring knowledge automatically. A learning system makes decisions based on accumulated experiences through successful solutions and related previous problems[5]. The machine learning systems have individual and common characteristics that enable to form cluster. A characteristic problem of this step is to identify what or which clustering methods can actually provide a useful and valid rating for the dataset under study. Among the concepts for screening and study of cluster analysis techniques[6] is the knowledge of Euclidean distance and Statistics techniques, well known algorithms which are suitable to the problem of interest. The dendrogram[7] has been used in the field of chemical and principal component analysis (PCA) also.

II. DETAILS EXPERIMENTAL

2.1. Methodology

Density Function Theory

Quantum mechanical calculations with DFT method using functional B3LYP and basis 6- 31G (d, p) are employed. Vibrational properties such
as I.R spectra were simulated at the above mentioned level of theory using optimized geometric parameters to separate compounds with your fingerprint. All calculations were performed using Gaussian 09 and results are achieved with Gauss View5.

Density Functional Theory (DFT) B3LYP / 6-31G (d, p) has been used to optimize the structures alcohol and aliphatic hydrocarbon (Figure 1).It was found a linear and growing dependence on energy with mass. And then the calculated energy by quantum mechanics with the enthalpy of combustion.It can be computed just knew the class of the molecule and its molecular weight to find the enthalpy of combustion. All correlations for this study were maximum.

The vibrational properties in the region of Infra-Red and Raman spectra gives structural molecular information of the density of states. Geometry of already optimized were used at the same basis set,

like DFT/B3LYP method at 6-31 G (d,p) level of theory[8], calculated computationally by gaussian09, and it was effective to separate alcohol of hydrocarbon and even arrange for size. Even if the compound has the same mass, but differences of organic class, conformational energy and spectrum is important to differentiate from other isomers, and the cheminformatics methods using DFT combining with PCA and Dendrogram are powerful tool for predicting the properties and separation of compounds.

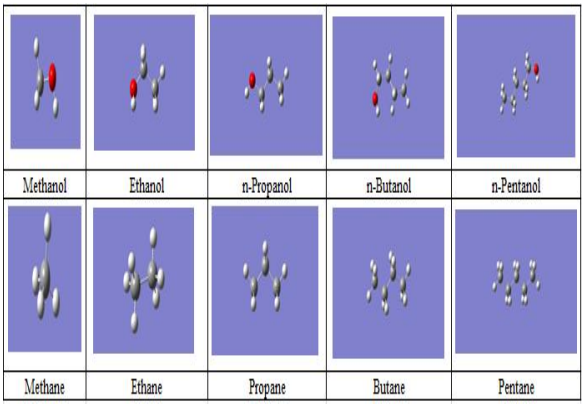

Fig.1. Chemical Compounds

Regression (or approximation of functions) is the process of determining the model of a real system that produces a response or output *y* variable when actuated by a set of input variables*x*.

Assuming that the actual system output as energy can be represented by a mathematical function of the inputs of mass. The goal of regression is to build a model that corresponds to the best approximation of the actual system[9].

The linear model is a relationship between the output variable *y*(Energy) and an input variable, was written.The linear regression objective is to estimate the parameters of a linear model of the form:.$y = ax + b$, as follow.

Principal Component Analysis (PCA)

The Principal Component Analysis(PCA) is the most widely used multivariate method[10]forreducing the sample space where the maincomponentsare linearly independent, making anoptimallinear transformation sometimes called the Karhunen-Loève Transform(KLT) or Hotelling Transform, widely used in pattern recognition. Calculate PCA is an important part of the cluster methodology in order to prevent misleading results. The PCA scores, is sufficient to confirm the clustering tendency.

Considering a database of vibrational frequency computed in gaussian09 as variance matrix (x), it was expressed in terms of scores and variables. Hence the data is modeled as a multivariate space can be written vectorially as:

$$\vec{x'} = R.\vec{x}$$

The goal is to find the maximum rotation matrix that allows to obtain the largest amount of information in a small space. The constraints of the optimization of this matrix is orthogonal and maximization the matrix variance of data, given by:

$$\arg\limits_{\vec{r_2}}\left[\max\left[\mathrm{var}(x_2')\right]\right]$$

subject to the constraints

$$\vec{r_2}^{\,t}.\vec{r_2} = 1$$

$$\vec{r_1}^{\,t}.\vec{r_2} = 0$$

Dendrogram

The hierarchical method is an unsupervised and agglomerative technique that examines the distances between all samples in the data set, and represents this information in the form of a two-dimensional graph calleddendrogram[11]. Through the dendrogram, one can visualize clusters and similarity between variables.It is a deterministic hierarchical method that aims to cluster data simultaneously in different scales by creating a tree of groups called dendrogram. The tree is a hierarchy of multiple levels where the groups at one level can be grouped as items of groups of higher levels , the criterion of the minimum distance.

III. RESULTS AND DISCUSSION

Table 1 shows the energies calculated using B3LYP/ 6-31 G (d,p):

Table1: Energy

n#	Compound	Molar Mass	Energy Abs	Energy Gaussian (a.u.)
1	Methanol	32	115,72396	-115,724
2	Ethanol	46	155,04667	-155,047
3	n-Propanol	60	194,3627	-194,363
4	n-Butanol	74	233,67966	-233,680
5	n-Pentanol	88	272,99629	-272,996
6	methane	16	40,524019	-40,524
7	ethane	30	79,838738	-79,839
8	propane	44	119,15537	-119,155
9	butane	58	158,47195	-158,472
10	Pentane	72	197,78849	-197,788

Figures 2 and 3 show the maximum correlation between the size of the chain with the molecular electronic property, the result of an addictive nature. When size increases, increases physical property directly proportional.

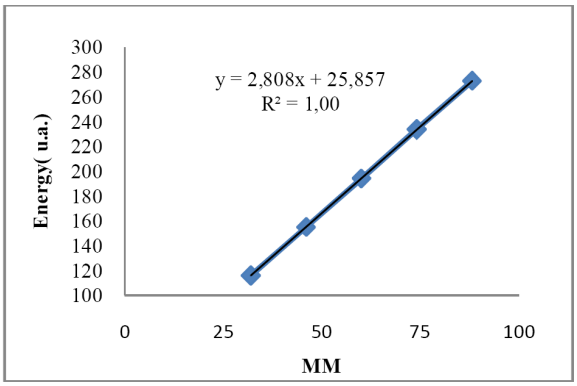

Fig.2. Alcohol

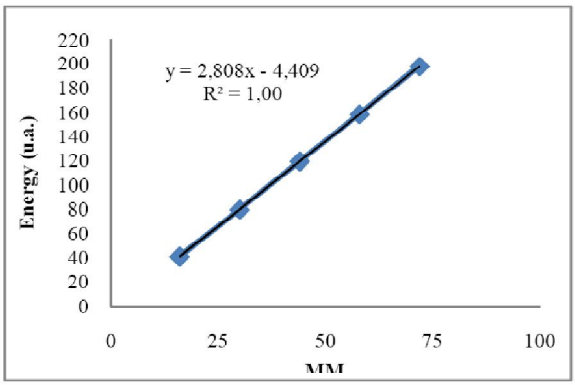

Fig.3. Hydrocarbon

The correlation energy calculated by quantum mechanics necessary to optimize the hydrocarbons was also shown to be directly proportional to the enthalpy of combustion, as it can be seen in Figure 4:

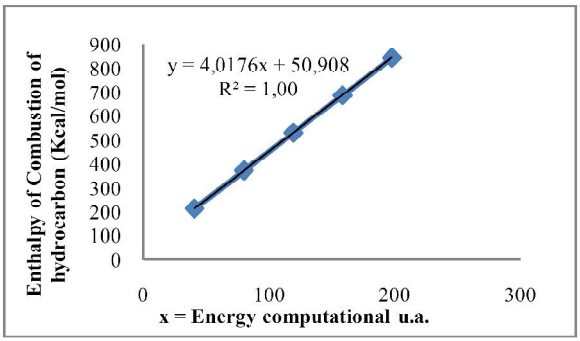

Fig.4. Enthalpy of combustion of hydrocarbon (Kcal/mol)

This correlation can be useful equation to be used the higher the carbon chain in aliphatic hydrocarbons.
However, it is first necessary to separate the compounds of interest into similar groups, and vibrational properties was used as **Table 2**. And after to start correlation with data set Quantum Chemistry. This study was separated by statistical using the concept of maximum variance of PCA method, the organic classes of compounds, as can be seen the result of Figure 5. The Figure 6 corroborates the results using a hierarchical method that using method of Euclidean distance, forming two groups in each dendrogram branch.

After grouped is only possible by knowing the molecular weight and the class to determine the enthalpy of combustion.

Table2: Energy and Vibrational Properties

		Energy	OH	C-O	C-H
1	Methanol	115,7239635	3034,48	1061,21	2988,88
2	Ethanol	155,0466724	3107,17	1083,08	2987,74
3	n-Propanol	194,3626953	3107,57	1070,1	2988,49
4	n-Butanol	233,6796637	3108,49	1089,78	2979,85
5	n-Pentanol	272,9962929	3087,52	1083,61	2977,26
6	methane	40,52401948	0	0	3162,96
7	ethane	79,83873847	0	0	3121,92
8	propane	119,1553651	0	0	3108,41
9	butane	158,4719537	0	0	3105,56
10	Pentane	197,7884873	0	0	3104,72

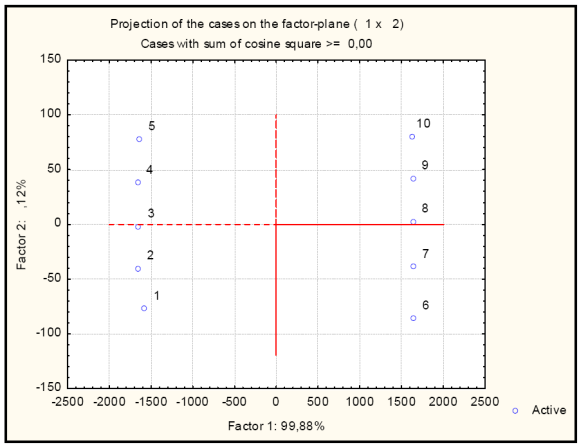

Fig.5. PCA

The separation using PCA and dendrogram was able to statistically separate the compound the other, only data using quantum mechanics, with a powerful tool for separating compounds of a different nature for their vibrational properties.

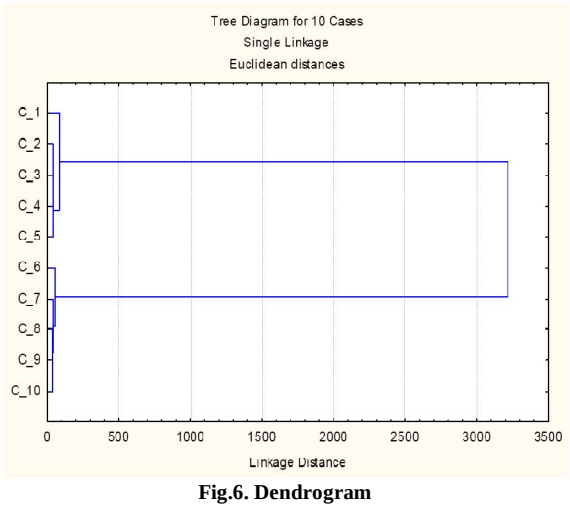

Fig.6. Dendrogram

CONCLUSIONS

The computational statistical study showed that the correlation is direct for energy computed using DFT and functional B3LYP with the enthalpy and increased carbon chair by arranging according to organic class, and the major conclusions were:

1. In all cases studies, alcohol and hydrocarbon, the correlations of energy with the size of the chain confirmed the expectations of increased physical property due to the aument of intermolecular forces of electrostatic nature involving the compounds.
2. It is possible to separate compounds according to their class using machine learning, with the techniques of PCA and dendrogram.
3. Use the vibrational properties as molecular descriptors are efficient to apply machine learning techniques and separate compounds using the fingerprints of those components.

ACKNOWLEDGMENTS

PhD Sustainable Chemistry Course and classes in Porto University where I learned Statistical Software and Aveiro University, Gaussian Software.

REFERENCES

[1] J. Gasteiger., T. Engel. *Chemoinformatics : A Textbook*. John Wiley & Sons, 2004

[2] P. Hohenberg; Walter Kohn . "Inhomogeneous electron gas". Physical Review 136 (3B): B864–B871,1964.

[3] W. Kohn,; L. Sham. "Self-consistent equations including exchange and correlation effects". Physical Review 140 (4A): A1133–A1138, 1965.

[4] J. Perdew, ; J. Chevary; S. Vosko; Jackson, Koblar, A; Pederson, Mark R; Singh, D J; Fiolhais, Carlos. "Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation". Physical Review B46 (11): 6671, 1992.

[5] Gaetano Ragno; Michele De Luca; Giuseppina Ioele; An application of cluster analysis and multivariate classification methods to spring water monitoring data, Microchemical Journal 87, 119-127, 2007.

[6] Nørskov-Lauritsen, L.; Bürgi, H. B. , Cluster analysis of periodic distributions; application to conformational analysis, New York, NY [u.a.] : Wiley-Blackwell Journal of Computational Chemistry 6 (1985), S. 216-228, 1985.

[7] Adamson, G.; Bawden, D.; Comparison of Hierarchical cluster Analysis Techniques for Automatic Classification of Chemical Structures, Journal of Chemical Information and Computer Sciences, 21, 204-209, 1981

[8] Xiao Dong , Kevin E. Gilbert , Rajarshi Guha , Randy Heiland , Jungkee Kim , Marlon E. Pierce , Geoffrey C. Fox , and David J. Wild, Web Service Infrastructure for Chemoinformatics, J. Chem. Inf. Model., 47 (4), pp 1303–1307, 2007.

[9] P. Geladi; B. Kowalski; "Partial least squares regression: a tutorial" Analytica Chimica Acta 185: 1-17, 1986.

[10] K. Varmuza; P. Filzmoser; Introduction to Multivariate Statistical Analysis in Chemometrics; CRC Press; USA, Pp. 34-40, 2009.

[11] S. Yidana; D. Ophori; B. Yakubo; A multivariate statistical analysis of surface water chemistry data – the Ankobra Basin, Ghana, Journal of Environment Management V (86)1; pages 80-87, 2008.

★★★