Comparing Computational and Chemometric Approaches to Calculate Aromatic Bond Lengths: a Case of Nucleobases

Rudolf Kiralj (PQ), Márcia Miguel Castro Ferreira (PQ)*

marcia@iqm.unicamp.br

Instituto de Química, Universidade Estadual de Campinas, Campinas 13083-862, SP, Brasil

Palavras Chave: *Pauling p-bond order, HCA-PCA best model selection, dimer geometry optimization*

Introdução

Nucleobases (nucleic acid bases) are carbohydrate derivatives of heterocyclic and carbocyclic compounds, whether the attachment is through N, C or O. They can be classified as standard (A, T, G, C, U) and non-standard nucleobases and nucleobase analogues, and also as natural and synthetic. They can possess physical, chemical, biochemical, pharmacologic and physiologic effects desired in biotechnology, medicine and material chemistry. These properties can be well reflected through, or correlated with nucleobase bond lengths (CC, CN, CO). This work presents further development of initial report on nucleobase bond length calculation based on Pauling harmonic potential curve (employing Pauling π-bond orders), bond length-bond order relationships studied by chemometric methods (including the Pauling bond orders corrected to crystal packing effects and bond electrotopological indices), semi-empirical PM3 and *ab initio* HF 6-31G** methods.[1] The initial set of nucleobase geometrical data retrieved from Cambridge Structural database was extended, and semi-empirical MNDO and AM1, molecular mechanics MMFF94 and the inverse Gordy's curve calculations were performed.[2] The selection of the calculation method which reproduces the experimental bond lengths in the best way, was carried out by coupled Hierarchical Cluster Analysis (HCA) – Principal Component Analysis (PCA) using statistical indices for methods, experimental and calculated bond lengths, average experimental and calculated bond lengths for nucleobase rings, and experimental and calculated Julg's structural aromaticity indices with their errors for the same rings. The cytosine dimer geometry was optimized by computational methods used in this work.

Resultados e Discussão

The HCA – PCA procedure for finding the best model for calculation of bond lengths and structural parameters consists of inspection of corresponding PCA plots (with the first three principal components) and HCA dendograms to find out the clustering of the methods and their closeness to experiment or to some other reference method (*ab initio*, for example). This inspection showed that the prediction/calculation power of all the methods depends primarily on the nature of methods, and then on the property or set of parameters which are treated by the HCA-PCA procedure. In general, Pauling and Gordy curves as univariate models are the worst for bond length prediction for nucleobases, then follow multivariate models (Multiple Linear Regression, Principal Component Regression, Partial Least Squares Regression) and semi-empirical methods with MMFF94, and the best models is the *ab initio*. Hence, the practical advantage of multivariate models is that they are easy to use and compete with semi-empirical and molecular mechanics methods.

The cytosine dimer bond lengths, compared to experimental ones, even more clearly show that *ab initio* methods are the best, semi-empirical are in the middle, while MMFF94 is the worst. Structural interpretation for that is the fact that a cytosine dimer includes several hydrogen bonds which are coupled with cytosine π-electron delocalization. These hydrogen bonds are badly reproduced by semi-empirical methods and MMFF94.

Conclusões

This extensive study on CC, CN and CO nucleobase bond lengths employing various chemometric and computational approaches shows *via* a coupled HCA-PCA procedure that promising multivariate models compete with semi-empirical and molecular mechanics methods, although *ab initio* are still the best. Hydrogen bonds seem to be important in nucleobase π-electron delocalization.

Agradecimentos

The authors thank to FAPESP.

[1] .Ferreira, M. M. C., Kiralj, R., XI SBQT, Caxambu, 18 – 21 November, 2001, P228.

[2] Kiralj, R., Ferreira, M. M. C., *J. Chem. Inf. Comput. Sci.*, in press