Use of the Cambridge Structural Database in Study of Single and Partial Double C-X (X=C,N,O) Bonds in Organic Molecules in Crystalline State

Rudolf Kiralj and Márcia M. C. Ferreira, Laboratório de Quimiometrica Teórica e Aplicada, Instituto de Química, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, 13084-971, Brazil

E-mails: marcia@iqm.unicamp.br, rudolf@iqm.unicamp.br,

URL: lqta.iqm.unicamp.br

The number of chemical compounds in the CAS (Chemical Abstract Service) Registry Database, 50% of which are peptides and proteins, is growing superexponentially (Figure 1).

A similar trend is observed for the number of organic & organometallic crystal structures in the Cambridge Structural Database (CSD, Figure 2) [1,2]. Today there are 272066 entries in the CSD, from X-ray, neutron and synchrotron radiation.

Consequently, CSD enables data mining for many high-quality crystal structures of many classes of compounds, as for example:
-simple organic compounds with single C-C bonds [3,4]
-planar benzenoid polycyclic aromatic hydrocarbons (PB-PAHs) and their aza-, diaza- and polyaza-derivatives [3-9]
-picrate-like systems [5]
-nucleobases [9]
Although different, these classes of organic compounds have something in common: CC bonds which vary from pure single to double, and sometimes CO or CN bonds are acting similarly.

The quantitative relationships between experimental bond lengths in crystalline state and the Pauling π-bond orders were studied in all these classes of molecules in univariate and sometimes in multivariate way.

Data mining, computational, chemometric and other details can be found in the cited works. However, a few moments should be pointed out:
-data mining took into account desirable molecular geometry of the compounds, excluded metal complexes, disordered and low-quality structures
-gas-phase molecular structures and structures from other non-crystallographic sources, as well as *ab initio* B3LYP 6-31G calculations have been done only for simple organic compounds; this set of compounds was extended to carbon allotropes and some other species like molecular complexes including PAHs to have the complete idea on C-C multiple, single and electron-defficient bonds**
-univariate relationships were based on linear regression with bond lengths and Pauling π-bond orders P or bond numbers $M = P + 1$ in the form: $D/\text{Å} = a + b\,P$ or $D/\text{Å} = a + b\,M$
-multivariate regression included more variables: P bond orders corrected to crystal packing effects, the sum of atomic numbers, and topological indices counting the number of neighbouring bonds or rings around a particular bond

Figure 3 shows the whole range of C-C bond multiplicity, from acethylene and acetylide anion, to saturated hydrocarbons and diamond, and even more to intermolecular complexes bound by weak C-C bond.
It is important to point out that Novoa *et al*. [10] discovered the longest C-C bond even known: such bonds are electron defficient bonds with very small bond number M, and the bond lengths go up to the graphite interlayer distance. Various carbon allotropes are spread over the entire range of M values. This discovery shakes the old concept of the C-C bond and "nonbonding interatomic or intermolecular interactions".

The resulting univariate regression coefficients a/b exhibit expected similarities and differences, as follows:

	C-C	C-N	C-O	all bonds
PB-PAHs	1.468(2)/-0.147(5)	-	-	-
Aza-PAHs	1.462(6)/-0.143(13)	1.444(10)/-0.184(18)	-	-
Diaza-PAHs	1.458(3)/-0.143(8)	1.415(4)/-0.152(8)	-	-
Polyaza-PAHs	1.421(30)/-0.087(81)	1.431(20)/-0.128(42)	-	-
Picrates	1.497(11)/-0.212(25)	-	1.326(5)/-0.198(18)	-
Nucleobases	1.487(7)/-0.202(16)	1.398(3)/-0.127(7)	1.295(16)/-0.101(26)	1.429(5) /-0.199(13)

Figures 4 and 5 use three simple structural parameters: the mean CC bond length, CC bond length variation, and the CX bond fraction, to visualize similarities/differences between these π-systems, relatively to benzene as the standard aromatic system.

The use of multivariate methods – Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) for bond length classification, and Partial Least Squares (PLS) regression, showed to be more adequate than univariate statistics for the study of C-C, C-N and C-O bond lengths in PB-PAHs and nucleobases.

Much better PLS regression models were obtained using more bond variables!

PB-PAHs: $d\,/\,\text{Å} = 1.431 - 0.060\,P - 0.063\,P_{cc} + 0.006\,n + 0.004\,m + 0.001\,l$
Nucleobases: $d\,/\,\text{Å} = 2.304 - 0.080\,P - 0.078\,P_{s} - 0.068\,Q - 0.006\,n$

where n, m, l are topological indices, Q is the sum of atomic numbers, and Ps are Pauling π-bond orders including crystal packing effects (intermolecular bonds and interactions in crystalline state).

Figure 6 shows bond types and the effects of neighbourhood (topological indices) and crystal packing to $D - P$ plot.
Figure 7 reveals similarity between PB-PAHs and nucleobases in PCA, confirmed also by HCA (not shown here): analogous clustering of bonds in classes (roman numerals) and the reduction of original bond variables into three principal components (PCs).

CONCLUSIONS!

DATA MINING + CHEMOMETRIC ANALYSIS + STRUCTURAL & COMPUTATIONAL METHODS = A VERY POWERFUL MEANS TO STUDY BOND LENGTHS IN ORGANIC CRYSTALS + INTERESTING AND USEFUL RESULTS ON BASIC CHEMICAL CONCEPTS (WHAT IS A BOND?? C-C BOND??) + DIRECTIONS FOR FUTURE STUDIES (intrinsic π-system properties, substitution effects, crystal packing effects, (hetero)aromaticity, crystal packing, etc.)

Literature: 1) F. H. Allen, *Acta Cryst*, B58 (2002) 380; 2) CCDC site: *http://www.ccdc.cam.ac.uk/*; 3) M. M. C. Ferreira, R. Kiralj, *Hem. Pregl*., submitted; 4) M. M. C. Ferreira, R. Kiralj, unpublished; 5) R. Kiralj *et al*., *Acta Cryst*., B52 (1996) 823; 6) R. Kiralj *et al*., *J. Mol. Struct. – Theochem*, 427 (1998) 25; 7) R. Kiralj *et al*., *Acta Cryst*., B55 (1999) 55; 8) R. Kiralj, M. M. C. Ferreira, *J. Chem. Inf. Comput. Sci*., 42 (2002) 508; 9) R. Kiralj, M. M. C. Ferreira, *J. Chem. Inf. Comput. Sci*., in press; 10) J. J. Novoa *et al*., *Angew. Chem. Int. Ed*., 40 (2001) 2540.

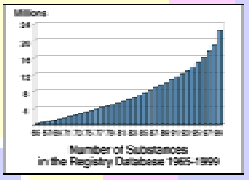

Figure 1. Cumulative growth of the CAS-Registry Database. http://www.cas.org/casdb.html#regdb

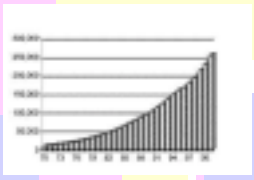

Figure 2. Cumulative growth of the CSD from 1970 to 2001. CCDC Annual Report for 2001. http://www.ccdc.cam.ac.uk/annrep2001/report.html

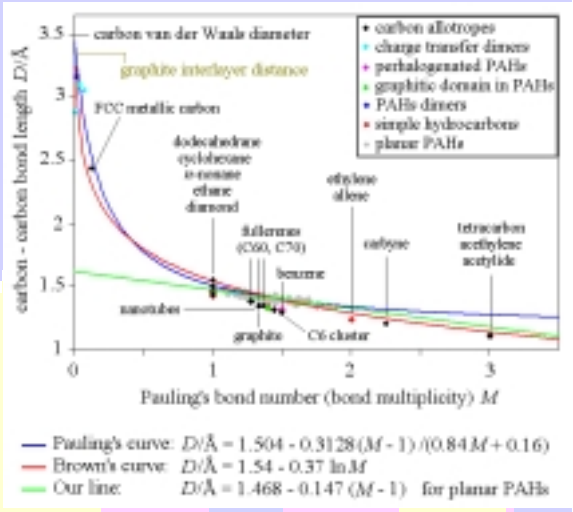

Figure 3. C-C bond length – Pauling's bond number relationship for carbon allotropes, hydrocarbons and their derivatives. The relationship covers wide range of bond multiplicity, from pure triple to intermolecular bond. Some analytical and regression curves are overlapped with the plot.

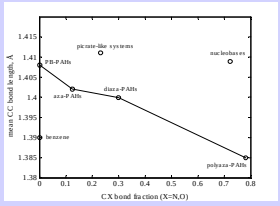

Figure 4. The mean C-C bond length in various (hetero)aromatic systems.

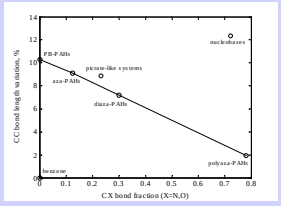

Figure 5. The C-C bond length variation in various (hetero)aromatic systems.

Figure 6. The bond neighbourhood effects and the crystal packing effects on C-C, C-N and C-O bond length – bond order relationships for nucleobases.

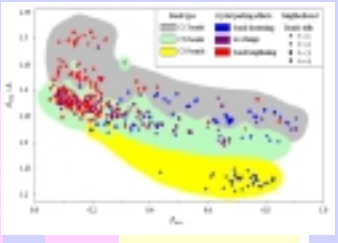

Figure 7. 3D scores plots from PCA show similarity and dissimilarity between PB-PAHs and nucleobases. Roman numerals denote various classes of bonds defined by topological indices: (*nml*) for PB-PAHs, (*Qn*) for nucleobases. The colouring for the packing effects is the same as in Figure 6.