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

Rudolf Kiralj (PQ) and Márcia M. C. Ferreira (PQ)*, Laboratório de Quimiometria Teórica e Aplicada, Instituto de Química, Universidade Estadual de Campinas, Campinas, SP, 13083-970, Brazil.

marcia@iqm.unicamp.br, rudolf@iqm.unicamp.br, http://lqta.iqm.unicamp.br

INTRODUCTION

Nucleobases (nucleic acid bases) are carbohydrate derivatives of heterocyclic and carbocyclic compounds, and can be classified as standard (A, T, G, C, U) and non-standard, also as natural and synthetic. They can possess physical, chemical, biochemical, pharmacologic and physiologic effects desired in biotechnology, medicine and material chemistry, which can be well observed through, or correlated with bond lengths (CC, CN, CO). This work represents more advanced development of initial nucleobase bond length calculation based on Pauling harmonic potential curve, bond length-bond order relationships studied by chemometric methods, semi-empirical PM3 and *ab initio* HF 6-31G** methods.[1] The initial set of nucleobase 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 best calculation method for bond lengths was carried out by coupled Hierarchical Cluster Analysis (HCA) – Principal Component Analysis (PCA) The cytosine dimer geometry was optimized by computational methods used in this work.

[1]Ferreira, M. M. C., Kiralj, R., XI SBQT, Caxambu, 18 – 21 Nov. 2001, P228. [2]Kiralj, R., Ferreira, M. M. C., *J. Chem. Inf. Comput. Sci*., online: April 19, 2003.

METHODS

Database Mining: the search for crystal structures of nucleosides in Cambridge Structural Database[1,2] (CSD).

Resonance Structures: drawing resonance structures for nucleobases, and calculation of Pauling π-bond orders for nucleobases.

Bond Length-Bond Force Relationships (BLBFR): updated for CC, CN, CO bond lengths.

The Extended HOSE[3] Model: updated by new BLBFR and applied for calculation of weighted Pauling π-bond orders were calculated.

Molecular Mechanics and Quantum Chemical Calculations: geometry optimization of nucleosides at MMFF94, MNDO, AM1, PM3 and HF 6-31 G** level.

Chemical Bond Descriptors: Pauling π-bond orders corrected to crystal packing effects, and bond topological and electrotopological indices.

Simple Bond Length-Bond Order Relationships (BLBOR): linear regression (LR), Pauling's harmonic and Gordy's inverser curves.

Chemometric Analysis: PCA and HCA for analysis of bond lengths data; regression models MLR, PCR and PLS for prediction of CC, CN, CO bond lengths.

HCA-PCA Procedures for the Best Prediction Models: all results analyzed by PCA-HCA to find out the best prediction model (analytical, regression or computational).

Cytosine Geometry Optimization: geometry optimizaton of isolated cytosine and cytosine clusters at MM, semi-empirical and *ab initio* level.

[1] Cambridge Structural Database, NCSA ChemViz., Univ. Urbana-Champaign. [2]Berman H. M. *et al*., *J. Am. Chem. Soc*. **1996**, *118*, 509-518. [3]Krygowski T. M. *et al*., *Acta Crystallogr*. **1983**, *B39*, 732-739.

SOME RESULTS & COMMENTS

To see related Figures and Tables below the text.

Database mining: 24 crystal structures of nucleosides used as the training set, with 309 bond lengths: 86 CC, 185 CN, 38 CO bonds.

The HCA – PCA procedure for finding the best model: 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.

Cytosine geometry: 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.

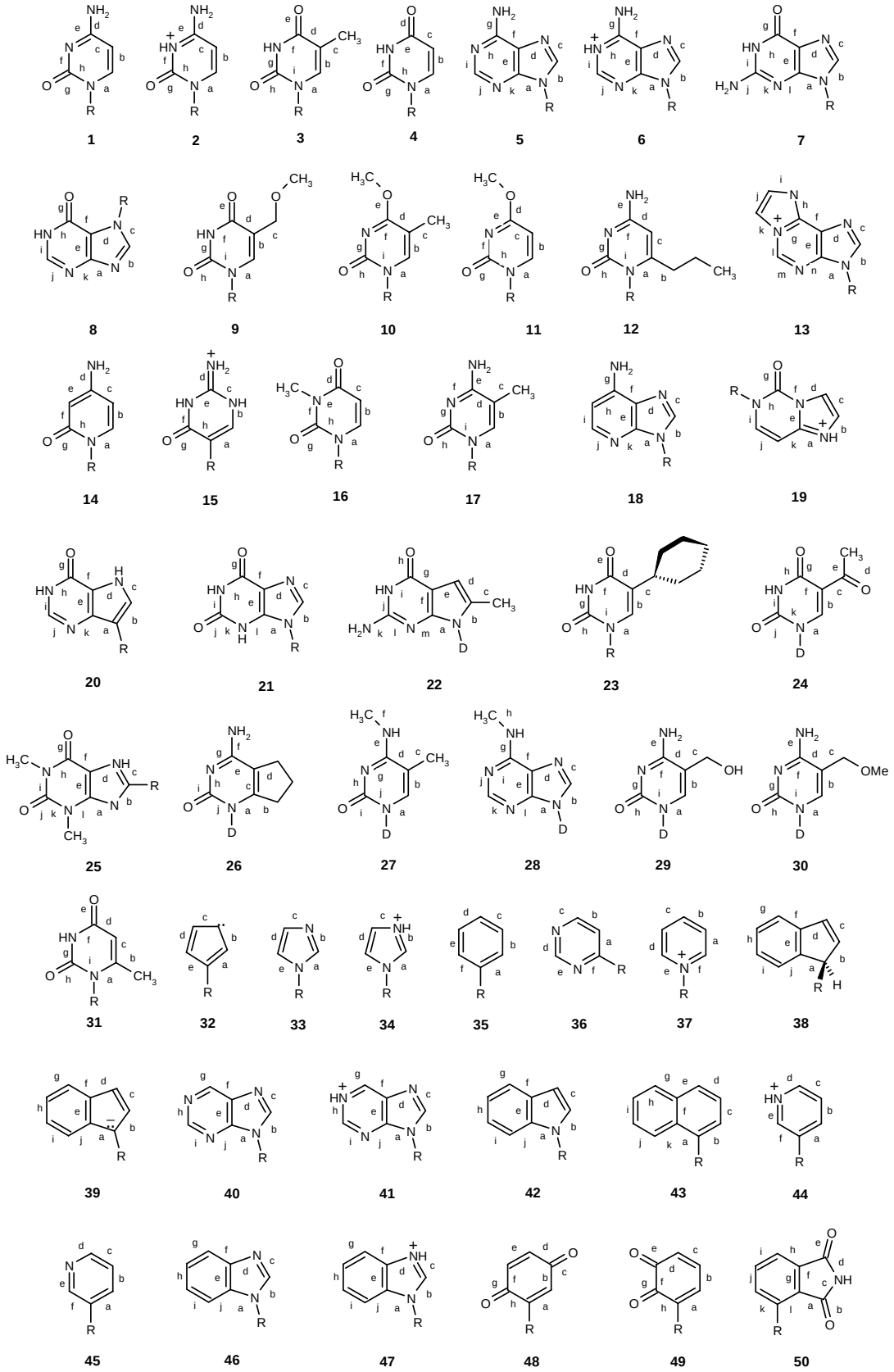
CONCLUSIONS

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 (resonance–assisted hydrogen bonds).

CC, CN, CO nucleobase bond lengths depend on π-bond orders, electrotopological and topological indices accouting for π-electron delocalization effects, bond type and neighborhood, respectively. The bond lengths are 3D phenomenon and can be classified in 9 classes based the indices.Distinction between purines and pyrimidines, and also among five classes of nucleobases (C, T, U, A, G) can be clearly observed based on bond lengths only.

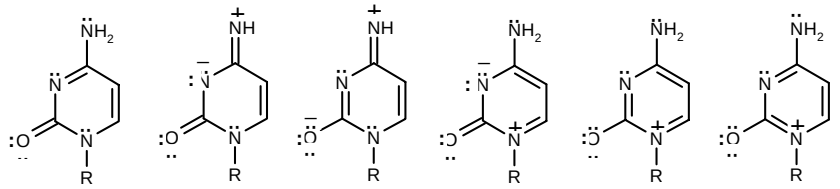
ACKNOWLEDGEMENT

The authors thank to FAPESP.



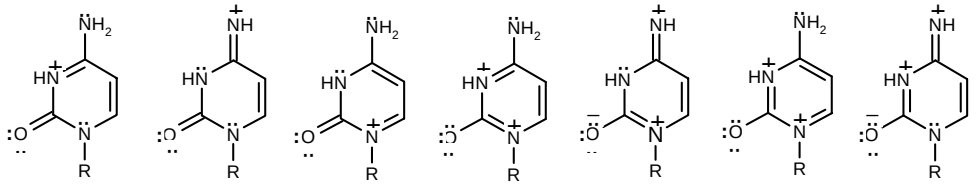
Nucleobases: training (**1**-**31**) and prediction (**32**-**50**) set.

C



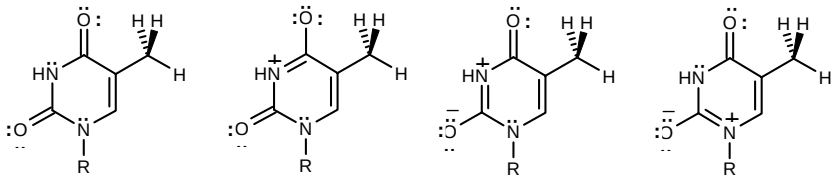
I (24.3%) II (24.3%) III (15.2%) IV (13.2%) V (9.9%) VI (13.1%)

C^+

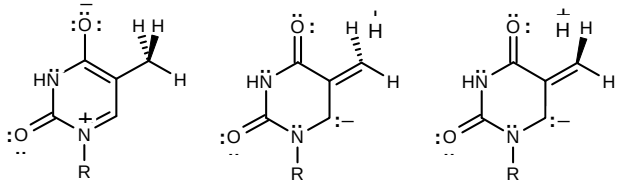


I (20.1%) II (26.1%) III (13.2%) IV (10.1%) V (11.4%) VI (7.9%) VII (11.3%)

T

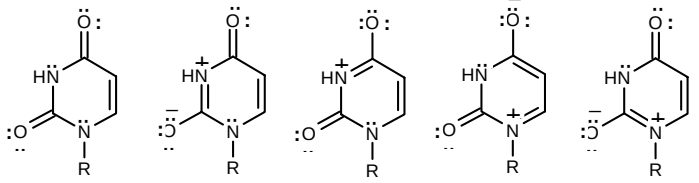


I (30.2%) II (12.6%) III (12.2%) IV (12.1%)



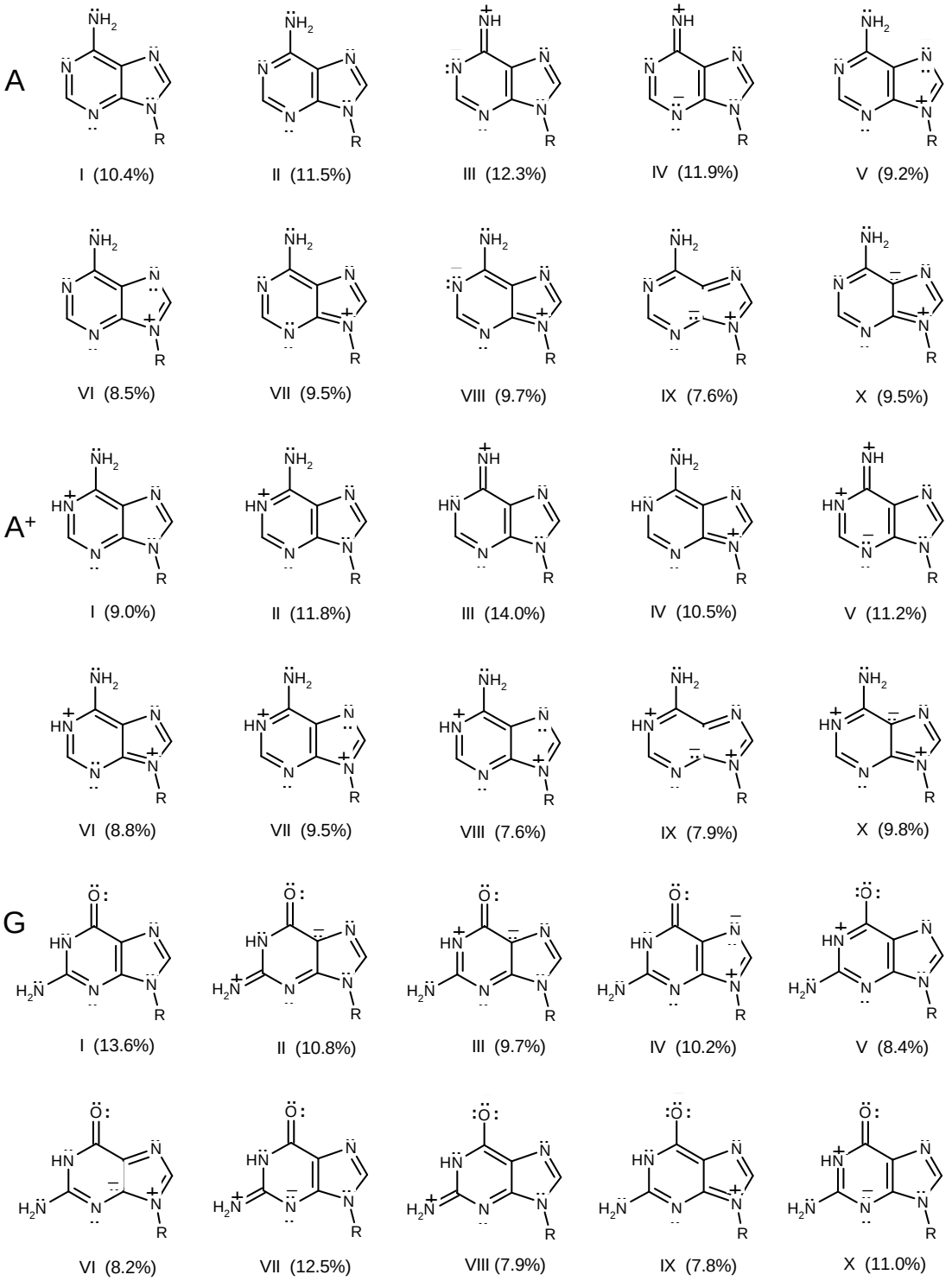
V (8.6%) VI (12.1%) VII (12.1%)

U

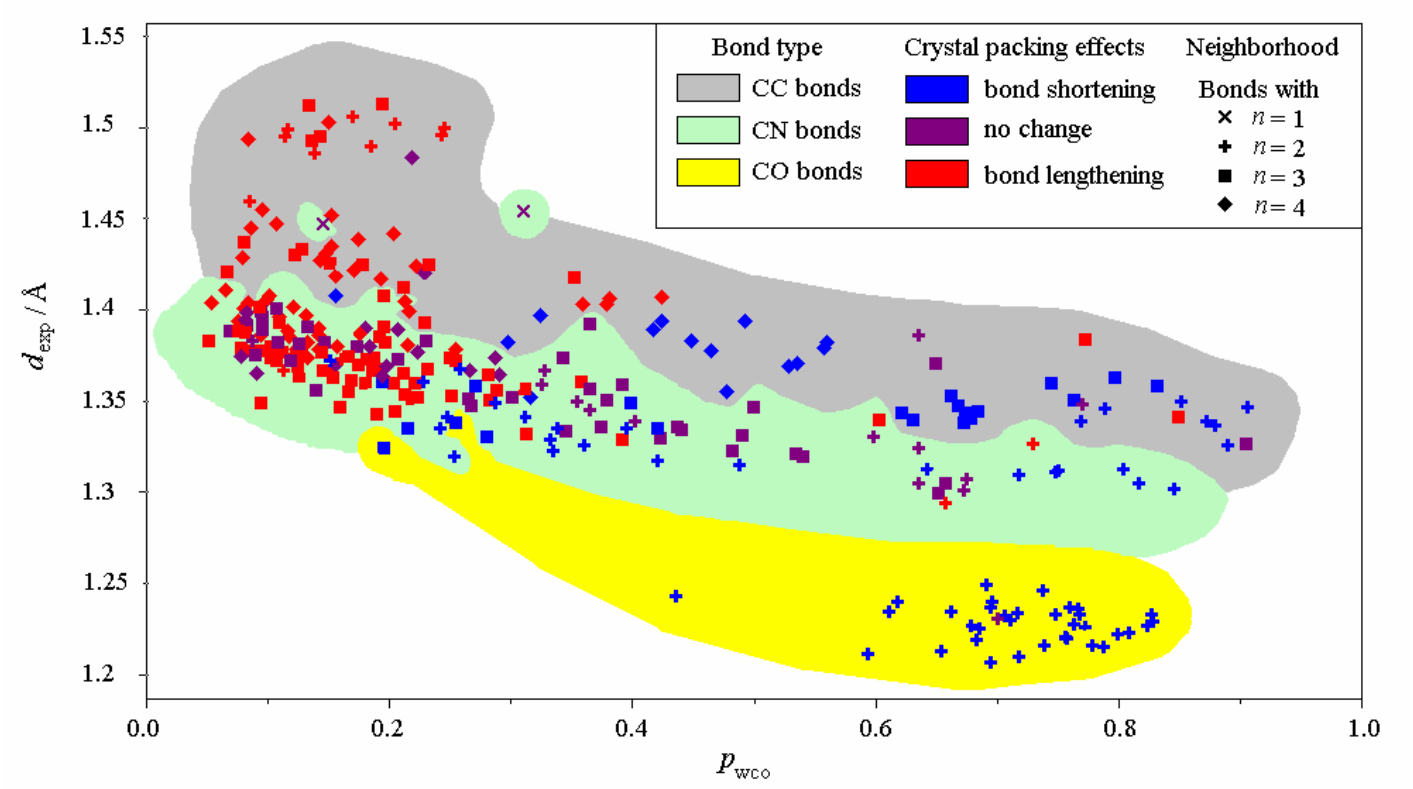


I (38.9%) II (16.1%) III (17.3%) IV (12.1%) V (15.6%)

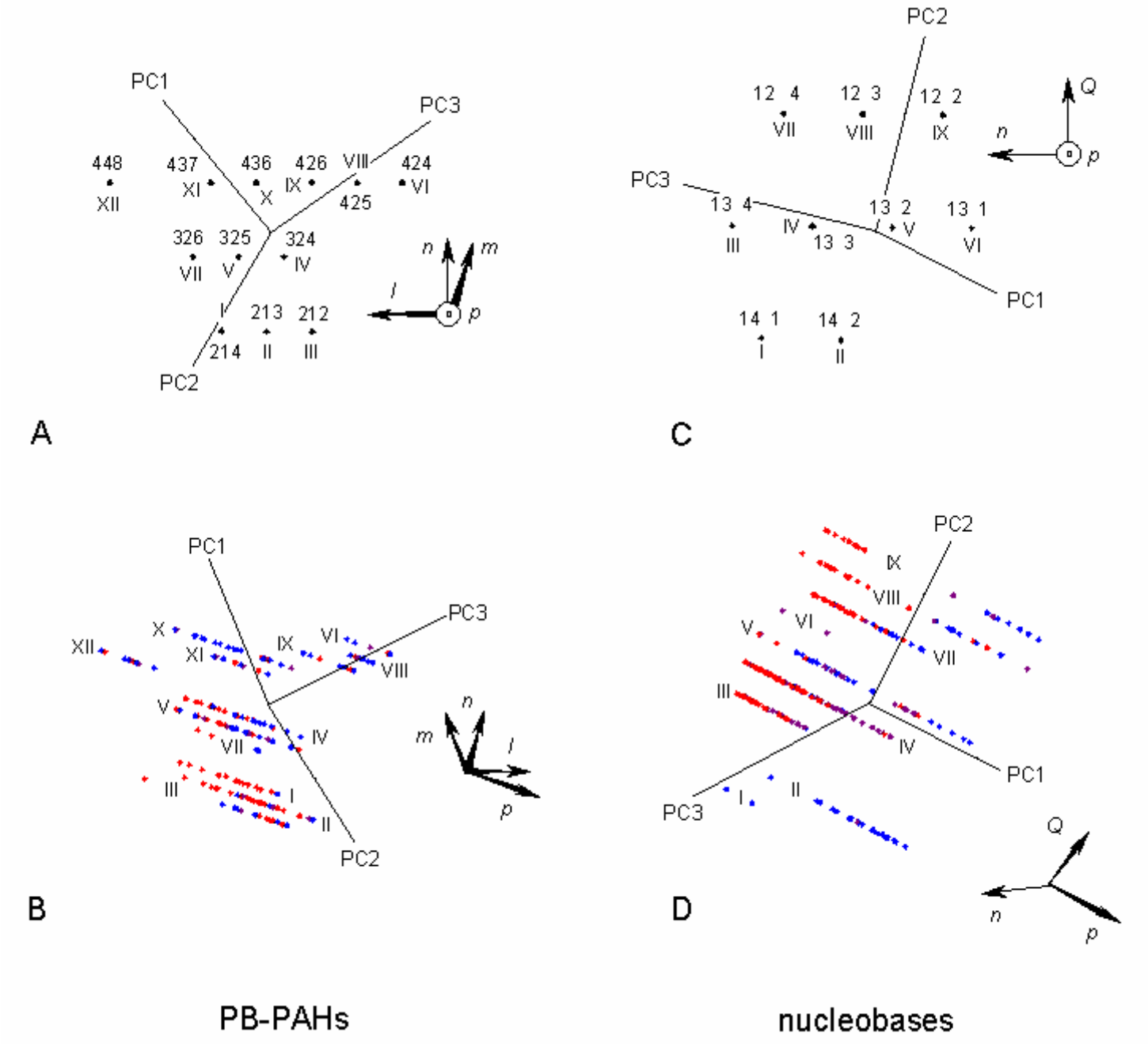
Resonance structures and their weights for standard pyrimidines.



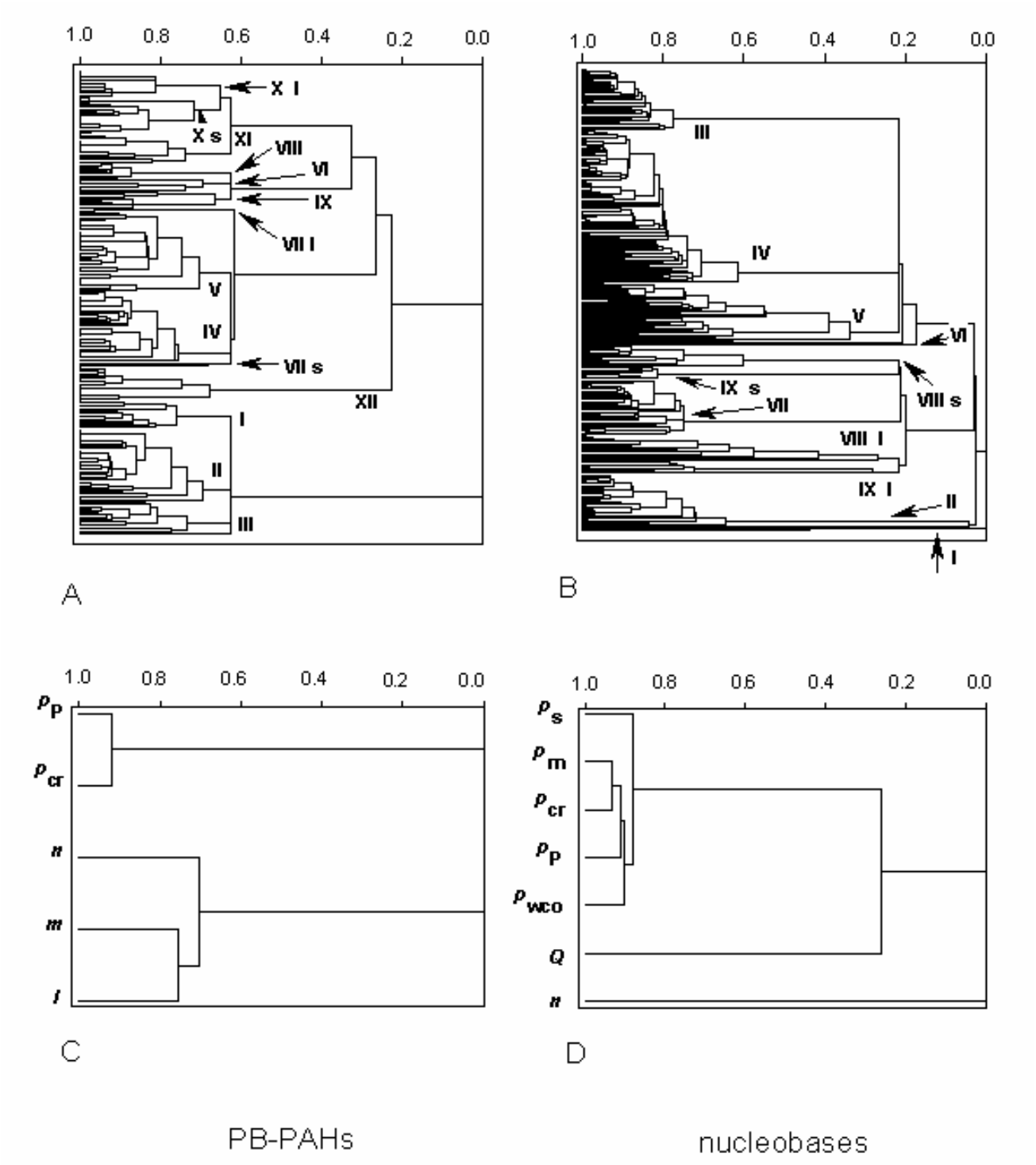
Resonance structures and their weights for standard purines.



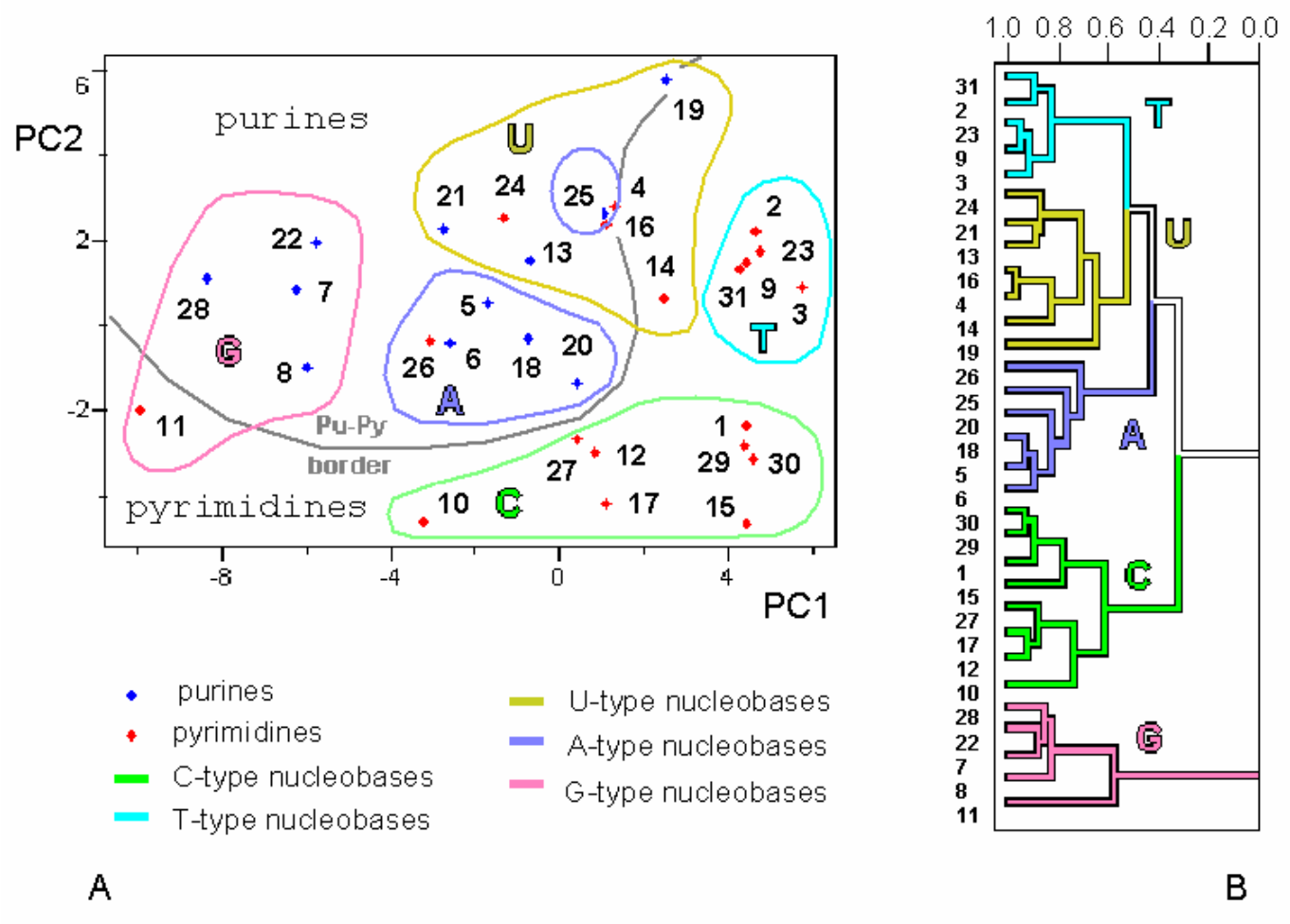
A weighted Pauling π-bond order corrected to crystal packing effects (p_{wco}) *versus* experimental nucleobase bond length (d_{exp}). Crystal packing effects refer to predicted d_{pred} values obtained from d / Å = a + b p_{wco} regression equation; variable n is the number of neighbouring non-H atoms around a particular bond.



3D scores plots for PB-PAHs and nucleobases. Groups of bonds (rows of points) are projected perpendicularly to the plane of paper, and are visible as points in A and C, and marked with the group number (roman numerals) and with integer descriptors: A: *nml* for PB-PHs; B: *Qn* for nucleobases. Crystal packing effects, with the same coloring as in the previous figure, are presented in B and D. The coordinate systems with original axes (*p* – bond orders) are qualitatively oriented.



HCA dendograms for samples (A, B) and variables (C, D) for PB-PAHs and nucleobases (p_{wco}). The groups from the 3D PCA scores plots are marked in the same way, adding **s** (for short bonds) and **l** (for long bonds) for subgroups of VII and X (PB-PAHs), and VIII and IX (nucleobases).



A. Scores plot for Julg's structural aromaticity index data set. B. The corresponding HCA dendogram. Distinction between Pu and Py is visible. Five types of nucleobases, C-, T-, U-, A- and G-type, are marked with different color.

Statistical parameters[a] for cytidine clusters[b] optimized by various methods.[c]

Cluster*	***Method***	Δ_{max}/Å	<Δ>/Å	σ/Å	<Δ/σ>	Δ*HOMA*	Δ*A*
Monomer	MMFF	0.087	0.043	0.052	42.63	0.599	0.468(98)
(0)	MNDO	0.069	0.038	0.043	37.50	0.249	0.354(93)
	AM1	0.063	0.027	0.03	26.75	0.138	0.223(87)
	PM3	0.071	0.038	0.046	37.75	0.278	0.390(94)
	HF	0.048	0.020	0.026	20.13	0.427	0.298(92)
	B3LYP	0.048	0.020	0.024	19.88	0.173	0.218(87)
Dimer	MMFF	0.086	0.042	0.050	41.88	0.605	0.472(98)
(2)	MNDO	0.067	0.037	0.041	36.50	0.246	0.345(92)
	AM1	0.055	0.025	0.030	24.75	0.091	0.166(84)
	PM3	0.071	0.035	0.042	34.75	0.259	0.359(93)
	HF	0.035	0.014	0.019	14.38	0.337	0.198(87)
	B3LYP	0.028	0.010	0.013	10.00	0.059	0.088(80)
Trimer	MMFF	0.057	0.027	0.032	27.13	0.182	0.066(80)
(4)	MNDO	0.064	0.034	0.039	34.25	0.215	0.311(91)
	AM1	0.055	0.025	0.030	25.00	0.074	0.153(83)
	PM3	0.059	0.030	0.036	30.25	0.169	0.261(88)
Tetramer	MMFF	0.052	0.021	0.026	21.00	0.388	0.269(90)
(6)	MNDO	0.063	0.035	0.040	34.88	0.212	0.311(91)
	AM1	0.050	0.024	0.028	24.00	0.026	0.111(81)
	PM3	0.055	0.028	0.033	28.38	0.119	0.214(86)
Pentamer	MMFF	0.044	0.024	0.028	24.25	0.384	0.263(90)
(7)	MNDO	0.063	0.035	0.039	34.63	0.204	0.305(91)
	AM1	0.051	0.024	0.028	24.13	0.031	0.116(81)
	PM3	0.054	0.028	0.032	27.75	0.105	0.202(86)
Hexamer	MMFF	0.051	0.021	0.027	21.38	0.453	0.289(91)
(9)	MNDO	0.063	0.034	0.039	34.13	0.197	0.299(90)
	AM1	0.050	0.025	0.029	25.25	0.023	0.112(81)
	PM3	0.053	0.028	0.034	28.38	0.089	0.192(85)

[a]Statistical parameters based on d_{exp} and d_{cal} for the eight cytosine bonds, experimental and calculated *HOMA* and *A* for cytosine+. Errors for HOMA are less than 0.001. $HOMA_{exp}$ = 0.465 and A_{exp} = 0.682(53). [b]H-bonding cluster consisting of the reference cytidine molecule and from zero to five neighbouring molecules. [c]Computational methods: molecular mechanics MMFF94; semi-empirical MNDO, AM1, PM3; *ab initio* – Hartree-Fock (HF) and DFT with B3LYP functional (B3LYP). *The number of hydrogen bonds between the referent and neighboring molecules is given in the brackets.