On Heteroaromaticity of Nucleobases. Bond Lengths as Multidimensional Phenomena[†]

R. Kiralj and M. M. C. Ferreira*

Laboratório de Quimiometria Teórica e Aplicada, Instituto de Química, Universidade Estadual de Campinas,
Campinas, SP, 13083-862, Brazil

Received September 6, 2002

Three hundred and nine carbon−carbon, carbon−nitrogen, and carbon−oxygen π-bond lengths in high precision crystal structures of 31 purine and pyrimidine nucleobases were related to the Pauling π-bond order, its analogues corrected to crystal packing effects, the numbers of non-hydrogen atoms around the bond, and the sum of atomic numbers of the bond atoms. Principal Component Analysis (PCA) and Hierachical Cluster Analysis (HCA) demonstrated that the bond lengths in the nucleobases are three-dimensional phenomenon, characterized by nine distinct classes of bonds. Bond lengths predicted by Linear Regression models, Pauling Harmonic Potential Curves, Multiple Linear Regression, Principal Component, and Partial Least Squares Regression were compared to those calculated by molecular mechanics, semiempirical, and *ab initio* methods using PCA-HCA procedure on the calculated bond lengths, statistical parameters, and structural aromaticity indices. Incorporation of crystal packing effects into bond orders makes multivariate models to be competitive to semiempirical results, while further improvement of quantum chemical calculations can be achieved by geometry optimization of molecular clusters.

1. INTRODUCTION

Nucleobases (nucleic acid bases) are carbohydrate derivatives of natural or synthetic heterocyclic[1] and carbocyclic[2] compounds, whether the attachment is through N, C, or O. Adenine (A), guanine (G), cytosine (C), thymine (T), and uracile (U) are the common (standard) nucleobases in natural DNA and RNA. There are also many naturally occurring nucleobases, incorporated in various biomolecules or participating in biochemical processes.[1,3] Among them, modified or nonstandard nucleobases are derivatives of A, G, C, T, or U; over 100 were found in RNA and DNA.[4] Synthetic nucleobases comprise even larger structural diversity, including modified and nonnatural nucleobases, and nucleobase analogues. They can possess physical, chemical, biochemical, pharmacologic, and physiologic effects desired in biotechnology, medicine, and material chemistry, as for example: peptide nucleic acids,[5] highly specific receptors for base pairs,[6] manipulation of gene expression in the DNA supramolecular complex,[7] nonpolar nucleobases,[2,8,9] conductors/semiconductors of stacked nucleobases,[10] etc. Six years ago[11] there were only 350 nucleobase containing structures in the Nucleic Acid Database (NDB).[12] Today there are some 1500 entries in the NDB, over 400 in the NMR-Nucleic Acid Database,[13] around 800 in the DNA-Binding Protein Database,[14] over 1700 in the Protein Data Bank (PDB),[15] and over 500 in the Cambridge Structural Database (CSD).[16]

The main noncovalent forces important for stabilization of nucleic acids are aromatic $\pi ... \pi$ stacking interactions and hydrogen bonds.[3,17−20] The adjacent overlapping nucleobases in stack are mutually parallel at a vertical distance 3.3−3.6 Å[20] (double of the van der Waals radius for carbon, 1.70 Å[21]). In general, crystal packing effects can be around 0.01−0.02 Å for bond lengths.[22] The bond lengths between a given pair of atoms in similar environment are the same within the standard uncertainties of the measurements (estimated standard deviations, esds).[22] Standard esds for bond length and angles from high-quality crystal structure determination have reached 0.005 Å and 0.5°, respectively. Thus the quantification of substition and crystal packing effects on the standard molecular geometry is possible. Bond lengths in π-systems are a good measure of aromaticity, although not providing a complete idea on the aromaticiy of the system.[23,24] As the heteroaromaticity[25] is the aromaticity of heterocycles, all said above on bond lengths is valid for nucleobases also. Electron delocalization, conjugation, and hyperconjugation[26] are the major electronic factors making bonds to be partial double. The Resonance Theory[27−29] describes well these phenomena by resonance structures. From these structures, the Pauling π-bond orders, p_P, can be easily calculated,[27,29,30] and bond lengths d are expressed as a simple function of p_P (Bond Length-Bond Order Relationship, BLBOR).[29−33] Instead of p_P, some integer as the number of adjacent non-hydrogen atoms, n, can be used also.[34,35] Harmonic Oscillator Stabilization Energy approach (HOSE)[36] determines the weights of resonance structures and calculates corrected, weighted p_P.[37] BLBORs for planar benzenoids (planar benzenoid polycyclic aromatic hydrocarbons, PB-PAHs),[31,32,38] azabenzenoids,[31,32,37] diazabenzenoids,[33,37] poliazabenzenoids,[33,37] and picrates[31,32] exhibit the similarity among CC, CN, and CO π-bonds. There is practically one BLBOR (from linear regression, LR) for CC bonds in PB-PAHs and all their aza-derivatives and another for CN bonds in all aza-PAHs. BLBOR for CC bonds in picrates seems to be unique. Figure 1a shows that the mean CC bond length decreases from PB-PAHs to polyazabenzenoids with the number fraction of CN bonds (with respect

* Corresponding author phone: +55 19 3788 3102; fax: +55 19 3788 3023; e-mail: marcia@iqm.unicamp.br.

[†] Part of this work was presented at XI Simpósio Brasileiro de Química Teórica, November 18−21, 2001, Caxambu, MG, Brazil.

10.1021/ci0200523 CCC: $25.00 © 2003 American Chemical Society
Published on Web 04/19/2003

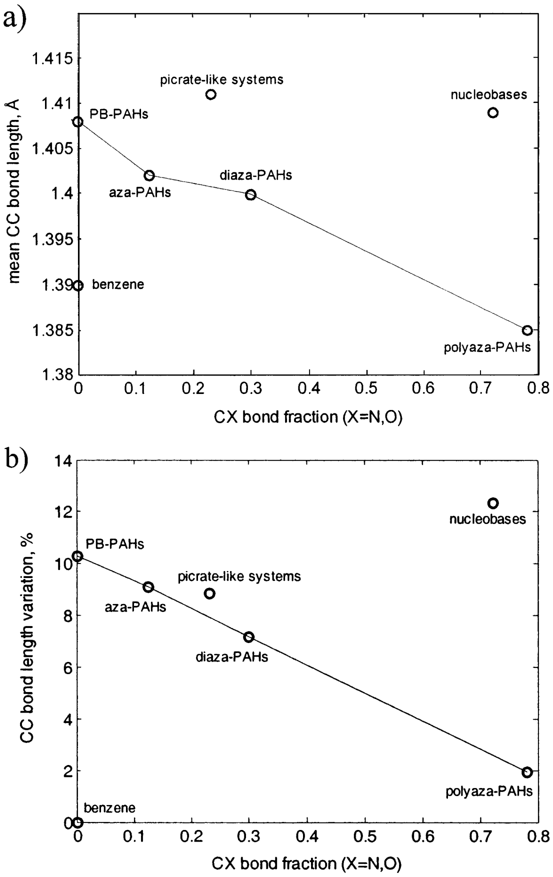

Figure 1. (a) The mean CC bond length (in Å) and (b) the CC bond length variation (in %) in various (hetero)aromatic classes, depending on the number fraction of CX (X = N,O) bonds.

to all bonds). The variation of d, (max d − min d)/max d, for CC bonds is 10% in PB-PAHs and only 2% in polyazabenzenoids; this variation depends linearly on the number fraction of CN bonds (Figure 1b). The CN bond length variation in aza-PAHs is 5−9%. Picrates exhibit CC and CO bond length variation as 9% and 8%, respectively; these CC bonds do not fit to the line in Figure 1b. Picrates represent π-systems with exocyclic C=O groups. Figure 1a,b reveals that CN and CO bonds affect CC bonds. CN bonds cause shortening and enhance equalization of CC bond lengths, CC aromatic character is more pronounced. CO bonds have the opposite effect on CC bonds (Figure 1). Nucleobases could have some properties in common with aza-PAHs and picrates.

Recently,[38] BLBOR for PB-PAHs was extended to multivariate analysis including bond orders corrected for the effects of intermolecular interactions in the crystalline state (from here on: crystal packing effects), n and topological indices of the bond neighborhood. Exploratory data analysis was performed by using Hierarchical Cluster (HCA) and Principal Component Analyses (PCA).[39] Parsimonius Multiple Linear Regression (MLR) and Partial Least Squares Regression (PLS)[39,40] models were built and bond lengths for some PB-PAHs were predicted satisfactorily. Julg's structural aromaticity index A[41,42] was calculated from the predicted bond lengths. Analogous study of experimental CC, CN, and CO bond lengths in nucleobases is presented in this work for the first time. The standard geometry data[43] for the neutral and protonated standard nucleobases (set I), C (**1**), HC$^+$ (**2**), T (**3**), U (**4**), A (**5**), HA$^+$ (**6**), and G (**7**) (Figure 2), were used. The geometry data for nonstandard and modified (natural and synthetic) nucleobases (set II), neutral and protonated **8**−**31** (Figure 2), were also used. The third set (set III, the prediction set) comprised nucleobases **32**−**50** (Figure 2) with simple molecular structure, low quality crystal structures, or no nucleoside crystal structure. Bond orders and topological indices were calculated, and the chemometric analysis was performed as for PB-PAHs.[38] Analytical curves,[27,29,32,37,38] molecular and quantum mechanics methods were employed for bond length calculations also. HCA and PCA were further applied to select the best methods to predict nucleobase bond lengths and some structural aromaticity indices. Quantum and molecular mechanics calculations on cytidine clusters were performed in order to deepen the knowledge about the relationships between intermolecular interactions and molecular properties of nucleobases. The list of frequently used mnemonics is provided in Table 1.

2. METHODOLOGY

a. Database Mining. The 1996 nucleic acid geometry standards by Berman et al.[11] and their Internet update[43] based on a survey of high-resolution small-molecule crystal structures contained in the CSD were the source of experimental bond lengths (the mean bond lengths with standard errors of the means[11]) for the standard nucleobases **1**−**7** (Table 2). Crystal structures of the corresponding nucleosides and their hydrochlorides (ribo- or deoxyribonucleosides, REFCODEs: CYTIDI02, DOCYTC, THYDIN01, BEURID, ADENOS01, ADOSHC, GUANSH10) were retrieved from CSD October 2001 release.[44] A list of REFCODEs with references is in the Supporting Information. The database mining for the structures containing nucleobases **8**−**31** (Table 2) satisfied the following criteria: crystallographic $R \leq 6.0\%$, esds on bond lengths ≤ 0.005 Å, publication year ≥ 1975, no disorder nor errors in crystal geometry. Structures of the following species were not retrieved: free nucleobases; nucleotides; nucleobases bound to metal; nucleosides without β-D-ribose or β-D-2′-deoxyribose; nucleosides with atoms other than C, H, N, O; nucleosides with other chemical bonds besides C−C, C−N, C−O, X−H (X = C,N,O), or with triple or partial triple bonds. Nucleobases **32**−**50** (Table 2, Figure 2) are simple aromatic and heteroaromatic (N, O) systems. Only structures of nucleosides RIBFIM, TAWMUZ, TUPQOK, FUJWOW (Table 1) are available for these nucleobases. The structures of **32**−**50** did not satisfy the searching criteria R factor, esds, and publication year and were substituted (BENZEN06, CALBOG) or bound to metal (SURLOG, SIQBAV, SEHXAE, INDYLI). The esds σ_{exp} for bond lengths were from literature, estimated as the mean of the σ range (Table 2), or calculated by PLATON.[45] Bond lengths corrected to thermal motion in crystal (libration correction) were from literature or calculated by PLATON. Available bond lengths for molecules in gas-phase[46] were also used. The bond lengths for the data retrieved from the CSD were measured by PLATON or Titan.[47] Structure TUPQOK had no atomic coordinates available, so taking σ = 0.005 Å and bond lengths g, h, i (Figure 2) from NAPHTA10, molecular graphics methods for determination of molecular dimensions[48] were applied on Figure 1 from Kool et al.[8]

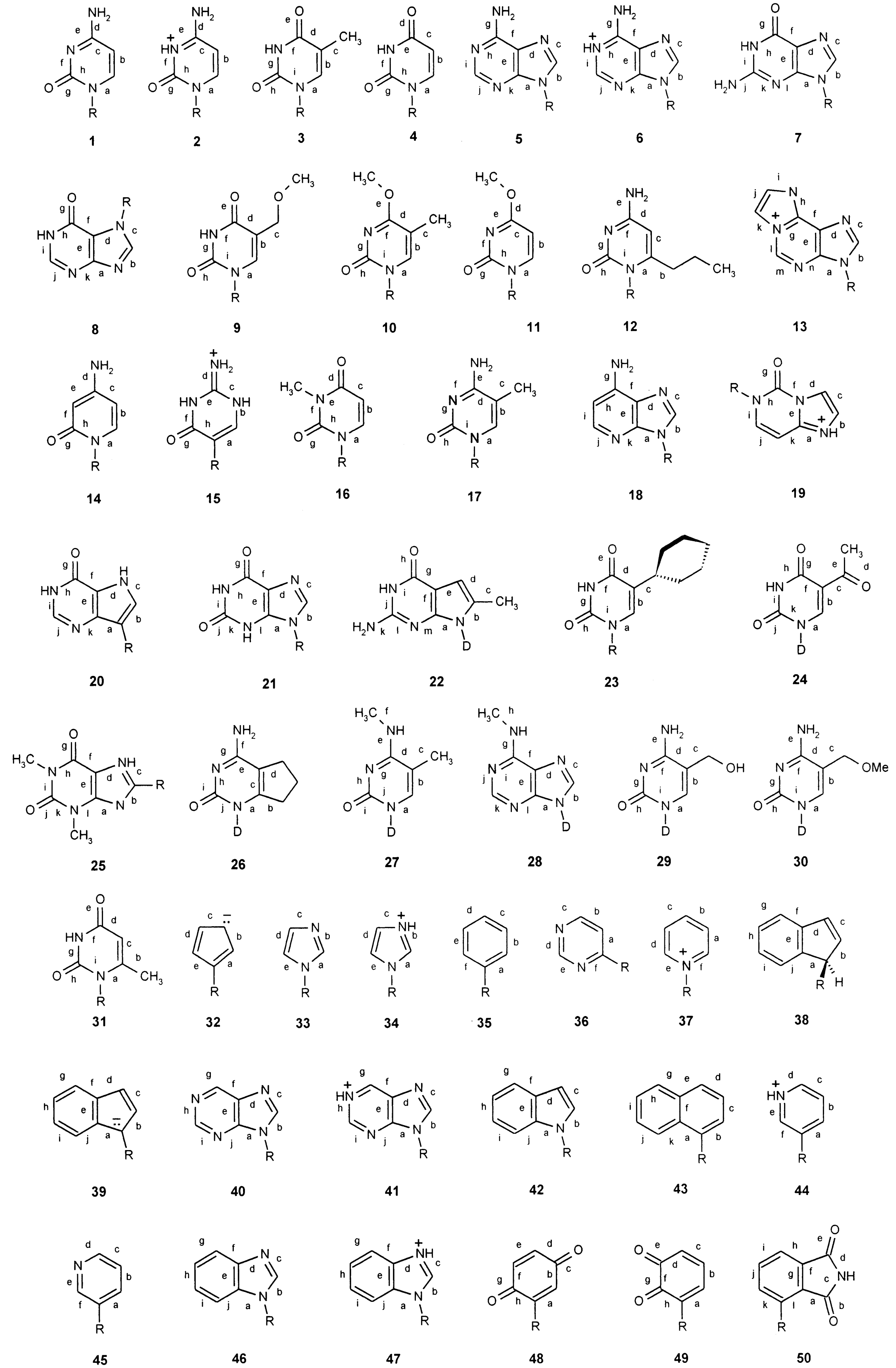

Figure 2. Formulas with the bond numbering for the studied nucleobases atached to ribose (R) or 2′-deoxyribose (D).

Table 1. Frequently Used Abbreviations in This Work

abbreviation	meaning
	Molecules
A, T, G, C, U, HA^+, HC^+	adenine, thymine, guanine, cytosine, uracil, protonated adenine, protonated cytosine
PAHs	fused polycylic aromatic hydrocarbons
PB-PAHs	planar benzenoid fused polycyclic aromatic hydrocarbons
Pu, Py	purine or purine-like nucleobases, pyrimidine or pyrimidine-like nucleobases
PuC, PyC	purine- or pyrimidine-like carbocycles
PuN, PyN	purine- or pyrimidine-like N-hetercocycles
PuO, PyO	purine- or pyrimidine-like O-heterocycles
	Data
NDB	The Nucleic Acid Database
PDB	The Protein Data Bank
CSD	The Cambridge Structural Database
	Concepts, Methods, Approaches
esd	estimated standard deviation
HOSE	The Harmonic Oscillator Stabilizing Energy approach
QSSR	Quantitative Structure-Structure Relationship
BLBOR	Bond Length-Bond Order Relationship
BLBDR	Bond Length-Bond Descriptor Relationship
QSBLR	Quantitative Structure-Bond Length Relationship
PHC	Pauling harmonic potential curve
PC	Principal Component
LR	Linear Regression
MLR	Multiple Linear Regression
PCA	Principal Component Analysis
HCA	Hierarchical Cluster Analysis
PCR	Principal Component Regression
PLS	Partical Least Squares Regression

b. Resonance Structures. Resonance structures were drawn in the light of the Resonance Theory, by means of chemical knowledge and measuring certain geometrical parameters: exocyclic substituent parameters (bond lengths, bond, and torsion angles), and the glycosilic bond parameters (bond lengths and torsion angles). Based on these measurements, it was assumed that sugar unit, Me bound to endocyclic N, and some other side groups do not affect the electron delocalization in the ring. The relevant resonance structures were drawn with the following assumptions. (1) There is no hyperconjugation through the glycosilic bond; (2) The delocalized electrons are from electron-rich groups: double CO and CC bonds; lone pairs from primary, secondary, or tertiary amine N atoms; CH and CO bonds participating hyperconjugation (as Me- in uracil and $-CH_2-$ in its derivatives); (3) 1−3 Kekulé structures for neutral molecules or cations with the positive charge at −NH− were drawn for each nucleobase in set I+II (Table 2); more Kekulé structures were drawn for nucleobases in set III. Ionic resonance structures included one charge separation in the most cases; double charge separation was for **27** and **28** (Table 2); (4) These ionic structures have positive charge at primary ($=N^+H_2$), secondary ($=N^+H$-), or tertiary ($=N^+$-(R)-) amine N; negative charge at carbonyl O, aromatic N ($-{:}N^-{=}$); and high electron density carbons at the point of fusion of the rings in purines (Pu). The resonance structures for pyrimidines (Py) **1**−**4** and Pu **5**−**7** are shown in Figures 3 and 4, respectively; (5) The side chain bonds in **9**, **13**, **16**, **23**, **25,** and **26** do not contribute to electron delocalization (see Figure 2); (6) The carbon atoms at the point of fusion of the rings in Pu are negatively charged or neutral; (7) Hyperconjugation exists for CH and CO side chain bonds perpendicular or significantly inclined to the nucleobase ring. Such hyperconjugation has been already detected as a significant deviation of corresponding bond lengths from X-ray and neutron diffraction standards.[49] Thus the measurement of geometrical parameters revealed which resonance structures are important. Pu nucleobases have in average more resonance structures (8−17) than Py (4−10), Table 2. The ionic structures are prevalent and necessary for calculation of π-bond orders:[27,29,30] p_P was calculated as $p_P = C_d/C$, C is the number of all resonance structures, and C_d is the number of structures in where a particular bond appears as double. Atomic negative charge population n_p was defined as $n_p = C_n/C$, C_n is the number of the resonance structures in which a particular atom is in the electron-rich state. The electron-rich and electron-poor states are defined for these atoms: carbonyl O (neutral in C=O and negative in $C-O^-$), ether O (neutral in −C−O−C and negative in $-C-O^{-}\cdots C^{+}$), aromatic N (neutral as $-N{:}=$ and negative as $-{:}N{:}^{--}$), aliphatic N (positive as $=N^+(X)(Y)$ and neutral as $-{:}N(X)$-(Y)), and tertiary C at point of fusion of the Pu rings (neutral as −C(X) = and negative in $-{:}C^-(X)$-). $C_n = 0$ for an atom with zero nonbonded electrons both in ground and excited state. $C_n = 1$ if unsaturated carbon in a certain structure becomes saturated due to hyperconjugation. The sum of all p_P and n_p around an atom is one. In alternant PAHs Kekulé structures are sufficient (no charge separation) and the sum of all p_P and n_p is one ($n_p = 0$).

c. Bond Length-Bond Force Relationships. Coefficients a and b from $d/\text{Å} = a + b\,f$ where f is the force constant, and single (s_0) and double (d_0) bond lengths for CC, CN, and CO bonds, had to be determined for the HOSE approach. Data for s_0 and d_0 were from gas-phase structure determinations[46] (Table 3) and (a, b) from Krygowski et al.[36] The other (a, b) set was obtained by LR on bond parameters from Cornell et al.[50] The third (a, b) set was from Dewar and Gleicher.[51] These three (a, b) sets are called *kr*, *co*, and *dg* sets (Table 3). The sets are based on works developed by three independent groups working in different areas (crystallography and structural chemistry,[36] semiempirical methods,[51] molecular mechanics[50]) and times (the mid 1960s, mid 1980s,

Table 2. CSD Crystal Structure and Resonance Structure Data for Studied Nucleobases

	nucleobase[a]	NB[b]	source[c]	R[d]	$10^3\ \sigma$/Å[d]	year[d]	C[e]	K[e]	I[e]	D[e]	H[e]	B[e]
1	cytosine	Py	ref [11]	<0.06	1−2	1996	6	1	5	0	0	0
			ref [43]	<0.06	1−2	2000						
			CYTIDI02	0.054	1−5	1995						
2	protonated cytosine	Py	ref [11]	<0.06	1−2	1996	8	3	5	0	0	0
			ref [43]	<0.06	1−2	2000						
			DOCYTC[f]	0.035	3	1970						
3	thymine	Py	ref [11]	<0.06	1	1996	7	1	6	0	2	0
			ref [43]	<0.06	1−2	2000						
			THYDIN01	0.044	1−5	1995						
4	uracil	Py	ref [11]	<0.06	1	1996	5	1	4	0	0	0
			ref [43]	<0.06	1	2000						
			BEURID10[f]	0.033	3−4	1975						
5	adenine	Pu	ref [2]	<0.06	1	1996	10	2	8	0	0	0
			ref [43]	<0.06	1	2000						
			ADENOS01	0.024	2	1991						
6	protonated adenine	Pu	ref [11]	<0.06	1−2	1996	10	3	7	0	0	0
			ref [43]	<0.06	1−2	2000						
			ADOSHC[f]	0.037	4−5	1973						
7	guanine	Pu	ref [11]	<0.06	1−2	1996	10	1	9	0	0	0
			ref [43]	<0.06	1−2	2000						
			GUANSH10[f]	0.036	6	1970						
8	hypoxanthine	Pu	ZOZXOB	0.027	1−2	1996	8	1	7	0	0	0
9	5-methoxymethyluracil	Py	FUXBIJ01	0.017	1	1994	6	1	5	0	1	2
10	O^4-methylthymine	Py	DOXPOV	0.042	4−6	1986	9	1	8	0	4	0
11	O^4-methyluracil	Py	CEFJUS	0.028	3−5	1983	6	1	5	0	3	0
12	6-propylcytosine	Py	SECKEQ	0.029	2−4	1998	8	1	7	0	2	2
13	protonated 1,N^6-ethenoadenine	Pu	BIMFIM	0.045	1−5	1984	16	2	14	0	0	8
14	3-deaza-cytosine	Py	DAZCYT10	0.035	2	1977	4	1	3	0	0	0
15	protonated pseudo-isocytosine	Py	PSCYTD	0.040	4−5	1980	7	3	4	0	0	1
16	3-methyluracil	Py	ZAYTIC	0.035	3−4	1995	5	1	4	0	0	1
17	5-methylcytosine	Py	TALJAR	0.036	3−5	1991	8	1	7	0	2	0
18	1-deaza-adenine	Pu	DEHQOW	0.039	3−4	1999	9	2	7	0	0	0
19	protonated 3,N^4-ethenocytosine	Py	ETCYTC	0.045	1−5	1976	7	1	6	0	0	0
20	9-deaza-hypoxanthine	Pu	VOVJIZ	0.042	1−5	1992	9	1	8	0	0	0
21	xanthine	Pu	CUTVAO	0.039	1−5	1984	10	1	9	0	0	0
22	8-methyl-7-deazaguanine	Pu	NEDDIJ	0.043	2−3	1997	13	1	12	0	0	0
23	5-cyclohexyluracil	Py	PULVIB	0.052	1−5	1996	6	1	5	0	1	5
24	5-acetyluracil	Py	ACURID	0.034	2−4	1980	8	1	7	0	2	0
25	1,3-dimethylxanthine	Pu	KABVEO	0.042	4−6	1987	10	1	9	0	0	2
26	4-amino-6,7-dihydro-1*H*,5*H*-*c*yclopentapyrimidine-2-one	Pu	TEJNIF	0.049	4−5	1996	11	1	10	0	4	2
27	N^4-5-dimethylcytosine	Py	SEDQEX	0.036	2−3	1998	10	1	7	2	4	0
28	N^6-methyladenine	Pu	DEFPOT	0.037	3	1985	17	2	9	6	6	0
29	5-hydroxymethylcytosine	Py	HEVXOV	0.030	2−3	1994	7	1	6	0	1	0
30	5-methoxymethylcytosine	Py	VEXDOR	0.044	3−4	1990	7	1	6	0	1	0
31	6-methyluracil	Py	MEDOUR	0.044	3−5	1980	7	1	6	0	2	0
32	cyclopentadienyl	PyC	SURLOG	0.036	1−5	1995	5	5	0	0	0	0
33	imidazole	PyN	IMAZOL06	0.026	1	1979	3	1	2	0	0	0
			RIBFIM[f]	0.030	3	1973						
34	imidazolinium	PyN	SIQBAV[f]	0.043	6−10	1998	5	2	3	0	0	0
35	benzen	PyC	BENZEN06	0.036	1	1987	2	2	0	0	0	0
			TAWMUZ	0.055	3	1996						
36	pyrimidine	Py	PRMDIN01	0.042	2	1979	2	2	0	0	0	0
37	pyrimidinium	Py	SEHXAE[f]	0.038	11−30	1997	2	2	0	0	0	0
38	indene	PuC	CALBOG	0.031	4	1998	4	2	2	0	0	0
39	indenyl	PuC	INDYLI	0.055	1−5	1975	6	6	0	0	0	0
40	purine	Pu	PURURE[f]	0.075	4	1977	8	2	6	0	0	0
41	purinium	Pu	CLPRCV[f]	0.040	6−010	1981	7	2	5	0	0	0
42	indole	PuN	ZIRFOV02	0.050	3−5	1997	4	2	2	0	0	0
43	naphthalene	PuC	NAPHTA10	0.035	2	1983	3	3	0	0	0	0
			TUPQOK[f]	0.039	13−18	1996						
44	pyridinium	Py	JOZCIK	0.035	3	1992	2	2	0	0	0	0
45	pyridine	Py	PYRDNA01	0.044	3	1981	2	2	0	0	0	0
			FUJWOW	0.031	5	1987						
46	benzoimidazole	PuN	BZIMBF10	0.050	3	1976	6	2	4	0	0	0
47	benzoimidazolinium	PuN	BZIMBF10	0.050	3	1976	6	4	2	0	0	0
48	*p*-benzoquinone	PyO	BNZQUI02[f]	0.074	3	1978	4	1	3	0	0	0
49	*o*-benzoquinone	PyO	OBNZQU[f]	0.039	2−4	1973	4	1	3	0	0	0
50	phthalimide	PuO	PHALIM01	0.038	3−4	1992	6	2	4	0	0	0

[a] The nucleobase name. [b] TNB: The nucleobase type: Py/Pu, PyC/PuC, PyN/PuN, PyO/PuO. See Table 1. [c] The CSD retrieval: literature search[11,43] and the search in this work (REFCODEs are in italics for data not used in further chemometric analysis). [d] The quality indices of the crystal structure data: R, the range of esds σ, and the year of publication. [e] The resonance structure data: C − the total number of resonance structures, K − the number of Kekulé structures, I − the number of ionic structures with one ion pair, D − the number of ionic structures with two ionic pairs, H − the number of resonance structures with hyperconjugation, B − the number of bonds not included in the electron delocalization. [f] The structures which do not satisfy all the searching criteria.

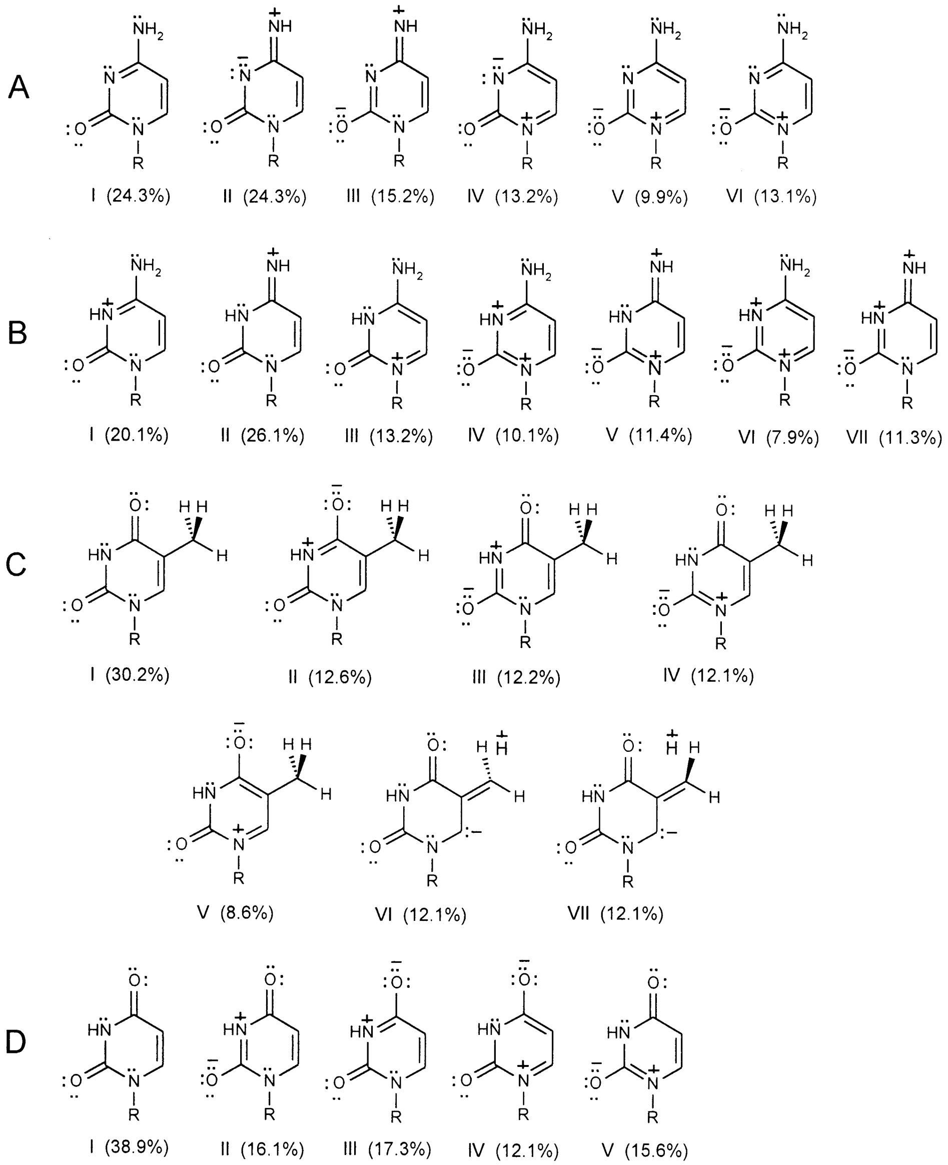

Figure 3. Resonance structures of nucleobases **1** (A), **2** (B), **3** (C), and **4** (D) in ribonucleosides. Electrons from double bonds and free pairs, centers of positive and negative charge are shown. The number and weight of the resonance structures are also given.

mid 1990s). It is quite surprising that *a* correlates extremely highly with *b* (corr coeff −0.992) when all the data are included, so *kr*, *co*, and *dg* are equivalent sets and CC, CN, and CO bond are instrinsically similar. Cornell et al. introduced an improved version of AMBER force field, which was successfully designed primarily for proteins and nucleic acids.[51−54]

d. The Extended HOSE Model. The extended HOSE was reported by Krygowski et al.[36] as

$$E_k = c[\Sigma_i(s_i - s_0)^2(a_1 + b_1 s_i) + \Sigma_j(d_j - d_0)^2(a_2 + b_2 d_j)],$$
$$w_k = E_k^{-1}/(\Sigma_k E_k^{-1}), \quad E^{-1} = C^{-1}(\Sigma_k E_k^{-1})$$

where E_k − the HOSE energy of the *k*th resonance structure; c − a constant; a_1, b_1, a_2, b_2 − empirical constants from $d - f$ LR models (Table 3); s_i, d_j − the bond lengths of all the bonds which appear as single or double, respectively, in the *k*th resonance structure; s_0, d_0 − the standards for *kr*, *co*, *dg* sets (Table 3); w_k − the weight of the *k*th resonance structure; and E − the total HOSE energy of the resonance hybride. Local F77 program[55] was used to calculate these quantities with errors.[37] The weighted Pauling π-bond order p_w was calculated as $p_w = \Sigma_l w_l$ where w_l are the weights of the canonical structures in which the bond appears as double. Using the three sets *kr*, *co*, and *dg*, three bond orders p_w (p_{wkr}, p_{wco}, p_{wdg}) were obtained.

e. Molecular Mechanics and Quantum Chemical Calculations. Conformational search and geometry optimization by Titan at molecular mechanics (MMFF94[56]), semiempirical (MNDO,[57] AM1,[58] PM3[59]), and HF ab initio (6-31G**) level were peformed on experimental/modeled structures for all nucleosides.

f. Incorporation of Crystal Packing Effects into Bond Orders. According to the current knowledge on intermolecular interactions recorded in the CSD,[60] nucleoside crystals are built by stronger hydrogen bonds (common and resonance assisted), weak hydrogen bonds as C−H...X or X−H...π (X = C, N, O), interactions not mediated by hydrogen (as π...π stacking interactions), and van der Waals interactions. A few strong directional interactions or many weak interactions may influence bond length directly as bond length stretching or indirectly through torsional effects. Even

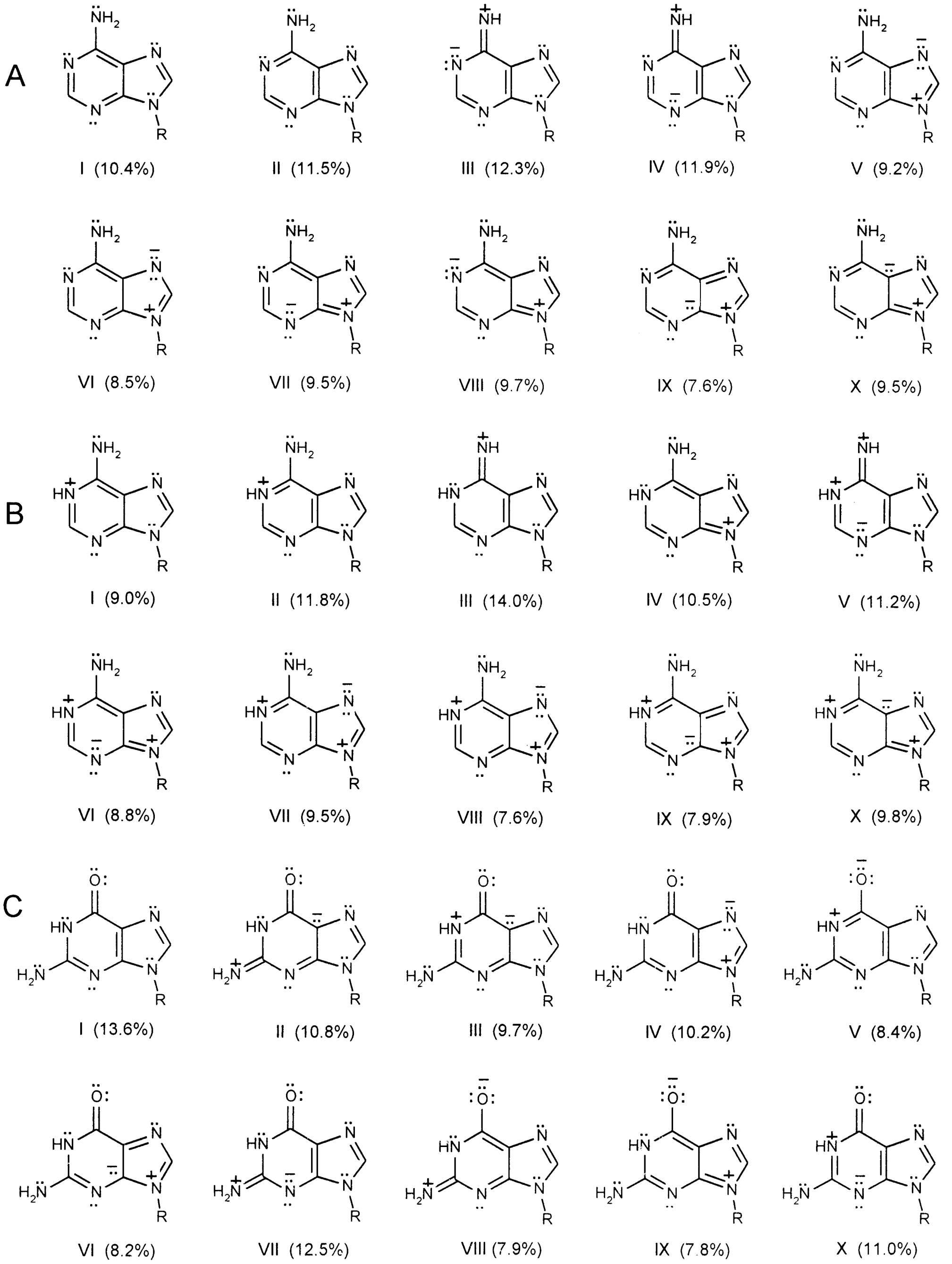

Figure 4. Resonance structures of nucleobases **5** (A), **6** (B), and **7** (C) in ribonucleosides. Electrons from double bonds and free pairs, centers of positive and negative charge are shown. The number and weight of the resonance structures are also given.

strong carbonyl bond suffers bond length change >0.01 Å due to stronger hydrogen bond.[61−63] Weak crystal packing effects on CC bond lengths in PB-PAHs have beeen observed[38] being <0.01 Å and quantified in a way which is not applicable for nucleosides. There are two reasons why to include crystal packing effects for nucleobases. First, in the case of resonance-assisted hydrogen bonds, there is a mutual influence of hydrogen bonding and conjugation of heteroaromatic systems,[64] which might not be detected in bond lengths due to large esds.[65] Second, crystal packing effects on bond lengths in nucleoside crystals, although being much smaller than intrinsic electron delocalization in the nucleobase rings, can be statistically significant and determine the best models for predicting bond lengths.[38] Maximum foreshortening of CC, CN, and CO bond lengths in structures determined by X-ray diffraction is 0.01 Å.[66] In this work, these foreshortenings are included in variables for crystal packing effects, and throughout chemometric techniques in large part, if not completely, are eliminated. The average esd in set I+II is 0.003 Å, which can be used as the limit for the packing effects. As s_0 - d_0 (Table 3) corresponds to the increment of p_P, $\Delta p = 1$, 0.002 Å is equivalent to $\Delta p > 0.01$ for crystal packing effects. The minimum HF-experimental bond length differences in **1**−**7** were assumed to be 0.005, 0.015, and 0.025 Å for CC, CN, and CO bonds, respectively, and to increase with *n* as *n*-tuple multiples (Table 3). p_{wco} exhibited practically equal correlation with experimental *d* ($r = -0.7092$) as p_{wkr} ($r = -0.7087$) and p_{wdg} ($r = -0.7070$) and a slightly better fit to *d*. Comparing p_P and p_{wco} ($\Delta p > 0.01$) for bonds in set I+II, a systematic

Table 3. HOSE[a,b] and Crystal Packing Correction[c] Parameters

bond	s_0/Å	d_0/Å	$a/10^4$ Pa	$b/10^4$ Pa	Δp			
					$n = 4$	$n = 3$	$n = 2$	$n = 1$
CC	1.535	1.339	44.39	−26.02	0.0255	0.0510	0.0765	0.1020
	H_3C-CH_3	$H_2C=CH_2$	34.40	−19.86				
			52.88	−32.23				
CN	1.471	1.276	43.18	−25.73	0.0760	0.1026	0.1282	0.1538
	H_3C-NH_2	$H_2C=N(OH)$	25.41	−14.01				
			57.42	−36.95				
CO	1.425	1.208	52.35	−32.88		0.1382	0.1070	
	H_3C-OH	$H_2C=O$	41.27	−25.92				
			60.60	−40.49				

[a] The lengths of single (s_0) and double (d_0) bonds from gas-phase structure determination of small molecules.[46] [b] The three sets of (a, b) constants scaled to Pa units (one above the other): *kr*[36] (plain text), *co*[50] (bold), *dg*[51] data set (italics). [c] Bond order increments [51]$\Delta p = c_0\ (h - n)/(s_0 - d_0)$, c_0 is the minimal observed crystal packing effect on CX (X = C,N,O) bond length, and h is a multiplicity constant ($h = 5$ for CC, CN; $h = 4$ for CO).

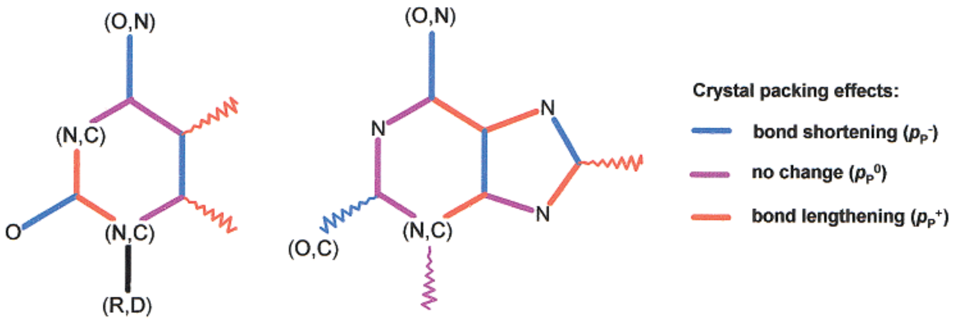

Figure 5. Crystal packing effects observed for Py (left) and Pu (right) in sets I+II. Solid bonds are common for all nucleobases, and wave bonds just for some. Two possible atom types are in parentheses.

predominant distribution of shortened ($p_P^- = p_P + \Delta p$), lengthened ($p_P^+ = p_P - \Delta p$), and unchanged ($p_P^0 = p_P$) bonds was obtained (Figure 5). The bonds not characterized in Figure 5 are considered as unchanged ($p_P^0 = p_P$). For $p_P^+ < 0$, it was set $p_P = 0$. The Pauling π-bond order corrected to maximum packing effects, $p_s = p_P^-$, p_P^+ or p_P^0, was defined and applied to BLDORs for set I+II. The bond order including maximum packing effects, $p_{cr} = p_P + \Delta p$ if $p_{wco} - p_P > 0.01$; $p_{cr} = p_P - \Delta p$ if $p_P - \mathrm{p}_{wco} > 0.01$; $p_{cr} = p_P$ if $|p_{wco} - p_P| \leq 0.01$, was calculated. The average packing effects were incorporated into bond order p_m, (equal to p_{cr}, with $\Delta p/2$ instead of Δp). p_{cr} and p_m, according to their definition could not be calculated for set III.

g. Other Chemical Bond Descriptors. Q for bond A−B was defined as $Q = Q(A) + Q(B)$ where $Q(X)$ is the atomic number of atom X. Bond descriptor n was counted as the number of non-hydrogen atoms directly attached to a particular bond and shown to be useful in predicting lengths in PB-PAHs.[38]

h. Pauling and Gordy Curves. Ratios $t = f_d/f_s$, with force constants f_s and f_d for single and double bond, respectively, were used for Pauling harmonic potential curves (PHC)[27,29] $d = s_0 - (s_0 - d_0)tp/(tp + 1)$; $p = p_P$ or p_s. Regression Pauling logarithmic curves,[67] $d = a + b \log(p + 1)$, and Gordy's curves,[68] $d = a + b(p)^{-1/2}$, $p = p_P$ or p_s, were also studied for CC, CN, and CO bonds. All sets *kr*, *co*, *dg* were used.

i. Exploratory Data Analysis. Data set (p_P, p_{wkr}, p_{wco}, p_{wdg}, p_{cr}, p_m, p_s, n, Q) was autoscaled. Three analyses were performed, each one employing only one type of p_w, by using PCA and HCA[39] by means of Pirouette 3.01.[69] Weighted normal varimax rotation was performed on all Principal Components (PCs) to obtain PCs interpretable in terms of bond orders, n and Q. Single linkage was used for HCA.

j. Building and Validation of Regression Models. Data for 309 bonds were autoscaled for PLS and PCR and not for LR and MLR models. Sets I and I+II were training sets in 34 regression models which included n, Q, and π-bond orders: only one (p_P or $\mathrm{p_s}$), two (p_P and p_s), or five (p_P, p_s, p_{cr}, p_m, p_w where $p_w = p_{wco}$, p_{wkr}, or p_{wdg}). All the models were validated by leave-one-out cross-validation procedure. Three-variable models (p, n, Q) were treated with MLR, PLS, and PCR (with 3 PCs), giving the same results as theoretically expected. Models without the variables based on experimental data (p_{cr}, p_m, p_w) could be used for prediction of bond lengths. Other models are useful for discussion on nucleobase properties. LR and MLR were performed utilizing Matlab 6.1[70] and PCR and PLS by Pirouette 3.01.[69]

k. HCA-PCA Selection Procedures for the Best Prediction Models. To take into account the complexity of finding the best model for calculating bond lengths, statistical parameters (the first 15 in Table 4) were calculated.[38,71−74] For LR and PHC is $k = 1$, and also for molecular and quantum mechanics results (bond length from a computational method can be considered as one generated variable). The average number of C, N, and O atoms that define the bonds under study (9.55 per nucleobase), multiplied by the number of parameters required for computational method (three atomic coordinates and the atom type) and reduced by 6 (three translational and three rotational degrees of freedom), gives $NV = 57$. This way all models were compared in terms of their results or common statistical descriptors. In a previous study[75] it has been demonstrated how HCA and PCA can help to find out the best model, by

Table 4. Statistical Parameters Used in This Work

parameter[a]	symbol	expression[b]
maximum bond length deviation	Δ_{max}	$\Delta_{max} = \lvert d_{exp} - d_{cal}\rvert_{max}$
average bond length deviation	$\langle\Delta\rangle$	$\langle\Delta\rangle = 1/N\, \Sigma_i\, \lvert d_{exp,i} - d_{cal,i}\rvert$
average deviation-error ratio	$\langle\Delta/\sigma_{exp}\rangle$	$\langle\Delta/\sigma_{exp}\rangle = 1/N\, \Sigma_i\, (\lvert d_{exp,i} - d_{cal,i}\rvert/\sigma_{exp,i})$
LR coefficients	a, b	$d_{cal}/\text{Å} = a + b\, d_{exp}/\text{Å}$
minimum LR/MLR t-parameter	t	$[c_i/\sigma(c_i)]_{min}$
correlation coefficient from calibration	R	$R^2 = 1 - \Sigma_i\, (d_{cal,i} - d_{exp,i})^2/\Sigma_i\, (d_{exp,i} - \langle d\rangle)^2$
correlation coefficient from validation	Q	$Q^2 = 1 - \Sigma_i\, (d_{val,i} - d_{exp,i})^2/\Sigma_i\, (d_{exp,i} - \langle d\rangle)^2$
PRESS for calibration	pressc	$pressc = \Sigma_i\, (d_{cal,i} - d_{exp,i})^2$
PRESS for validation	pressv	$pressv = \Sigma_i\, (d_{val,i} - d_{exp,i})^2$
standard error of calibration	SEC	$SEC = [pressc/(N - k - 1)]^{1/2}$
standard error of validation	SEV	$SEV = [pressv/(N - k - 1)]^{1/2}$
weighted F-ratio	F	$F = R^2\, (N - k - 1)/[k(1 - R^2)]$
weighted FIT parameter	FIT	$FIT = R^2\, (N - k - 1)/[(N + k^2)(1 - R^2)]$
true no. of variables and parameters	NV	$NV = 1, 4, 57$ or k
standard deviation of bond lengths	σ	$\sigma^2 = \Sigma_i\, (d_i - \langle d\rangle)^2/(N - 1)$
HOMA deviation	$\Delta HOMA$	$\Delta HOMA = \lvert HOMA_{exp} - HOMA_{cal}\rvert$
average bond length	$\langle d\rangle$	$\langle d\rangle = 1/N\, \Sigma_i\, d_i$
Julg's index	A	ref 36
Julg's index error	$\sigma(A)$	ref 37
Julg's index deviation	ΔA	$\Delta A = \lvert A_{exp} - A_{cal}\rvert$

[a] PRESS − Predictive Residual Error Sum of Squares; HOMA − Harmonic Oscillator Measure of Aromaticity, see text. [b] Other symbols: d_{exp}, d_{cal}, d_{val}, d_i − experimental, calculated from calibration, calculated from validation, and any bond length, respectively; σ_{exp} − experimental esd; c_i, $\sigma(c_i)$ − any regression coefficient from LR or MLR and its statistical error, respectively; N − the number of bonds; k − the number of parameters: being 1 for PHC and LR, molecular and quantum mechanics (calculated bond length considered as one generated variable); NV equals to 4 for PHC, 57 for molecular and quantum mechanical models, and k for other models; A_{exp}, A_{cal} − experimental and calculated Julg's index. The error of is defined as $\sigma(\Delta A) = (\sigma(A_{exp})^2 + \sigma(A_{cal})^2)^{1/2}$.

treating the statistical parameters as molecular descriptors. Recently,[62,63] bond lengths and other structural parameters were treated by PCA and HCA. This approach was applied in this work. HCA (single linkage) and PCA (standard and weighted normal varimax rotated) were performed on autoscaled statistical parameters (32 × 11 matrix) and bond lengths (309 × 33 matrix). The best models were compared to analogous LR, MLR, PCR, and PLS models for data for one bond type (CC, CN, CO).

l. Julg's and Other Structural Aromaticity Indices. Julg's structural aromaticity index A[41,42] and its error $\sigma(A)$[38] and the average bond length $\langle d\rangle$ were calculated from d_{exp}, d_{cal}, and esds (from the previous 32 models) for **1**−**31**. d_{exp} were not corrected to librational motion due to minor contribution of libration to A.[38] Chemometric analysis of the three data matrices 31 × 32 (for A, $\sigma(A)$, and $\langle d\rangle$) and the matrix 309 × 32 (for d_{exp}, d_{cal}) was carried out by means of HCA-PCA procedure just described, to qualify the heteroaromaticity of nucleobases and to establish an additional method for finding the best prediction model for structural indices of aromaticity.

m. Molecular and Quantum Mechanics Estimation of Packing Effects on Nucleobase Aromaticity. It has been noticed[37] that structural and electronic properties significantly differ for a molecule in a free state and in van der Waals or charge-tranfer cluster. Krygowski[76] observed that hydrogen bond and other crystal forces significantly affect *HOMA* (Harmonic Oscillator Measure of Aromaticity) and other aromaticity indices. The geometry optimization of molecular complexes could aid in quantifying these effects. Crystal structure for this purpose was CYTIDI02, prepared by PLATON[45] as an isolated molecule and as clusters of 2−6 cytidine molecules. The geometry was optimized at MMFF94, MNDO, AM1, and PM3 level using Titan[47] and for dimers at HF and B3LYP level (6-31G** set) by Gaussian 98.[77] The bond lengths of the central molecule were studied in terms of statistics: Δ_{max}, $\langle\Delta\rangle$, σ, $\langle\Delta/\sigma_{exp}\rangle$, $\Delta HOMA$, ΔA (Table 4). The fourth C residue from the 3′-terminus in B-DNA double helix 5′-d(CpCpApGpTpApCpTpGpG)-3′ retrieved from NDB[12] (structure BD0023:[78] $R = 10.5\%$, res. 0.74 Å) was compared with C from CYTIDI02 and from the ab initio geometry optimized dimers.

3. RESULTS AND DISCUSSION

a. Database Mining Results. The standard nucleobases **1**−**7** (set I) were retrieved from CSD as well as 24 high precision crystal structures of natural and synthetic nucleobases in nucleosides (set II). Nineteen nucleosides (set III) were retrieved as crystal structures of nucleobases; 13 are high precision structures. Four of them are in the form of crystal structures of nucleosides, from which only two are of high quality (TAWMUZ and FUJWOW, Table 2). Seventeen from 50 nucleobases have bond length data corrected to libration motion in crystal. Gas-phase data were found only for three nucleobases (**32**, **35**, **37**). Nucleobases **1**−**31** have 2−5 nitrogen and 0−3 oxygen atoms participating π-electron delocalization, being similar to triaza-PAHs and oxi-PAHs. There are 21 Pu (set I: 3; set II: 9; set III: 9) and 29 Pu nucleobases (set I: 4; set II:15; set III:10). Set III consists of five carbocycles (PyC, PuC), 4 *N*- (PyN, PuN) and three *O*-heterocycles (PyO, PuO), neutral and protonated purine and pyrimidine. Altogether there are 463 delocalized bonds (set I: 87; set II: 222; set III: 154), from which 188 are CC (set I: 20; set II: 66; set III: 102), 231 CN (set I: 57; set II: 128; set III: 46), and 44 CO (set I: 10; set II: 28; set III: 6), normally distributed around their means. High-precision CC bond lengths in set I+II are characterized by relatively large variation (12%) and mean (1.409 Å). Nucleobases are differentiated from aza-PAHs series and picrates due to these characteristics (Figure 1a,b). Thus the high exocyclic CX (X = N,O) and endocyclic CN bond fraction causes enhanced bond length alternation and the decrease

Table 5. Correlation Statistics for Nucleobases[a]

descriptor	Set I	Set II	Set III	Set I+II	Set I+II (CC)	Set I+II (CN)	Set I+II (CO)
p_P	−0.634	−0.657	−0.443	−0.653	−0.812	−0.736	−0.546
p_{wkr}	−0.705	−0.710	−0.650	−0.709	−0.833	−0.806	−0.799
p_{wco}	−0.704	−0.711	−0.649	−0.709	−0.837	−0.806	−0.799
p_{wdg}	−0.704	−0.708	−0.646	−0.707	−0.839	−0.804	−0.797
p_{cr}	−0.713	−0.706	−0.550	−0.706	−0.836	−0.823	−0.718
p_m	−0.686	−0.690	−0.510	−0.689	−0.826	−0.800	−0.648
p_s	−0.711	−0.709	−0.416	−0.708	−0.837	−0.759	−0.568
n	0.596	0.479	0.420	0.502	0.0229	0.467	-
Q	−0.750	−0.756	−0.688	−0.754	-	-	-

[a] Nonexisting when a descriptor has only one or two distinct values.

of π-electron delocalization with respect to aza-PAHs. Similarly, the CO and CN bond length variation (10% and 11%, respectively) is greater than those in aza-PAHs and in picrates. Two bond lengths d_1 and d_2 from crystal structure are considered not to be significantly different (or to be "equal") at 0.99 probability level (normal distribution of bond lengths in crystal is assumed) if $q = |d_1 - d_2|/[\sigma^2(d_1) + \sigma^2(d_2)]^{1/2} < 2.58$, and $\sigma(d_1)$ and $\sigma(d_2)$ are esds for d_1 and d_2, respectively.[79] In general, due to underestimated esds, $q < 5.0$ is recommended.[65] No difference can be observed between bond lengths from Berman et al. in 1996[11] and 1999[43] (set I, Table I in Supporting Information). Significant differences exist between these surveys and nucleoside crystal data (CYTIDI02, DOCYTC, THYDIN01, BEURID, ADENOS01, ADOSHC, QUANSH10) mainly due to substitution effects. The maximum observed thermal corrections in set I−III are 0.003, 0.005, and 0.012 Å for CC, CN, and CO bonds, respectively, larger than in PB-PAHs (0.003 Å[38]). The both length shortening/lengthening between benzene/pyridine and their ribonucleosides can be also observed (Table 1 in Supporting Information). Bond lengths in gas-phase differ from those from X-ray diffraction for cyclopentadienyl. Crystal packing effects cannot be statistically observed in every case, contrary to substitution effects.

b. Data Set Degeneration and Correlation Statistics. LR Models. CC, CN, and CO bonds in PB-PAHs, aza-PAHs, and picrates[32,35,38] exhibit degeneracy i.e. the phenomenon that a certain value of a bond descriptor corresponds to significantly different values of experimental bond lengths. Bonds with such property form a degenerative group. Introducing crystal packing effects as p_{cr}, topological indices n, m (the number of benzenoid rings around a particular bond) and l (the number of neighboring atoms around those defined by n), the degeneracy in PB-PAHs was substantially reduced. The degeneracy in nucleobases (Tables I−III in Supporting Information) can be observed also. If degeneration groups with more than five bonds are considered, 82% CO, 87% CC, 94% CN bonds belong to only 10, 17, and 27 such groups. CO bonds are rather short (partially double to double), because $p_P = 0.43-0.80$. CC bonds are concentrated in quite a broad region, varying from pure double to pure single bond ($p_P = 0.10-0.80$). The variation of CN bonds is also large ($p_P = 0.08-0.88$). The degenerations mainly occur at partial double to single bonds (CO), double to partial double (CC), or with no preference (CN). Crystal packing corrections in bond orders reduce although do not eliminate the degeneracy. There is also degeneracy of bond lengths with respect to other descriptors: 45%, 24%, 30%, 1% bonds have $n = 4$, 3, 2, 1, respectively; 28%, 60%, and 12% bonds correspond to $Q = 12$, 13, and 14, respectively.

The correlation analysis for BLBDRs (Table 5) confirms the findings of the degeneracy analysis. In general, pronounced degeneracy corresponds to low correlation. Among all p_w orders none is preferred. p_m is less correlated to d than p_{cr} and p_s. Q-d relationship has the highest correlation. Sets I and II do not differ in correlation coeffiecients (except for p_P and n, Table 5), which enables them to be treated as one set. Set III shows lower correlations systematically, due to low quality data and heterogenicity. p_P, n, and p_{cr} for PB-PAHs were higher correlated to d than for nucleobases: −0.895 (p_P), −0.929 (p_{cr}), and 0.735 (n). Table 3 contains also the correlation coeffiecients regarding only CC, CN, and CO bonds from set I+II. None of the coefficients is greater than 0.84, while they reached 0.94 for PB-PAHs.[38] From the correlation analysis emerge multivariate QSBLR models instead of LR models.

c. Exploratory Data Analysis. Bond Classification. Figure 6 shows distribution of bond type (referring to Q), crystal packing (counting for the $p_{wco} - p_P$ difference) and neighborhood effects (included in n) on bond lengths. Three distinct regions pertaining to CC (gray), CN (green), and CO bonds (yellow) are visible. The CC−CN and CN−CO mixing is mostly due to formally single bonds participating in hyperconjugation: two CO (Figure 2: e in **10**, and d in **11**) including ether oxygen and two CN bonds (Figure 2: f in **27**, h in **28**) where N is monomethylated. Most of the data for these three bond types are arranged along straight to slightly curvilinear lines. The exceptions are formally single CC bonds (ranging from 1.484(3) to 1.513 (4) Å) participating in hyperconjugation: all CH_3-C_{sp2}, CH_2-C_{sp2}, and $CH-C_{sp2}$ base-substituent bonds or bonds in substituents. The distribution of crystal packing effects in Figure 6 exhibits high concentration of bond shortening cases at medium to high bond orders (>0.4, blue), bond lengthenings at low bond orders (0−0.25, red), and uniform distribution of bonds without correction for crystal packing effects at medium and low bond orders (purple). The major mixing area is at the bond orders 0.25−0.4. The neighborhood also influences bond lengths. The two longest CN bonds have $n = 1$ (× in Figure 6), which is in accordance with the chemical structures in Figure 2. It can be observed that all formal double CO bonds (i.e. all CO bonds except the two longest ones) have $n = 2$ (marked as + in Figure 6) as well as formal double CN and CC bonds (in this way represented in Kekulé structures in Figure 2) at high bond orders (> 0.6), formal single CN bonds in the mentioned mixing area, and half of the longest CC bonds influenced by hyperconjugation. There is a great mixing of CC and CN bonds having $n = 3$ and $n = 4$ over large bond order range (0−0.6); at $p_{wco} > 0.6$ there are only bonds with $n = 3$. These relationships between bond

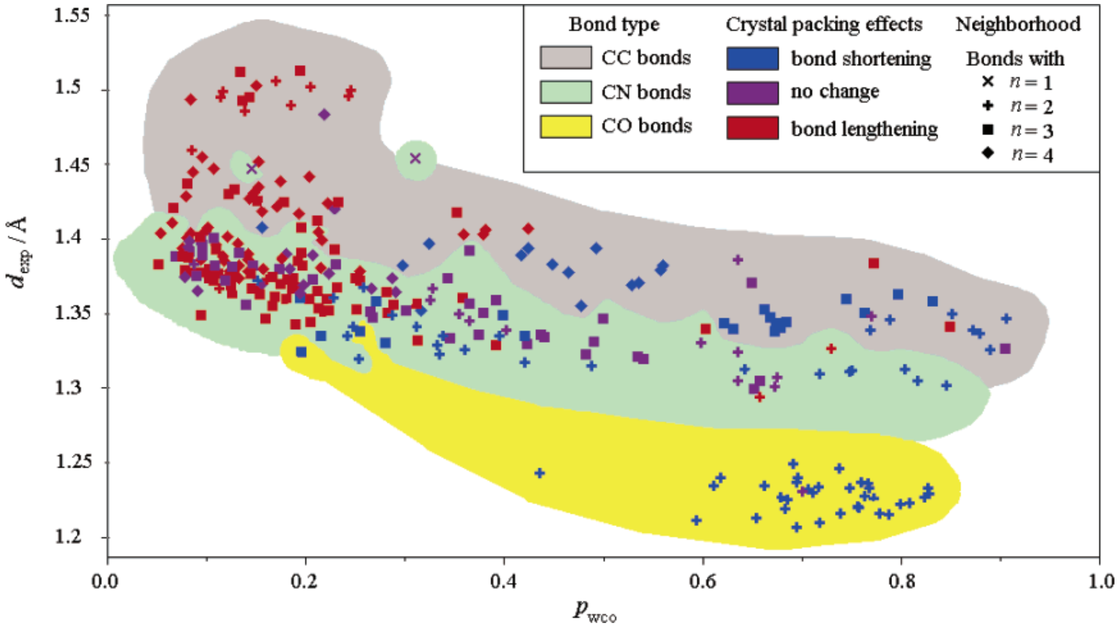

Figure 6. Weighted π-bond order p_{wco} vs experimental bond length d_{exp} (set I+II). Bond types, crystal packing, and neighborhood effects are showed by distinct colors and symbols.

Table 6. Varimax Rotated PCA Results for the Training/Validation Set (Set I+II)

PC	% variance	% cum. var	p_P	p_{wco}	p_{cr}	p_m	p_s	n	Q
PC1	69.77	69.77	−0.441	−0.441	0.442	0.445	0.435	−0.160	0.054
PC2	14.62	84.39	−0.029	−0.063	−0.069	−0.052	−0.092	0.209	−0.967
PC3	14.43	98.82	−0.162	−0.165	−0.165	−0.165	−0.204	0.907	−0.168
PC4	0.49	99.31	0.809	−0.116	−0.567	0.022	−0.103	−0.004	−0.004
PC5	0.41	100.00	−0.150	−0.233	−0.205	−0.189	0.919	−0.009	0.014
PC6	0.29	100.00	−0.078	0.874	−0.381	−0.261	−0.133	−0.001	−0.001
PC7	0.00	100.00	0.148	0.026	0.629	−0.763	0.002	0.000	0.001

orders, bond type, and bond neighborhood are reflected in correlation coefficients r between the bond descriptors. Bond orders highly correlate with each other ($r > 0.94$), weakly to moderately with n ($r = -0.48$ to -0.54), and weakly with Q ($r = 0.17$ -0.24). There is also weak to moderate correlation between Q and n ($r = -0.403$). Descriptors for PB-PAHs exhibited similar correlations: high between bond orders ($r > 0.99$) and weak to moderate between the bond orders and topological indices (n, m, l; $r = -0.33$ to -0.64), but much higher correlations between topological indices ($r = 0.84{-}0.92$) than it is between Q and n. PCA analysis could show some similarities between PB-PAHs and nucleobases. Varimax rotated PCA on autoscaled data for PB-PAHs[38] (p_P, p_{cr}, n, m, l) gave the first three PCs explaining 98.93% of the total variance. The main contribution for PC1 comes from the topological indices (sum of the loading coefficient squares reaches 95%), PC2 is mainly a linear combination of the bond orders (86% contribution), and PC3 is practically depending only on n (94% contribution). Furthermore, PC1 highly correlates with the topological indices ($r > 0.8$), PC2 correlates with the bond orders ($r > 0.95$), and the highest correlation of PC3 is with n (only $r = 0.45$). On the other side, experimental bond lengths have very weak (0.16), moderate to weak (0.42), and high (−0.83) correlation with PC3, PC1, and PC2, respectively. Varimax rotated PCA for set I+II (Table 6) shows that the first three PCs contain 98.82% of original data, similarly to PB-PAHs. Moreover, the contribution of bond orders (p_P, p_{wco}, p_{cr}, p_m, p_s) to PC1 is almost exclusive (97%). Indices n (topological) and Q (electrotopological) have the major contribution to PC2 (94%, Q) and PC3 (82%, n). Extremely high correlation of the bond orders with PC1 ($r > 0.97$) and Q with PC2 (−0.98) and n with PC3 (0.91) just confirms these findings. d_{exp} correlates moderately to PC1 (−0.64) and PC2 (0.67) and weakly to PC3 (0.14). The only significant difference between the PAHs and the nucleobases is that the main PC1/PC2/PC3 characteristic is topological/electronic/topological for the PAHs and electronic/electrotopological/topological for nucleobases. In both cases PC3 is determined by n describing the closest bond neighborhood. The substituent effects in nucleobase provoke minor sterical and substantial electronic effects. In PB-PAHs consisting only of hydrogen and sp^2 carbon, the arrangement of fused hexagons is the main factor determining topology and being predominant to electronic effects. That is why topological indices have always been working well for PAHs in general, in BLBOR, and other theoretical studies.[80] 3D scores plots for PB-PAHs and nucleobases (Figure 7) reveal more properties in common. CC bonds in PAHs were classified as 12 distinct groups depending on the values of the topological indices n, m, l (Figure 7A). The number of these groups in PB-PAHs is theoretically the number of bond types that can be obtained by drawing molecular fragments around a particular bond. For nucleobases, the number of bond types is theoretically 10 (see Table 3 for crystal packing correction parameters). From there, one group is not present in set I+II neither in set III: CO bond with $n = 1$, in functional group $H_2N(H)-C=X$, ($X = C,N$). There are 3 groups of CC bonds, 4 groups of CN bonds, and 2 of CO bonds (Figure 7C) in nucleobases. Numbers 12 and 10 are based on atomic

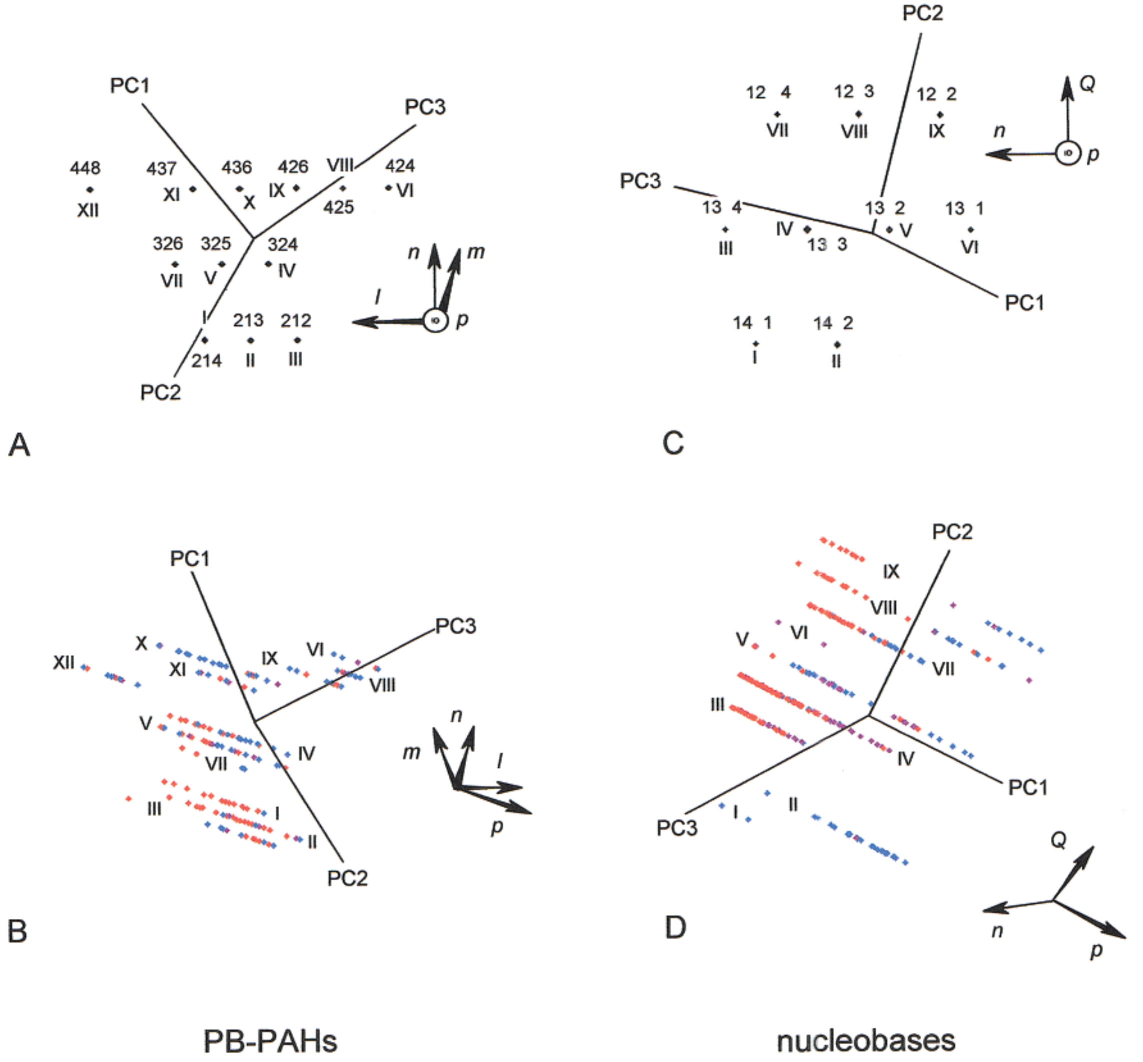

Figure 7. 3D scores plots for PB-PAHs and nucleobases. Groups of bonds (rows of points) projected perpendicularly to the plane of paper, are visible as points in A and C, and are marked with the group number (roman numerals) and with integer descriptors: A: *nml* for PB-PHs; B: *Qn* for nucleobases. Crystal packing effects, with the same coloring as in Figures 5 and 6, are presented in B and D. The coordinate systems with original axes (p − bond orders) are qualitatively oriented.

properties (valence, hybridization, atomic electron structure, atomic size, etc.) responsible for formation of π-bonds. The groups can be better viewed in Figure 7B,D, as parallel lines containing at least two bonds. The crystal packing effects show pronounced regularity in these plots; even not presented in all groups, bond lengthening/shortening occurs at low/high bond orders in both PB-PAHs and nucleobases. All the discussion on PCA results for nucleobases included only *co* data; *kr* and *dg* data exhibit practically the very same trends. Even including the prediction data set (set III), thus comprising 463 bond lengths, the varimax rotated PCA results are very much similar to the previous ones (with any p_w), which confirms that the studied properties of nucleobases are in common for all three data sets. Moreover, if varimax rotated PCA is performed using the data from set I (67 bonds, Table 1), the results are almost identical to those for set I+II (3PCs account for 98.95% total variance). Only the scores matrix shows the absence of groups I and VI (Figure 7C and D), because nucleobases **1**−**7** do not possess bonds $-NH-CH_3$ and $-O-CH_3$.

HCA results (single linkage method) for set I+II are briefly summarized in Figure 8. The PAHs−nucleobases paralelism from PCA is reconfirmed. Comparing the sample dendograms with the 3D scores plots (Figure 7), one can notice three "layers" containing 2−4 rows (groups of bonds, just discussed in PCA). In PAHs (Figure 7A) groups I−III make the bottom layer, then IV, V, and VII are in the middle layer, and the other groups are in the top layer. The dendogram (Figure 8A) consists of the cluster I−III and subcluster IV, V, and VII (similarity index > 0.6), while the rest are in a few subclusters. The nucleobase 3D scores space (Figure 7C) contains groups I−II in the bottom layer, III−VI in the middle, and VII−IX in the top layer. Analogously, the corresponding dendogram (Figure 8B) consists of well-defined subclusters (I, II), a big subcluster (it contains III−VI as four further subclusters), and the rest is in another subcluster (VII−IX). It is interesting to note that n in PB-PAHs and Q in nucleobases have similar function, to separate the data into the scores layers, while other classificatory variables (topological indices) spread the layers into distinct lines. Finally, inside each line are the bonds with the same neighborhood, so the electronic effects are the major factor to determine the bond length for these bonds. HCA dendogram for variables (Figure 8C,D) basically agrees with the PCA results. Bond orders, due to high correlation between each other (similarity index ≈ 0.9), make

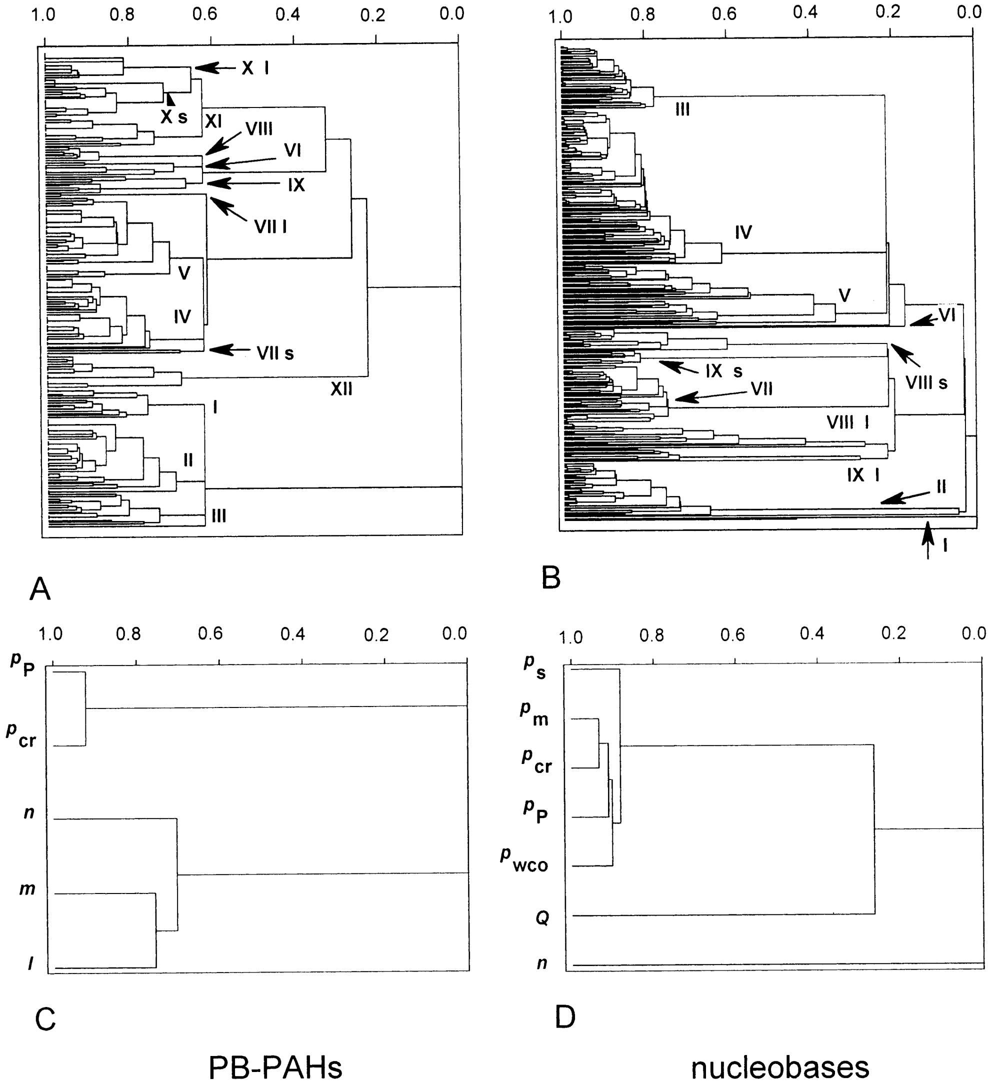

Figure 8. HCA dendograms for samples (A, B) and variables (C, D) for PB-PAHs and nucleobases (p_{wco}, set I+II). The groups from the 3D PCA scores plots are marked in the same way, adding **s** (for short bonds) and **l** (for long bonds) for subgroups of VII and X (PB-PAHs) and VIII and IX (nucleobases).

a cluster for the PAHs and distinct subcluster for nucleobases. Topological indices, although forming another cluster for the PAHs, are not so much correlated (similarity index < 0.8) which accounts for two PCs originated from these descriptors. In the case of nucleobases, Q and n even do not belong to the same cluster (similarity index zero). These differences in behavior of nonbond order descriptors between the PAHs and nucleobases can explain why there are two PCs (PC2 and PC3, Table 3) coming basically from one such descriptor in the case of nucleobases, and only one PC (PC3) is formed from a topological variable in PCA for PB-PAHs. There is no significant difference in orientation of original variables and new PC axes in the scores space for nucleobases (see Figure 7B). For PB-PAHs, the two coordinate systems are substantially different in orientation and dimensionality (Figure 7D). Besides compressing the bond order data into one PC, PCA also reduced the number of topological indices (3) into two main PCs for PB-PAHs. For nucleobases, this reduction did not occur since there were only two topological/electrotopological variables. HCA results are practically the same or not substantially different if *kr* or *dg* data are applied, and even if sets I and I+II+III are considered. This fact supports the conclusion that stuctural, topological, and electronic propeties of nucleobases, expressed through bond length data, are intrinsic characteristics of this class of compounds.

d. Finding the Best Prediction/Calculation Models by Means of PCA-HCA. In previous work[38] LR, MLR, and PLS models were constructed to predict CC bond lengths in PB-PAHs. A few parameters were calculated to validate the models and to propose the best, simplest, and the most appropriate: R, and Q, SEV, Δ/σ_{exp}. PLS model using three PCs and five bond descriptors (p_P, p_{cr}, n, m, l) was proposed. LR equations $d/\text{Å} = a + bf(x)$, where $f(x) = x$, $\ln(x + 1)$, or $(x + 1)^{-1/2}$ and $x = p_P$ or p_s, clearly showed that p_{cr} is preferred. By introducing crystal packing effects, $\langle\Delta\rangle$ and SEV decreased by 0.003 Å, R increased by 3%, and $\langle\Delta/\sigma\rangle$ became less than 2.58. There was no preference for linear, Pauling's log, or Gordy's equation. The Pauling analytical curves employing *kr*, *co*, and *dg* data, and p_P or p_s are systematically inferior to the LR models. One would expect such trends for nucleobases also. In Table 7 there are 50 models for prediction and calculation of CC, CN, and CO bond lengths in set I+II, presented by 15 statistical parameters for analytical (Pauling harmonic potential curves, PHC), LR (simple, Pauling and Gordy equations), MLR, PCR, and

Table 7. Statistics[a,b] for Models[c] for Calculation/Prediction of CC, CN, and CO Bond Lengths in Nucleobases

no	bond descriptors	set[d]	method	Δ_{max}/ Å	$\langle\Delta\rangle$/ Å	$\langle\Delta/\sigma_{exp}\rangle$	*a*/Å	*b*	*t*	*F*	*FIT*	R^2	*SEC*/ Å	*pressc*/ Å^2	Q^2	*SEV*/ Å	*pressυ*/ Å^2	*NV*
1	p_P	I+II	LR	**0.136**	**0.037**	**16.21**	**0.780**	**0.426**	15.11	**228**	**0.74**	**0.426**	**0.048**	**0.708**	0.418	**0.048**	**0.718**	1
2	p_s	I+II	LR	**0.128**	**0.035**	**15.72**	**0.679**	**0.501**	17.54	**308**	**0.99**	0.501	**0.045**	**0.616**	0.493	**0.045**	**0.625**	1
3	p_P	I+II	LR, P	**0.135**	**0.037**	**16.40**	**0.784**	**0.424**	15.03	**226**	**0.73**	**0.424**	**0.048**	**0.710**	0.416	**0.048**	**0.720**	1
4	p_s	I+II	LR, P	**0.130**	**0.036**	**16.08**	**0.704**	**0.483**	16.93	**287**	**0.93**	**0.483**	**0.046**	**0.638**	0.476	**0.046**	**0.647**	1
5	p_P	I+II	LR, G	**0.136**	**0.037**	**16.49**	**0.788**	**0.421**	14.94	**223**	**0.72**	**0.421**	**0.048**	**0.714**	0.413	**0.049**	**0.724**	1
6	p_s	I+II	LR, G	**0.132**	**0.037**	**16.27**	**0.720**	**0.471**	16.53	**273**	**0.88**	**0.471**	**0.046**	**0.653**	0.463	**0.046**	**0.662**	1
7	p_P	I+II	PHC, *co*	**0.105**	**0.055**	**25.27**	0.404	0.741		**51**	**0.16**	**0.141**	**0.059**	**1.059**				4
8	p_s	I+II	PHC, *co*	**0.114**	**0.058**	**26.83**	0.276	0.838		**18**	**0.06**	**0.054**	**0.062**	**1.167**				4
9	p_P	I+II	PHC, *kr*	**0.105**	**0.050**	**22.73**	0.412	0.731		**120**	**0.39**	**0.281**	**0.054**	**0.887**				4
10	p_s	I+II	PHC, *kr*	**0.114**	**0.053**	**24.61**	0.279	0.833		**69**	**0.22**	**0.184**	**0.057**	**1.006**				4
11	p_P	I+II	PHC, *dg*	**0.089**	**0.033**	**14.63**	0.279	0.815		516	**1.67**	0.627	**0.039**	**0.460**				4
12	p_s	I+II	PHC, *dg*	**0.114**	**0.038**	**17.48**	0.141	0.922		**315**	**1.01**	0.506	**0.045**	**0.609**				4
13	p_P, p_s, n, Q	I	MLR	0.062	0.013	**10.31**	0.125	0.908	**0.59**	**153**	7.36	0.908	0.018	0.018	0.886	0.019	0.023	4
E				**0.104**	0.017	**7.15**	0.245	0.817		**441**	**5.43**	0.853	0.024	0.018				
14	p_P, p_s, n, Q	I+II	MLR	0.092	0.017	**7.27**	0.180	0.868	**1.09**	499	**6.14**	0.868	0.023	0.163	0.863	0.024	0.170	4
15	p_P, p_s, p_{cr}, p_m, p_{wco}, n, Q	I	MLR, *co*	0.065	0.010	**8.00**	0.080	0.941	**0.21**	**134**	8.09	0.941	0.014	0.012	0.918	0.017	0.016	7
E				**0.606**	0.017	**8.04**	0.199	0.849		**62**	**1.22**	0.592	0.041	**0.503**				
16	p_P, p_s, p_{cr}, p_m, p_{wco}, n, Q	I+II	MLR, *co*	0.080	0.015	**6.15**	0.133	0.902	**0.20**	**398**	7.78	0.902	0.020	0.120	0.807	0.028	0.238	7
17	p_P, p_s, p_{cr}, p_m, p_{wkr}, n, Q	I	MLR, *kr*	0.066	0.010	**8.07**	0.081	0.940	**0.15**	**132**	7.94	0.940	0.014	0.012	0.916	0.017	0.017	7
E				**0.596**	0.017	**8.02**	0.200	0.848		**65**	**1.27**	0.601	0.040	**0.492**				
18	p_P, p_s, p_{cr}, p_m, p_{wkr}, n, Q	I+II	MLR, *kr*	0.082	0.015	**6.18**	0.135	0.901	**0.18**	**391**	7.65	0.901	0.020	0.122	0.804	0.028	0.242	7
19	p_P, p_s, p_{cr}, p_m, p_{wdg}, n, Q	I	MLR, *dg*	0.061	0.010	**8.03**	0.079	0.942	**0.27**	**137**	8.24	0.942	0.014	0.012	0.920	0.016	0.016	7
E				**0.592**	0.017	**8.01**	0.199	0.845		**66**	**1.29**	0.606	0.040	**0.486**				
20	p_P, p_s, p_{cr}, p_m, p_{wdg}, n, Q	I+II	MLR, *dg*	0.083	0.015	**6.16**	0.133	0.902	**0.13**	**397**	7.76	0.902	0.020	0.121	0.802	0.029	0.244	7
21	p_P, n, Q	I	AllR	0.078	0.014	**11.53**	0.177	0.869	**0.73**	**139**	**5.50**	0.867	0.020	0.026	0.843	0.022	0.031	3
E				**0.114**	0.018	**7.33**	0.245	0.817		547	**5.17**	0.843	0.025	0.019				
22	p_P, n, Q	I+II	AllR	0.097	0.018	**7.45**	0.200	0.853	**1.93**	589	**5.56**	0.853	0.024	0.182	0.848	0.025	0.188	3
23	p_s, n, Q	I	AllR	0.065	0.013	**10.22**	0.128	0.905	**0.45**	**201**	7.93	0.905	0.017	0.019	0.886	0.019	0.023	3
E				**0.105**	0.017	**6.97**	0.234	0.825		616	**5.81**	0.858	0.024	0.017				
24	p_s, n, Q	I+II	AllR	0.091	0.017	**7.30**	0.181	0.867	3.25	664	**6.27**	0.867	0.023	0.164	0.863	0.024	0.169	3
25	p_P, p_s, n, Q	I	PCR(3)	0.071	0.013	**10.37**	0.145	0.892	-	**174**	**6.87**	0.892	0.019	0.021	0.871	0.020	0.026	3
E				**0.109**	0.017	**6.90**	0.232	0.827		**605**	**5.71**	0.856	0.024	0.018				
26	p_P, p_s, n, Q	I+II	PCR(3)	0.093	0.017	**7.25**	0.182	0.867	-	657	**6.19**	0.866	0.023	0.165	0.863	0.023	0.170	3
27	p_P, p_s, p_{cr}, p_m, p_{wco}, n, Q	I	PCR(3), *co*	0.069	0.012	**9.50**	0.127	0.906	-	**203**	8.02	0.906	0.018	0.019	0.887	0.018	0.022	3
E				**0.102**	0.016	**6.44**	0.215	0.839		696	**6.56**	0.873	0.023	0.157				
28	p_P, p_s, p_{cr}, p_m, p_{wco}, n, Q	I+II	PCR(3), *co*	0.087	0.016	**6.80**	0.162	0.881	-	753	7.11	0.881	0.022	0.147	0.877	0.022	0.152	3
29	p_P, p_s, p_{cr}, p_m, p_{wkr}, n, Q	I	PCR(3), *kr*	0.069	0.012	**9.49**	0.127	0.906	-	**202**	8.01	0.906	0.018	0.019	0.887	0.018	0.022	3
E				**0.102**	0.016	**6.46**	0.215	0.839		694	**6.55**	0.872	0.023	0.158				
30	p_P, p_s, p_{cr}, p_m, p_{wkr}, n, Q	I+II	PCR(3), *kr*	0.087	0.016	**6.81**	0.162	0.881		751	7.08	0.881	0.022	0.147	0.877	0.022	0.152	3
31	p_P, p_s, p_{cr}, p_m, p_{wdg}, n, Q	I	PCR(3), *dg*	0.068	0.012	**9.46**	0.126	0.907		**205**	8.09	0.907	0.018	0.018	0.888	0.018	0.022	3
E				**0.102**	0.016	**6.44**	0.214	0.840		698	**6.58**	0.873	0.023	0.157				
32	p_P, p_s, p_{cr}, p_m, p_{wdg}, n, Q	I+II	PCR(4),*dg*	0.083	0.016	**6.76**	0.156	0.886		588	7.23	0.886	0.020	0.141	0.881	0.022	0.147	4
33	p_P, p_s, n, Q	I	PLS(3)	0.069	0.013	**10.29**	0.141	0.896		**180**	7.12	0.896	0.018	0.021	0.874	0.019	0.025	3
E				**0.107**	0.017	**6.90**	0.230	0.828		614	**5.79**	0.858	0.024	0.175				
34	p_P, p_s, n, Q	I+II	PLS(3)	0.093	0.017	**7.25**	0.182	0.867		657	**6.19**	0.867	0.023	0.165	0.861	0.024	0.171	3
35	p_P, p_s, p_{cr}, p_m, p_{wco}, n, Q	I	PLS(4), *co*	0.060	0.011	**8.47**	0.091	0.933		**215**	10.38	0.933	0.015	0.013	0.914	0.016	0.017	4
E				0.087	0.015	**6.17**	0.215	0.839		588	7.23	0.886	0.022	0.141				
36	p_P, p_s, p_{cr}, p_m, p_{wco}, n, Q	I+II	PLS(4), *co*	0.078	0.015	**6.22**	0.134	0.901		695	8.55	0.901	0.020	0.122	0.897	0.020	0.128	4
37	p_P, p_s, p_{cr}, p_m, p_{wkr}, n, Q	I	PLS(4), *kr*	0.061	0.011	**8.43**	0.092	0.932		**212**	10.26	0.932	0.015	0.013	0.913	0.016	0.017	4
E				0.088	0.015	**6.17**	0.216	0.838		582	7.16	0.885	0.022	0.142				
38	p_P, p_s, p_{cr}, p_m, p_{wkr}, n, Q	I+II	PLS(4), *kr*	0.078	0.015	**6.24**	0.136	0.900		683	8.41	0.900	0.020	0.123	0.895	0.021	0.130	4
39	p_P, p_s, p_{cr}, p_m, p_{wdg}, n, Q	I	PLS(4), *dg*	0.059	0.011	**8.45**	0.088	0.935		**224**	10.78	0.935	0.015	0.013	0.917	0.016	0.016	4
E				0.089	0.015	**6.16**	0.213	0.840		595	7.32	0.887	0.022	0.140				
40	p_P, p_s, p_{cr}, p_m, p_{wdg}, n, Q	I+II	PLS(4), *dg*	0.081	0.015	**6.23**	0.134	0.902		696	8.56	0.902	0.020	0.121	0.897	0.020	0.128	4
41	d_{cal}	I	MMFF	0.066	0.015	**11.61**	−0.050	1.040		**370**	**5.43**	0.850	0.021	0.030				57
42	d_{cal}	I+II	MMFF	**0.190**	0.021	**8.30**	−0.002	1.004		807	**2.60**	0.724	**0.033**	**0.340**				57
43	d_{cal}	I	MNDO	0.073	**0.031**	**24.75**	−0.194	1.165		**89**	**1.31**	0.578	**0.036**	0.084				57
44	d_{cal}	I+II	MNDO	0.080	0.030	**13.44**	**−0.313**	1.116		721	**2.33**	0.701	**0.035**	**0.368**				57
45	d_{cal}	I	AM1	0.080	**0.031**	**24.90**	−0.032	1.046		**92**	**1.35**	0.585	**0.036**	0.081				57
46	d_{cal}	I+II	AM1	0.080	0.028	**12.78**	0.081	0.960		836	**2.70**	0.731	**0.033**	**0.331**				57
47	d_{cal}	I	PM3	0.081	**0.033**	**26.62**	−0.186	1.160		**66**	**0.97**	0.502	**0.040**	0.099				57
48	d_{cal}	I+II	PM3	0.086	**0.031**	**14.02**	−0.082	1.081		603	**1.94**	0.662	**0.037**	**0.416**				57
49	d_{cal}	I	HF	0.044	0.014	**11.10**	−0.266	1.192		567	8.34	0.897	0.018	0.020				57
50	d_{cal}	I+II	HF	0.052	0.014	**6.35**	−0.229	1.164		3646	11.76	0.922	0.018	0.096				57

[a] Statistical parameters (see Table 4). [b] Bold values of the statistical parameters are out of acceptable limits: $\Delta_{max} \leq 0.100$ Å; $\langle\Delta\rangle \leq 0.030$ Å; $\langle\Delta/\sigma_{exp}\rangle \geq 2{,}58$; $|a| \leq 0.3$ Å; $|1 - b| \leq 0.3$ Å; $t \geq 2.58$; $F \geq 400$; $FIT \geq 7.00$; $R^2 \geq 0.500$; $SEC \leq 0.030$ Å; $pressc \leq 0.200$ Å^2; $Q^2 \geq 0.400$; $SEV \leq 0.040$ Å; $press\upsilon \leq 0.300$ Å^2. [c] The models: LR − simple (x), Pauling (P) and Gordy (G) log curves; PHC − PHC with *co, kr, dg* data set; MLR, PCR, PLS − MLR, PCR and PLS regressions with the number of PCs in brackets, and used data sets *co*, *kr* or *dg*; AllR − MLR, PCR, PLS regressions applied with the same results; MMFF − the molecular mechanics calculation; MNDO, AM1, PM3 − semiempirical MNDO, AM1 and PM3 calculations; HF − ab initio Hartree−Fock calculation. [d] The training set or the set which was used to calculate the statistical parameters. The regression models with set I as the training set were used to calculate bond lengths for set I+II, and the corresponding statistical parameters (except for validation) are in lines marked with E (extended).

PLS, molecular mechanics (MMFF94), semiempirical (MNDO, AM1, PM3), and ab initio (HF 6-31G**) models. A brief look at the statistical parameters shows that their variations are larger than of those for PB-PAHs.[38] $\langle \Delta/\sigma_{\mathrm{exp}} \rangle$ hardly approaches the limit 5.0 (range 6.2−26.8). Most of the correlations between the parameters are not high (0.42−1.00 excluding *NV*). It is not possible to find out the best model(s) by visual inspection of Table 7, where bold values denote not satisfied criteria. Should I or I+II be the training set? The PCA analysis revealed that two bond types (I and VI, Figure 7) exist in set II but not in set I. This fact is not in favor to use I as the prediction set. The statistical parameters, including those when set I was the training set and set II the training set (marked with E in Table 7), do not show clearly that models with I+II are better than models with I as the training set. High correlations among π-bond orders with those including crystal packing effects (set I+II: 0.97−1.00) result in practically no significant improvement if p_{cr}, p_{s}, p_{m}, and p_{w}s are added to the data set (p_{p}, p_{s}, n, Q). In fact, there are five data sets which can be used to find the best prediction/calculation models by means of chemometrics: (a) statistical parameters for I+II data set, $\mathbf{X}$(statistical parameters = 11, models = 32); (b) d_{cal} compared to d_{exp}, $\mathbf{X}$(bond lengths = 309, experiment + prediction models = 33); (c) calculated Julg's aromaticity index A_{cal} (referred to Py six-membered and Pu nine-membered rings) and compared to experimental A_{exp}, $\mathbf{X}$(nucleobases = 31, experiment + prediction models = 33); (d) $\langle d \rangle$ calculated and experimental (calculated for the same rings as A), $\mathbf{X}$(nucleobases = 31, experiment + prediction models = 33); (e) the standard deviation for A_{cal} and A_{exp}, $\mathbf{X}$(nucleobases = 31, experiment + prediction models = 33). PCA and HCA results (Figures 9−11) enable one to see which models are the best and where are the simplest models 24, 26, and 34. One model or data should be a reference. For case (a) (Figure 9) it is the HF model, for cases (b) (Figure 10A), (c) (Figure 10B), and (d) (Figure 10C) it is the experimental data. Case (e) (Figure 10D) is peculiar, as the best model should have the minimum $\sigma(A)$; it is observed that model 7 has this minimum, while computational models were the worst ones. PCA and HCA reflect intrinsic properties of various methodologies and their applicability for bond length calculation in set I+II, grouping those of the same nature into small clusters. PC1 describes only 53.2−87.5% of the total variance, while for PC2 and PC3 these percentages are 3.4−17.5% and 2.2−10.9%, respectively (Figures 9 and 10). HF and semiempirical models seem to be always among the best, except for calculation of A. On the other side, all LR models, in general, are the worst ones. PHC models give the smallest $\sigma(A)$ and are not poor in predicting $\langle d \rangle$. Models 11 and 12 show that there is some preference for *dg* data. Multivariate models 14, 22, 24, 26, and 34 always make a cluster which is very close to the reference model in cases (d) and (e), not so far in (a) and (b), and rather far in (c) (prediction of A). In average, HCA confirms these trends. Although this PCA-HCA analysis is semiquantitative, it helps to find out the best models for calculation of some property. HF is not always the best model, nor PHC is not always the worst model; simply, the best for prediction/calculation of all properties does not exist. The five multivariate models seem to be good in general, placed between computational and PHC models, and could be improved. The combined PCA-HCA analysis on model quality presented in this and previous works[62,63,75] can be used for (a) finding the best and most appropriate models to predict/calculate properties under study and (b) giving more insight into the nature of the properties, including their mapping to find out the most important property regions for the studied phenomena. Table 8 contains statistical parameters for regression models 51−71 referrring to only one

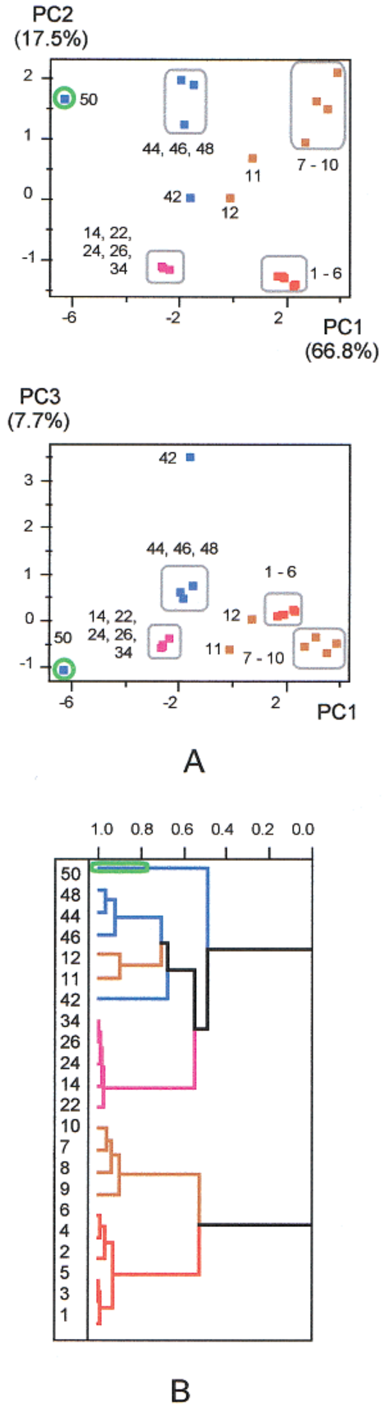

Figure 9. (A) PCA scores plots obtained by regular PCA on 11 statistical parameters (see text for details); (B) HCA dendogram on the prediction models, employing: the same 11 statistic parameters. Coloring refers to various types of prediction models: black box − experimental data, blue box − computational models (MMFF, semiempirical, HF), red box − LR models, brown box − PHC, magenta box − multivariate models with three or four bond descriptors, green open circle − the best and the reference model.

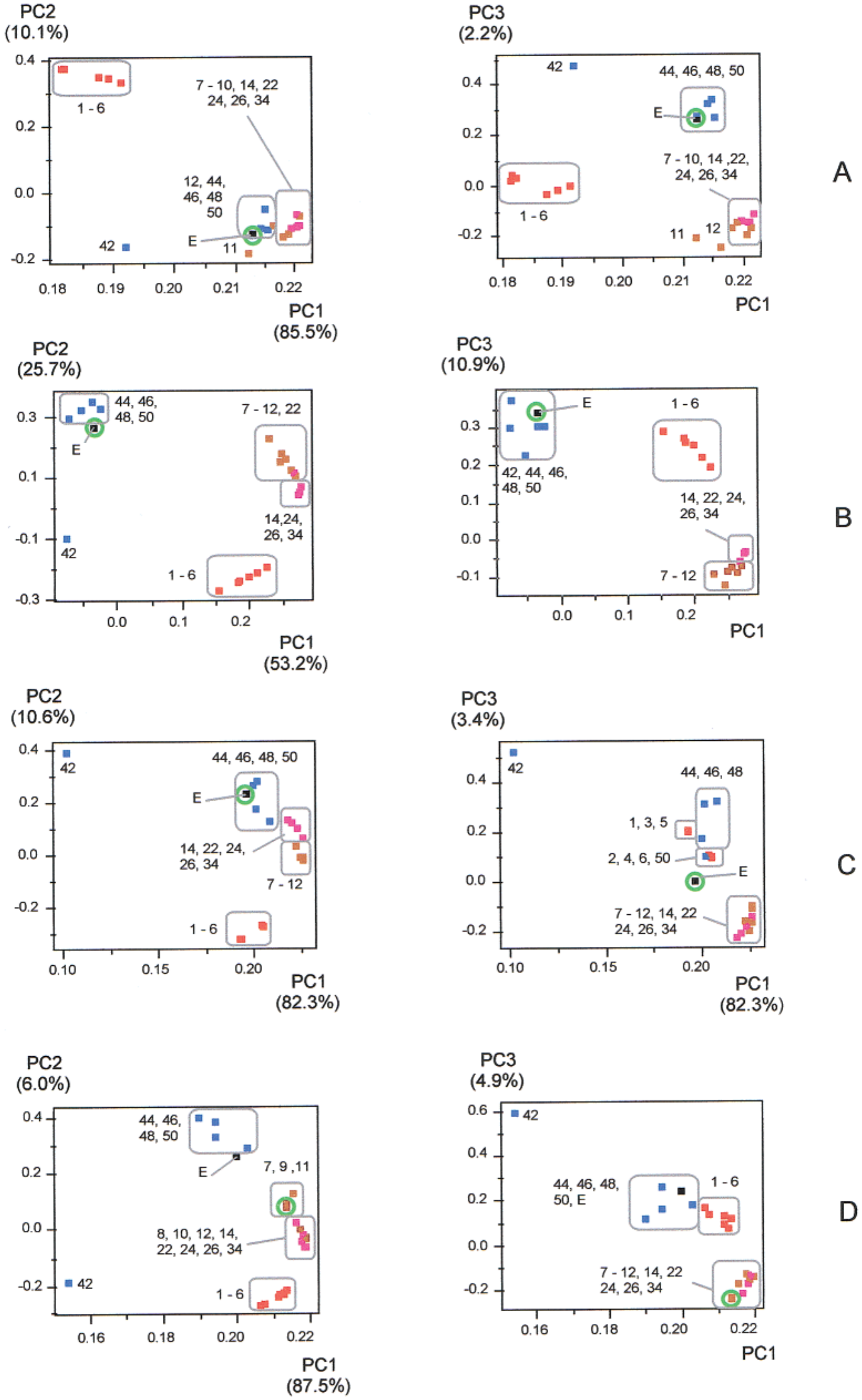

Figure 10. PCA loadings plots obtained by regular PCA on the following: (A) bond lengths d_{exp} and d_{cal}; (B) Julg's aromaticity index A; (C) average bond length $\langle d \rangle$; and (D) standard deviation $\sigma(A)$. Coloring is the same as in Figure 9.

bond type. These models, in general, are not better than analogous models (from Table 7) using the whole data set. They fail in *a*, *b*, *F*, *FIT*, R^2, and Q^2 and even bring some improvement in Δ_{max}, $\langle\Delta\rangle$, $\langle\Delta/\sigma_{exp}\rangle$, *SEC*, and *SEV*. According to these findings, the separation of the nucleobase data into CC, CN, and CO bond data is not recommendable. Rationale for this could be the cyclic nature of nucleobases and the π-electron delocalization phenomena which include the interaction between the ring and exocyclic bonds.

e. Prediction/Calculation of Bond Lengths in Set III. In Tables I–III in Supporting Information are the bond lengths calculated by HF, PM3, and PLS model 34. Only

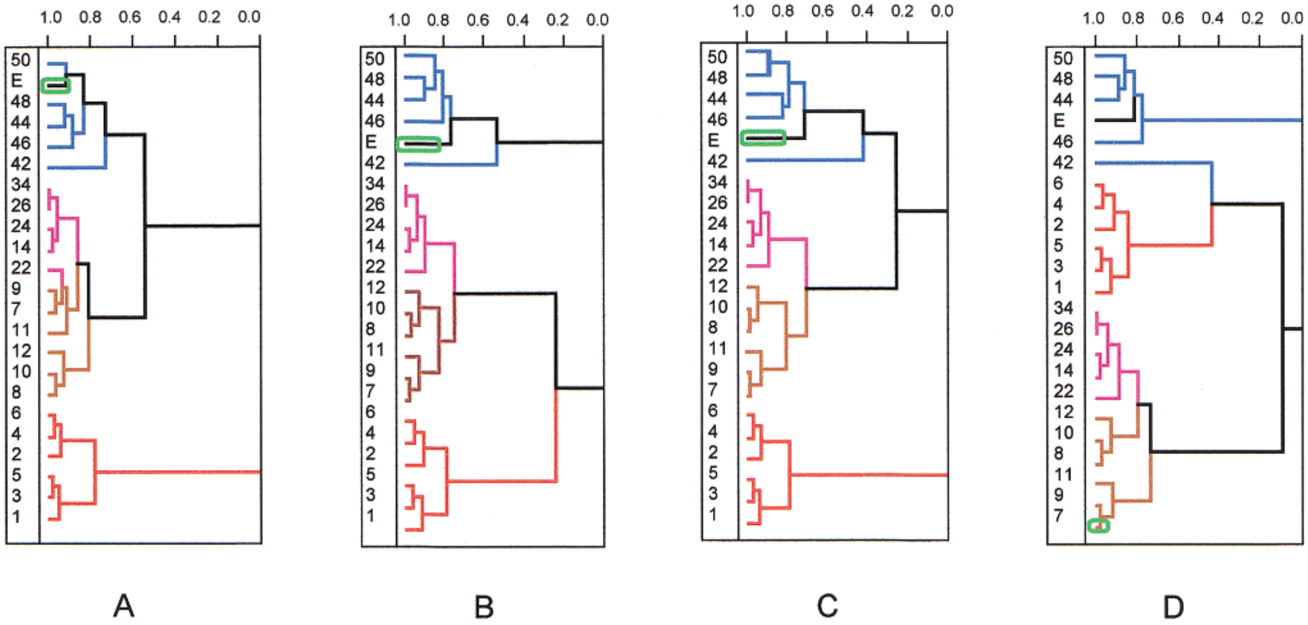

Figure 11. HCA dendograms on prediction models, employing the following: (A) bond lengths d_{exp} and d_{cal}; (B) Julg's aromaticity index A; (C) average bond length $\langle d \rangle$; and (D) standard deviation $\sigma(A)$, as variables. Coloring is the same as in Figures 9 and 10.

Table 8. Statistical Parameters[a,b] for Models[c] for Calculation/Prediction of CC, CN, or CO Bond Lengths in Nucleobases

no	bond descriptors	set[d]	co[e]	method	Δ_{max}/ Å	$\langle\Delta\rangle$/ Å	$\langle\Delta/\sigma_{exp}\rangle$	a/Å	b	t	F	FIT	R^2	SEC/ Å	$pressc$/ Å^2	Q^2	SEV/ Å	$pressv$/ Å^2	NV
51	p_P	(I+II)CC	1	LR	0.076	*0.025*	*9.69*	*0.481*	*0.659*	**12.74**	**162**	*1.86*	*0.659*	*0.031*	0.083	*0.644*	*0.032*	0.086	1
52	p_s	(I+II)CC	2	LR	0.071	*0.024*	*9.10*	*0.423*	*0.700*	**14.00**	**196**	*2.25*	*0.700*	*0.029*	0.073	*0.686*	*0.030*	0.076	1
53	p_P, p_s, n	(I+II)CC	14	MLR	0.068	**0.020**	**7.62**	**0.332**	**0.765**	*1.53*	**89**	**2.90**	**0.765**	**0.026**	0.057	**0.744**	**0.028**	0.062	3
54	p_P, p_s, n	(I+II)CC	26	PCR(2)	0.067	**0.021**	**8.37**	**0.358**	**0.746**		**122**	**2.71**	**0.746**	**0.027**	0.062	**0.729**	**0.028**	0.066	2
55	p_P, p_s, n	(I+II)CC	34	PLS(2)	0.066	**0.021**	**8.36**	**0.357**	**0.747**		**122**	**2.72**	**0.747**	**0.027**	0.062	**0.730**	**0.028**	0.066	2
56	p_P, n	(I+II)CC	22	AllR	0.069	**0.022**	**8.74**	**0.384**	**0.728**	*4.59*	**111**	**2.47**	**0.728**	**0.028**	0.066	**0.710**	**0.029**	0.071	2
57	p_s, n	(I+II)CC	24	AllR	0.066	**0.021**	**8.07**	**0.341**	**0.758**	*4.45*	**130**	**2.88**	**0.758**	**0.027**	0.059	**0.741**	**0.028**	0.063	2
58	p_P	(I+II)CN	1	LR	0.101	*0.014*	*5.80*	*0.626*	*0.541*	*14.69*	**216**	*1.16*	*0.541*	*0.019*	0.067	*0.533*	*0.019*	0.069	1
59	p_s	(I+II)CN	2	LR	0.102	*0.014*	*5.77*	*0.579*	*0.575*	*15.75*	**248**	*1.33*	*0.574*	*0.018*	0.062	*0.567*	*0.019*	0.064	1
60	p_P, p_s, n	(I+II)CN	14	MLR	0.106	*0.013*	*5.55*	**0.567**	**0.584**	**1.00**	**85**	**1.31**	**0.584**	*0.018*	0.061	**0.553**	*0.019*	0.066	3
61	p_P, p_s, n	(I+II)CN	26	PCR(2)	0.107	*0.013*	*5.52*	**0.570**	**0.582**		**127**	**1.34**	**0.582**	*0.018*	0.061	**0.555**	*0.018*	0.065	2
62	p_P, p_s, n	(I+II)CN	34	PLS(2)	0.107	*0.013*	*5.52*	**0.569**	**0.583**		**127**	**1.34**	**0.582**	*0.018*	0.061	**0.555**	*0.019*	0.065	2
63	p_P, n	(I+II)CN	22	AllR	0.111	*0.013*	*5.62*	**0.607**	**0.555**	*2.35*	**113**	**1.20**	**0.555**	*0.019*	0.065	**0.525**	*0.020*	0.070	2
64	p_s, n	(I+II)CN	24	AllR	0.105	*0.013*	*5.65*	**0.577**	**0.577**	**0.78**	**124**	**1.31**	**0.577**	*0.019*	0.065	**0.550**	*0.019*	0.066	2
65	p_P	(I+II)CO	1	LR	0.076	*0.014*	*5.23*	**0.865**	**0.299**	**3.92**	**15**	**0.39**	**0.299**	**0.217**	0.017	**0.076**	*0.025*	0.022	1
66	p_s	(I+II)CO	2	LR	0.074	*0.014*	*5.26*	**0.836**	**0.322**	**4.14**	**17**	**0.44**	**0.322**	*0.021*	0.016	**0.076**	*0.025*	0.022	1
67	p_P, p_s, n	(I+II)CO	14	MLR	0.021	*0.008*	*3.38*	**0.195**	**0.842**	**0.22**	**60**	**3.86**	**0.842**	*0.011*	0.004	**−22.031**	*0.128*	0.555	3
68	p_P, p_s, n	(I+II)CO	26	PCR(2)	0.021	*0.008*	*3.40*	**0.195**	**0.842**		**93**	**4.44**	**0.842**	*0.010*	0.004	**0.806**	*0.011*	0.005	2
69	p_P, p_s, n	(I+II)CO	34	PLS(2)	0.021	*0.008*	*3.40*	**0.195**	**0.842**		**93**	**4.44**	**0.842**	*0.010*	0.004	**0.806**	*0.011*	0.005	2
70	p_P, p_s, n	(I+II)CO	22	AllR	0.021	*0.008*	*3.40*	**0.195**	**0.842**	**0.261**	**93**	**4.43**	**0.842**	*0.010*	0.004	**0.804**	*0.012*	0.005	2
71	p_P, p_s, n	(I+II)CO	24	AllR	0.021	*0.008*	*3.39*	**0.195**	**0.842**	**0.311**	**93**	**4.44**	**0.842**	*0.010*	0.004	**0.806**	*0.012*	0.005	2

[a] Statistical parameters (see Table 4). [b] Bold/italics values of the statistical parameters which are better/worse than those in the analogous models for comparison. Plain values are not suitable for comparison, due to difference in the size of the data sets. [c] The predictive models (see Table 7): LR − simple; MLR, PCR, PLS − with the number of PCs in brackets; AllR − MLR, PCR, PLS regressions applied with the same results. [d] The training set or the set which was used to calculate the statistical parameters, with the bond type. [e] Analogous models from Table 7, used for comparison purposes.

nucleosides **35** and **45** possess 12 high quality values of d_{exp} (Table 2). These bond lengths are reproduced best by HF and then followed by PLS and PM3. When all 154 experimental bonds are considered, then HF is the best method in terms of Δ/d_{exp}, $\langle\Delta\rangle$, and r, and the second place is shared among PM3, PLS, and MMFF. Figure 12 illustrates the relations between d_{exp} and d_{cal} from model 34. Supposing that the cumulative effect of the crystal packing forces and substitution at/in the nucleobase ring on the 154 bond is at maximum ±0.05 Å, there are some samples out of this limit around the regression line $d_{cal} = d_{exp}$. At first, these are the CC bonds with low-quality d_{exp} in **34**, **36**, **45**, **48**, and **49** (rounded in cyan in Figure 12) and also some CC bonds with high-quality d_{exp} in **36**, **38**, **39**, **45**, and **47**. All CN and CO bonds are well concentrated around the regression line; all CO bonds are overpredicted. It is clear that the PLS model 34 and its analogues (14, 22, 24, 26, 34) can be used quite satisfactorily in bond length prediction for nucleobases.

f. Structural Aromaticity Indices for Nucleobases. The Effect of Intermolecular Interactions. (*a*) *The Resonance Structure Contributions*. The ionic resonance structures are predominant, and their contribution to resonance hybrid is rather high (Figures 3 and 4, *co* data, error 1%.): 40−60% for **2**, **4**, **6**, and >70% for **1**, **3**, **5**, **7**. Hyperconjugation contribution for **3** is 33%. Similar trends are observed for **8**−**31**. The crucial point to describe nucleoside CC, CN, and

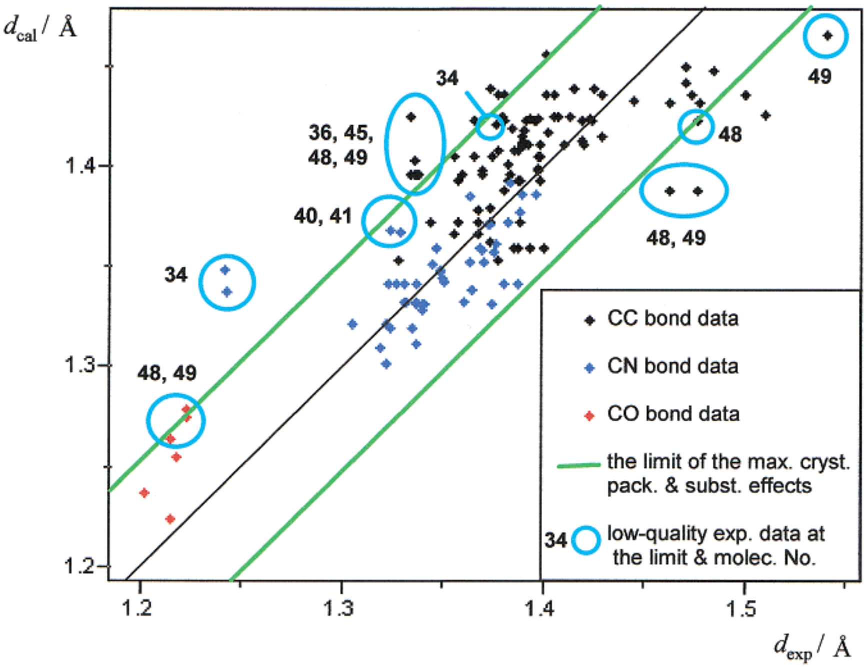

Figure 12. The PLS prediction results (model 34) for set III. The maximum limit 0.050 Å (green lines) includes crystal packing and substituent effects. Some low-quality (rounded in cyan) and high-quality data (not rounded) are out of the limit.

CO bond lengths in this work is the calculation of the Pauling π-bond orders, which then can be corrected to crystal packing effects; information on bond lengths not contained in bond orders can be added in simple variables such as Q and n. The construction of resonance structures for nucleobases **1**−**50** is shown to be based on correct theory.

(b) The Regression Coefficients. The regression coefficients which stand before π-bond orders are electronic aromaticity indices and can be expressed as b (the coefficient for only one bond order in the equation) or B (the sum of all coefficients for bond orders in the equation). Greater absolute value of b or B is related to a higher degree of aromaticity.[38] Table 9 compares these coefficients for PB-PAHs,[38] aza- and diaza-PAHs,[31,32,37] and nucleobases. There is no statistically relevant difference between PB-PAHs and their aza-derivatives in b, B coefficients, so the CC bond shortening due to presence of N or crystal packing effects cannot be observed. The packing effects in nucleobases are also unobserved. It is interesting to note that in terms of b and B, there is no difference between nucleobases and aza-PAHs series, although the differences in the proposed PLS equations and regression vectors are remarkable (Table 9).

Table 9. Selected Regression Parameters and BLBOR Equations for Various (Hetero)Aromatic Classes

class	bond type	bond descriptors	coefficient
		Regression Coefficients b and B	
PB-PAHs[a]	CC	p_{P}	−0.147(5)
		p_{cr}	−0.151(4)
		MLR: p_{P}, p_{cr}, n, m, l	−0.130(39)
		PLS: p_{P}, p_{cr}, n, m, l	−0.125
aza-PAHs[b]	CC	p_{P}	−0.143(13)
	CN	p_{P}	−0.184(18)
diaza-PAHs[b]	CC	p_{P}	−0.143(8)
	CN	p_{P}	−0.152(8)
nucleobases[c]	CC, CN, CO	p_{P}	−0.199(13)
		p_{s}	−0.168(10)
		MLR: p_{P}, p_{s}, n, Q	−0.149(26)
		PLS: p_{P}, p_{s}, n, Q	−0.159

PLS Regression Equations

deautoscaled equation for PB-PAHs:[a] $(d/\text{Å} - 1.409)/0.032 = -1.929(p_{\mathrm{P}\,-}\ 0.402) - 1.990(p_{\mathrm{cr}} - 0.405) + 0.196(n - 3.09) + 0.127(m - 2.08) + 0.004(l - 4.78)$

PLS regression vector for autoscaled equation: p_{P}: −0.357; p_{cr}: −0.385; n: 0.157; m: 0.114; l: 0.006

deautoscaled equation for nucleobases:[c] $(d/\text{Å} - 1.360)/0.063 = -1.271(p_{\mathrm{P}} - 0.347) - 1.241(p_{\mathrm{s}} - 0.346) - 1.075(Q - 12.85) - 0.102(n - 2.94)$

PLS regression vector for autoscaled equation: p_{P}: −0.264; p_{s}: −0.330; Q: −0.661; n: −0.078

[a] From a previous work.[38] [b] From another previous work.[39] [c] In this work.

(c) *The PCA-HCA Analysis on Some Experimental and Calculated/Predicted Structural Parameters*. In Table 10 are experimental and calculated aromaticity indices A, $\sigma(A)$, and $\langle d\rangle$ for Pu/Py rings in **1**−**31**. The relative error for A_{cal}, $\Delta/A_{exp} > 10\%$ can be encountered 19 times for HF and only 7 times for PM3 and PLS. A_{exp} for Pu ranges is 0.85−0.94 and for Py is 0.76−0.93. Scores plot from PCA analysis case (c) in section **d.**, illustrated by Figure 10B, reveals a quite well separation of Pu from Py rings in the PC1−PC2 plot (Figure 13A). In general, A increases with both PC1 and PC2. Some Py behave in the plot like Pu: **24** − with an extended heteroaromatic Py ring by a C═O group and **26** − with a cycloalkene fused with Py ring. Also, Pu **19** and **25** act as Py, mainly due to a tertiary amine N which disrupts the electron delocalization in the Pu ring. The HCA analysis of the same data set reveals additional details (Figure 13B) which can be observed also in the PCA plot (Figure 13A). There are five clusters and subclusters, presented by the standard nucleobases (C, T, U, A, G). G-type nucleobases are all Pu (**7**, **8**, **22**, **28**) except **11** (this Py ring is similar to Py fragment of **7** and **28**). A-type nucleobases include Pu (**5**, **6**, **18**, **20**, **25**) and a nontypical Py **26**. C-type nucleobases are only Py (**1**, **10**, **12**, **15**, **17**, **27**, **29**, **30**). T-type nucleobases are also Py (**2**, **39**, **23**, **31**). U-type cluster is a mixture of Py

Table 10. Experimental and Calculated Structural Aromaticity Indices[a]

no.	A_{exp}	A_{HF}	A_{PM3}	A_{PLS}	$\langle d_{\mathrm{exp}}\rangle$	$\langle d_{\mathrm{HF}}\rangle$	$\langle d_{\mathrm{PM3}}\rangle$	$\langle d_{\mathrm{PLS}}\rangle$
1	0.859(36)	0.745(49)	0.741(52)	0.922(25)	1.369	1.364	1.402	1.373
2	0.933(34)	0.920(38)	0.894(45)	0.911(36)	1.374	1.370	1.411	1.374
3	0.868(33)	0.785(41)	0.837(35)	0.930(24)	1.382	1.384	1.421	1.394
4	0.899(25)	0.805(35)	0.829(32)	0.880(28)	1.380	1.385	1.420	1.384
5	0.892(34)	0.817(45)	0.875(36)	0.821(44)	1.360	1.348	1.387	1.359
6	0.889(67)	0.796(91)	0.847(77)	0.821(83)	1.359	1.350	1.392	1.359
7	0.863(63)	0.710(92)	0.829(68)	0.766(80)	1.368	1.361	1.402	1.370
8	0.853(55)	0.739(74)	0.807(64)	0.789(69)	1.368	1.358	1.401	1.368
9	0.858(30)	0.790(37)	0.842(31)	0.919(23)	1.387	1.384	1.420	1.389
10	0.802(181)	0.661(236)	0.706(213)	0.844(159)	1.366	1.366	1.403	1.377
11	0.827(113)	0.735(143)	0.745(133)	0.786(133)	1.360	1.362	1.400	1.374
12	0.824(93)	0.719(114)	0.746(105)	0.852(82)	1.373	1.370	1.405	1.379
13	0.867(136)	0.816(161)	0.889(121)	0.848(145)	1.360	1.354	1.393	1.363
14	0.900(50)	0.833(65)	0.824(66)	0.902(50)	1.391	1.387	1.410	1.386
15	0.760(197)	0.582(259)	0.742(197)	0.929(107)	1.376	1.379	1.414	1.372
16	0.896(95)	0.797(132)	0.809(124)	0.874(105)	1.385	1.384	1.424	1.380
17	0.818(133)	0.675(175)	0.720(156)	0.876(109)	1.370	1.368	1.404	1.379
18	0.876(123)	0.806(157)	0.884(118)	0.825(150)	1.371	1.362	1.394	1.366
19	0.911(156)	0.865(192)	0.901(159)	0.901(164)	1.369	1.370	1.407	1.371
20	0.856(118)	0.757(154)	0.852(118)	0.859(117)	1.378	1.372	1.403	1.376
21	0.896(134)	0.764(202)	0.870(146)	0.803(184)	1.379	1.371	1.413	1.378
22	0.885(87)	0.757(129)	0.880(87)	0.792(124)	1.382	1.376	1.409	1.384
23	0.895(78)	0.765(116)	0.836(94)	0.918(69)	1.386	1.385	1.422	1.388
24	0.836(94)	0.808(103)	0.849(88)	0.841(93)	1.392	1.385	1.421	1.390
25	0.938(109)	0.853(170)	0.899(133)	0.870(158)	1.374	1.369	1.415	1.375
26	0.900(102)	0.734(167)	0.773(150)	0.767(155)	1.378	1.366	1.404	1.380
27	0.835(71)	0.679(98)	0.721(91)	0.871(61)	1.371	1.368	1.404	1.378
28	0.876(111)	0.815(137)	0.878(108)	0.746(158)	1.361	1.349	1.387	1.364
29	0.883(81)	0.674(135)	0.738(116)	0.912(70)	1.369	1.368	1.404	1.375
30	0.821(129)	0.684(174)	0.736(154)	0.912(92)	1.372	1.368	1.404	1.375
31	0.884(109)	0.832(131)	0.861(116)	0.905(98)	1.385	1.387	1.422	1.394

[a] Julg's aromaticity index based on experimental (A_{exp}) and calculated bond lengths from models 50 (A_{HF}), 48 (A_{PM3}), and 34 (A_{PLS}) from Table 7. Errors are in brackets, given at last 2−3 digits. Average bond lengths (in Å) from experiment ($\langle d_{\mathrm{exp}}\rangle$) and from models 50, 48, and 34 ($\langle d_{\mathrm{HF}}\rangle$, $\langle d_{\mathrm{PM3}}\rangle$, $\langle d_{\mathrm{PLS}}\rangle$, respectively). Their errors are at most 0.001 Å.

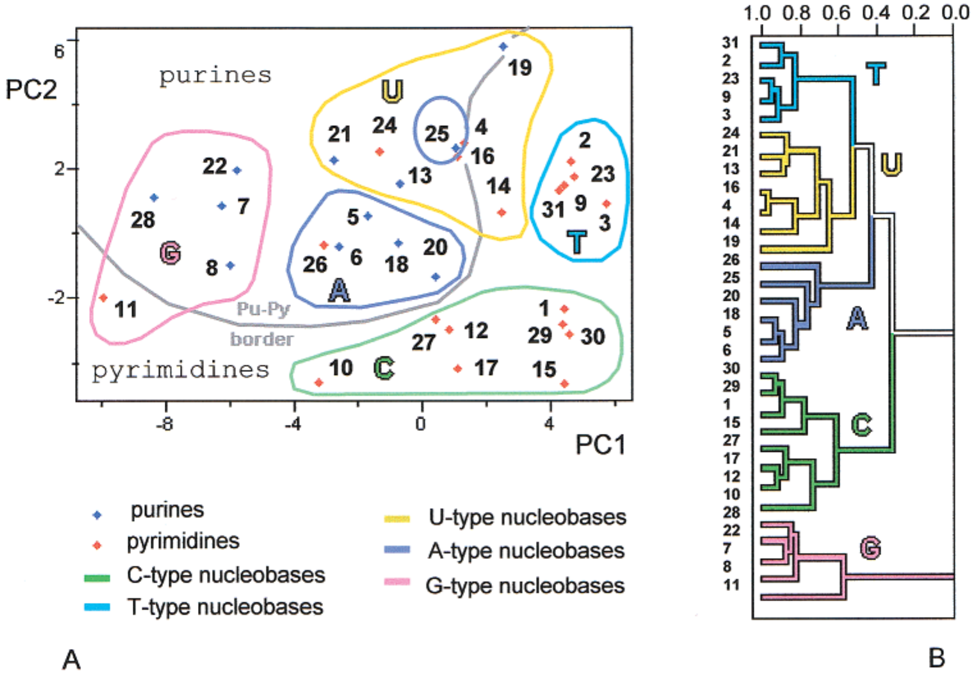

Figure 13. A. Scores plot for Julg's index data set. B. The corresponding HCA dendogram. Distinction between Pu and Py is visible. Five types of nucleobases, C-, T-, U-, A-, and G-type, are marked with different colors.

(**4**, **14**, **24**) and nontypical Pu (**19** − with tertiary amine in the ring; **13** − three fused rings; **21** − Py fragment equal to **4**). According to this PCA-HCA analysis of the index *A*, Py and Py-type (C, T, U) nucleobases possess a higher degree of π-electron delocalization than Pu and Pu-type (A, G) nucleobases. The very same analysis on $\langle d\rangle$ (see section **d.**) gives another classification, but there is still a good separation of Pu from Py and among the five clusters (C, T, U, A, G). The Varimax rotated PCA and HCA on experimental and calculated bond lengths data show that all the 309 bonds

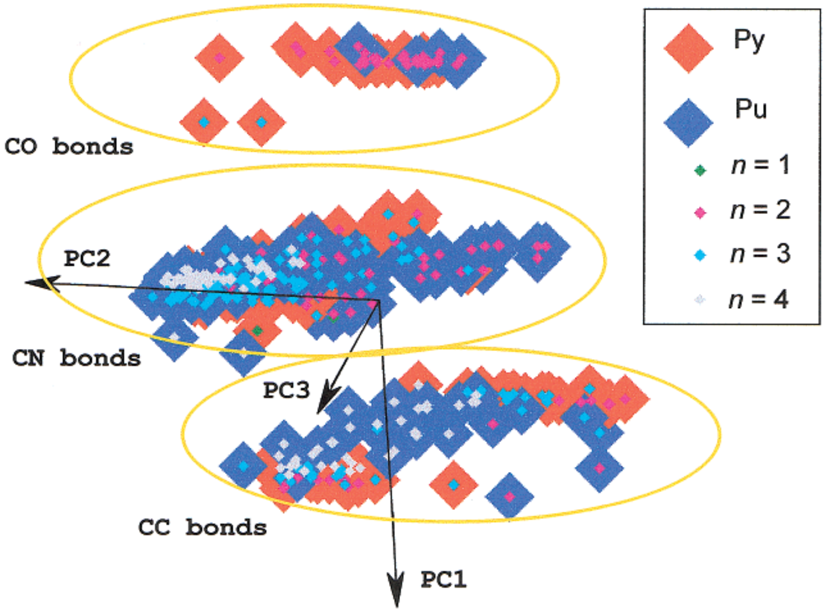

Figure 14. 3D scores plot on bond lengths data. Bond type, neighborhood, and Pu/Py distinction can be observed as clusters.

from set I+II can be described by two PCs (92.9% total variance). PC1 is highly correlated to Q (corr coeff $r = -0.975$) and can be interpreted as the bond length determined by the bond type (see the distinct clusters in Figure 14). PC2 is well related to π-bond orders ($r = -0.921$ with p_P) and slightly to n ($r = 0.383$). PC3 (3.7% total variance) is not significantly correlated to any bond descriptor ($|r| < 0.186$) and could be originated from properties of the calculation procedures. Thus, PC2 can be interpreted as a bond length defined by the local neighborhood effects (electronic, steric, and other properties). Bond lengths in Pu and Py nucleobases can be distinguished well in the CC cluster and much less in the CN and CO clusters. Bonds from Pu are more concentrated at lower PC2 than those from Py. HCA confirms the basic trends in this PCA.

(d) The Intermolecular Interactions in H−Bond Complexes of Nucleobases. Krygowski[76] showed that geometry and structural aromaticity indices can be affected by intermolecular H-bonding net, presence of cations, and push−pull cooperative substituent interactions. Molecular and crystal structures of cytidine in cytidine crystal (CYTIDI02) and in DNA decanucleotide (BD0023) reveal regular relations between molecular geometry, aromaticity, and intermolecular interactions. The reference cytidine molecule in CYTIDI02 is surrounded by 12 neighbors (through 86 contacts), being with two of them in $\pi...\pi$ stacking (interplanar distance between Py rings 3.740 Å). There are 10 hydrogen bonds established between the reference and six neighbors (C-1 to C-6, Figure 15A) and also an intramolecular H in the sugar. Groups $-NH_2$, C═O, and aromatic $-N=$ are involved in H-bonds. CC bond b is included in weak C−H...X interactions. Similarly, cytidine in DB0023 is surrounded by its phosphate and three nucleobases (T, A, G) and two waters; five H-bonds are established with the waters and the G base (Figure 15B). Comparison of CYTIDI02 and calculated (from dimer to hexamer) geometry can quantify the influence of packing effects on molecular properties (Table 11). Ab initio, especially B3LYP, are very close to the experiment, while MMFF is the worst. All the methods, especially semiempirical, are getting close to the experiment with the increase of the molecular cluster (more H-bonds and other intermolecular interactions). The bond length differences $\delta = d_{exp} - d_{B3LYP}$ for monomer (Figure 15A) outline the effect of the H-bond distribution around the cytosine system on its bond lengths. As H-bonds are the strongest intermolecular interactions (ranging 2−8 kcal mol^{-1} in this case, according to Gavezzotti[81]), the largest δ are observed for bonds with H-bond donors/acceptors (Figure 15A); acceptor-carbon bonds (C═O, −C═N-) are mainly prolongated, and the donor−carbon bond (C−NH_2) is shortened with respect to the free state. Large δ for bond h is caused by resonance assisted H-bonding, sugar-nucleobase steric interactions, and electronic effects in the Py ring; the ring adapts to the resonance changes provoked by the H-bonding. The CSD search by Slowikowska and Wozniak[82] for adenine residues also confirmed the existence of correlations between the structural parameters of adenine and of the H-bonding. Six-membered cytosine ring in CYTIDI02 is heteroaromatic (*HOMA*: 0.832, *A*: 0.890(27)), even including the two exocyclic bonds (*HOMA*: 0.456, *A*: 0.682(53)). Surprisingly, the B3LYP results for monomer, representing the free cytidine state, show lower degree of

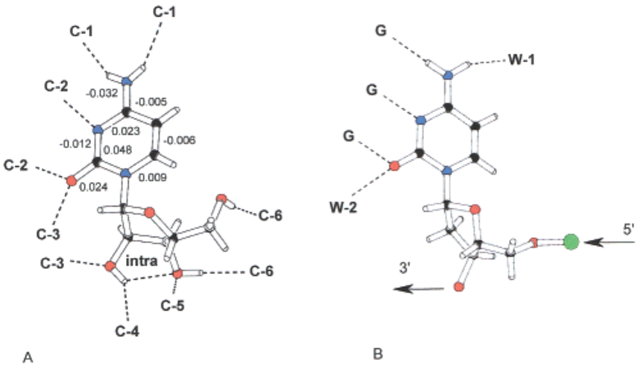

Figure 15. Cytidine residue in crystal structure (A) of cytidine (CYTIDI02) and (B) of a B-DNA decanucleotide (BD0023). Hydrogen bonds with neighboring cytidine (C-1 to C-6), guanosine (G), and water (W-1 and W-2) as well as an intra bond are marked with dashed lines. δ differences (see text) for cytosine (in Å) are presented also.

Table 11. Statistical Parameters[a] for Cytidine Clusters[b] with Geometry Optimized by Various Methods[d]

cluster[d]	method	Δ_{max}/ Å	$\langle\Delta\rangle$/ Å	σ/Å	$\langle\Delta/\sigma\rangle$	$\Delta HOMA$	ΔA
monomer (0)	MMFF	0.087	0.043	0.052	42.63	0.599	0.468(98)
	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 (2)	MMFF	0.086	0.042	0.050	41.88	0.605	0.472(98)
	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 (4)	MMFF	0.057	0.027	0.032	27.13	0.182	0.066(80)
	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 (6)	MMFF	0.052	0.021	0.026	21.00	0.388	0.269(90)
	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 (7)	MMFF	0.044	0.024	0.028	24.25	0.384	0.263(90)
	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 (9)	MMFF	0.051	0.021	0.027	21.38	0.453	0.289(91)
	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} (from CYTIDI02) 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 (whose geometry is studied) and neighboring molecules varying form zero (monomer) to five (hexamer). [c] Calculation methods: molecular mechanics MMFF94; semiempirical MNDO, AM1, PM3; ab initio − Hartree−Fock (HF) and DFT with B3LYP functional (B3LYP). [d] The number of hydrogen bonds between the referent and neighboring molecules is given in the brackets.

aromaticity for the Py ring excluding (*HOMA*: 0.752, *A*: 0.760(39)) and including (*HOMA*: 0.292, *A*: 0.464(69)) CO and NH_2 groups.

3. CONCLUSIONS

CC, CN, and CO nucleobase bond lengths depend on π-bond orders (p_P, p_s), electrotopological (Q), and topological (n) indices accounting for π-electron delocalization effects, bond type, and neighborhood, respectively. These bond lengths are 3D phenomenon (PC1 is strongly related to bond orders, PC2 to Q, PC3 to n) and can be classified in nine classes based on Q and n. The bond length prediction by traditional BLBORs can be improved by introducing crystal packing effects into bond orders and use of multivariate techniques and can compete with semiempirical results, as has been successfully shown in this work. The choice of the best model to predict a particular property can be performed easily by the PCA-HCA procedure. The nucleobase bonds are similar to PB-PAHs bonds in some aspects, as in data degeneration which can be reduced by suitable bond descriptors. There is no clear picture that the nucleobase bonds are less aromatic than those in PB-PAHs and picrates. It is certain that nucleobases are heteroaromatic systems rather distinct from picrates, PAHs and aza-PAHs. Nucleobase bond lengths in crystal, compared to a free state, are affected by substitution and crystal packing effects, and even calculation methods. Quantum mechanical treatment of molecular hydrogen bonding clusters results in nucleobase geometry close to the experimental. 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.

ACKNOWLEDGMENT

The authors aknowledge FAPESP for finantial support.

Supporting Information Available: List of CSD REFCODES with references for nucleobase crystal structures used and tables of bond descriptors, experimental and calculated bond lengths. This material is available free of charge via Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hurst, D. T. *An Introduction to the Chemistry and Biochemistry of Pyrimidines, Purines and Pteridines*; Wiley: Chichester, UK, 1980; p 104.

(2) Chaudhuri, N. C.; Ren, R. X.-F.; Kool, E. T. C-Nucleosides Derived from Simple Aromatic Hydrocarbons. *Synlett* **1997**, 341−347.

(3) Saenger, W. *Principles of Nucleic Acid Structure*; Springer: New York, 1984.

(4) Barciszewski, J.; Barciszewska, M. Z.; Siboska, G.; Rattan, S. I. S.; Clark, B. F. C. Some unusual nucleic acid bases are products of hydroxyl radical oxidation of DNA and RNA. *Mol. Biol. Rep.* **1999**, *26*, 231−238.

(5) Ray, A.; Norden, B. Peptide nucleic acid (PNA): its medical and biotechnical applications and promise for the future. *Faseb J.* **2000**, *14*, 1041−1060.

(6) Barnes, T. W., III.; Turner, D. H. Long-Range Cooperativity in Molecular Recognition of RNA by Oligodeoxynucleotides with Multiple C5(1-Propynyl) Pyrimidines. *J. Am. Chem. Soc.* **2001**, *123*, 4107−4118.

(7) Matulic-Adamic, J.; Beigelman, L.; Portmann, S.; Egli, M.; Usman, N. Synthesis and Structure of 1-Deoxy-1-phenyl-β-D-ribofuranose and Its Incorporation into Oligonucleotides. *J. Org. Chem.* **1996**, *61*, 3909−3911.

(8) Ren, R. X.-F.; Chaudhuri, N. C.; Paris, P. L.; Rumney, S., IV.; Kool, E. T. Naphthalene, Phenanthrene, and Pyrene as DNA Base Analogues: Synthesis, Structure, and Fluorescence in DNA. *J. Am. Chem. Soc.* **1996**, *118*, 7671−7678.

(9) Guckian, K. M.; Krugh, T. R.; Kool, E. T. Solution Structure of a Nonpolar, Non-Hydrogen-Bonded Base Pair Surrogate in DNA. *J. Am. Chem. Soc.* **2000**, *122*, 6841−6847.

(10) Schwogler, A.; Carell, T. Toward catalytically active oligonucleotides: Synthesis of a flavin nucleotide and its incorporation into DNA. *Org. Lett.* **2000**, *2*, 1415−1418.

(11) Clowney, L.; Jain, S. C.; Srinivasan, A. R.; Westbrook, J.; Olson, W. K.; Berman, H. M. Geometric Parameters in Nucleic Acids: Nitrogenous Bases. *J. Am. Chem. Soc.* **1996**, *118*, 509−518.

(12) Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S. H.; Srinivasan, A. R.; Schneider, B. The Nucleic-Acid Database: A Comprehensive Relational Database of 3-Dimensional Structures of Nucleic-Acids. *Biophys. J.* **2000**, *63*, 751−759. Website: The Nucleic Acid Database. The Nucleic Acid Database Project. Department of Chemistry and Chemical Biology, Rutgers, the State University of New Jersey, Piscataway, NJ, 1995−2001. [http://ndbserver.rutgers.edu/structure-finder/ndb/].

(13) The NMR−Nucleic Acid Database. The Nucleic Acid Database Project. Department of Chemistry and Chemical Biology, Rutgers, the State University of New Jersey, Piscataway, NJ, 1995−2001. [http://ndbserver.rutgers.edu/structure-finder/nmr/].

(14) The DNA-Binding Protein Database. The Nucleic Acid Database Project. Department of Chemistry and Chemical Biology, Rutgers, the State University of New Jersey, Piscataway, NJ, 1995−2001. [http://ndbserver.rutgers.edu/structure-finder/dnabind/].

(15) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242. Website: PDB − Protein Data Bank. The Research Collaboratory for Structural Bioinformatics (RCSB): Rutgers, the State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; the National Institute of Standards and Technology; and the University of Wisconsin-Madison. [http://www.rcsb.org/pdb/].

(16) Alen, F. H.; Kennard, O. 3D Search and Research Using the Cambridge Structural Database. *Chem. Design Autom. News* **1993**, *8*, 31−37. Website: The Cambridge Structural Database. Cambridge Crystal-

lographic Data Centre, University of Cambridge, Cambridge, UK. [http://www.ccdc.cam.ac.uk/prods/csd/csd.html].

(17) Guckian, K. M.; Schweitzer, B. A.; Ren, R. X.-F.; Sheils, C. J.; Paris, P. L.; Tahmassebi, D. C.; Kook, E. T. Experimental Measurement of Aromatic Stacking Affinities in the Context of Duplex DNA. *J. Am. Chem. Soc*. **1996**, *118*, 8182−8183.

(18) Guckian, K. M.; Schweitzer, B. A.; Ren, R. X.-F.; Sheils, C. J.; Paris, P. L.; Tahmassebi, D. C.; Kook, E. T. Factors Contributing to Aromatic Stacking in Water: Evaluation in the Context of DNA. *J. Am. Chem. Soc*. **2000**, *118*, 2213−2222.

(19) Voet, D.; Voet, J. G. *Biochemistry*, 2nd ed.; J. Wiley: New York, 1995; p 868.

(20) Lemieux, S.; Oldziej, S.; Major, F. Nucleic Acids: Qualitative Modeling. In *Encyclopedia of Computational Chemistry*; von Ragué Schleyer, P., Ed.; Wiley: Chichester, UK, 1998; Vol. 3, pp 1930−1941.

(21) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem*. **1964**, *68*, 441−451.

(22) Burgi, H. B. Structure correlation and chemistry. *Acta Crystallogr.* **1998**, *A54*, 873−885.

(23) von Ragué Schleyer, P.; Jiao, H. What is aromaticity? *Pure Appl. Chem*. **1996**, *68*, 209−218.

(24) Lewis, D.; Peters, D. *Facts and Theories of Aromaticity*, 1st ed.; MacMillan Press: London, 1975; pp 10−14.

(25) Bird, C. W. Heteroaromaticity. 13. Bond Alternation and Its Consequences for Conjugation Energies and Other Criteria of Aromaticity. *Tetrahedron* **1998**, *54*, 4641−4646.

(26) Cramer, C. J. Hyperconjugation. In *Encyclopedia of Computational Chemistry*; von Ragué Schleyer, P., Ed.; Wiley: Chichester, UK, 1998; Vol. 2, pp 1294−1298.

(27) Pauling, L. *The Nature of the Chemical Bond*, 3rd ed.; Cornell University Press: Ithaca, NY, 1972.

(28) Whealand, G. W. *The Theory of Resonance*; Wiley: New York, 1953.

(29) Pauling, L. Bond Numbers and Bond Lengths in Tetrabenzo[*de*, *no*, *st*, c_1d_1]heptacene and Other Condensed Aromatic Hydrocarbons: A Valence-Bond Treatment. *Acta Crystallogr.* **1980**, *B36*, 1898−1901.

(30) Herndon, W. C.; Párkányi, C. π Bond Orders and Bond Lengths. *J. Chem. Educ*. **1976**, *53*, 689−691.

(31) Kiralj, R.; Kojić-Prodić, B.; Žinić, M.; Alihodžić, S.; Trinajstić, N. Bond Length-Bond Order Relationships and Calculated Geometries for some Benzenoid Aromatics, Including Phenanthridine. Structure of 5,6-Dimethylphenanthridinium Triflate, [*N*-(6-Phenanthridinylmethyl)-aza-18-crown-6-κ^5*O,O′,O″,O‴,O⁗*](picrate-κ^2*O,O′*)potassium, and [*N,N′*-Bis(6-phenanthridinyl-κ*N*-methyl)-7, 16-diaza-18-crown-6-κ^4*O,O′,O″,O‴*] sodium Iodide Dichloromethane Solvate. *Acta Crystallogr.* **1996**, *B52*, 823−837.

(32) Kiralj, R.; Kojić-Prodić, B.; Nikolić, S.; Trinajstić, N. Bond lengths and bond orders in benzenoid hydrocarbons and related systems: a comparison of valence bond and molecular orbital treatments. *J. Mol. Struct. (THEOCHEM)* **1998**, *427*, 25−37, and the references therein.

(33) Kiralj, R.; Kojić-Prodić, B.; Piantanida, I.; Žinić, M. Crystal and molecular structures of diazapyrenes and a study of $\pi ... \pi$ interactions. *Acta Crystallogr.* **1999**, *B55*, 55−69.

(34) Stoicheff, B. P. Variation of Carbon−Carbon Bond Lengths With Environment as Determined by Spectroscopic Studies of Simple Polyatomic Molecules. *Tetrahedron* **1962**, *17*, 135−145.

(35) Kiralj, R. Structural Investigations of Macrocyclic Receptors with Phenanthridine Subunits. Master Thesis, University of Zagreb, Zagreb, Croatia, 1994.

(36) Krygowski, T. M.; Anulewicz, R.; Kruszewski, J. Crystallographic Studies and Physicochemical Properties of π-Electron Compounds. III. Stabilization Energy and the Kekulé Structure Contributions Derived from Experimental Bond Lengths. *Acta Crystallogr*. **1983**, *B39*, 732−739.

(37) Kiralj, R. Structural Studies of 4,9-Diazapyrene Derivatives. Doctoral Thesis, University of Zagreb, Zagreb, Croatia, 1999.

(38) Kiralj, R.; Ferreira, M. M. C. Predicting Bond Lengths in Planar Benzenoid Polycyclic Aromatic Hydrocarbons: A Chemometric Approach. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 508−523.

(39) Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics: A Practical Guide*; Wiley: New York, 1998.

(40) Beebe, K. R.; Kowalski, B. P. An Introduction to Multivariate Calibration and Analysis. *Anal. Chem*. **1987**, *59*, 1007A−1017A.

(41) Schleyer, P. R.; Freeman, P. K.; Jiao, H.; Goldfuss, B. Aromaticity and Antiaromaticity in Five-Membered C_4H_4X Ring Systems: ”Classical” and “Magnetic” Concepts May Not Be “Orthogonal”. *Angew. Chem., Int. Ed. Engl*. **1995**, *34*, 337−340.

(42) Lewis, D.; Peters, D. *Facts and Theories of Aromaticity*, 1st ed.; The MacMillan Press Ltd.: London, 1975; pp 10−14.

(43) Clowney, L.; Westbrook, J. D.; Berman, H. M. CIF applications. XI. A La Mode: a ligand and monomer object data environment. I. Automated construction of mmCIF monomer and ligand models. *J. Appl. Crystallogr*. **1999**, *32*, 125−133. See the Internet update of the nucleic acid standards at the following: À la mode: A Ligand And Monomer Object Data Environment. The Nucleic Acid Database. The Nucleic Acid Database Project. Department of Chemistry and Chemical Biology, Rutgers, the State University of New Jersey, Piscataway, NJ, 1995−2001. http://ndbserver.rutgers.edu/alamode/index.html [last accessed on 30 June 2001].

(44) Cambridge Structural Database, October Release 2001. The Chemistry Visualization Program (NCSA ChemViz) at the National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign. http://chemviz.ncsa.uiuc.edu.

(45) Spek, A. L. *PLATON, A Multipurpose Crystallographic Tool*, v. 31000, Utrecht University: Utrecht, The Netherlands, 2000. http://www.cryst.chem.uu.nl/platon/.

(46) *CRC Handbook of Chemistry and Physics*, 74th ed.; Lide, D. R., Ed.; CRC Press: Boca Raton, FL, 1993; Section 9.

(47) *Titan*, v. 1, Wavefunction, Inc.: Irvine, CA, 2000.

(48) Kiralj, R.; Ferreira, M. M. C. A priori molecular descriptors in QSAR: A case of HIV-1 protease inhibitors. II. Molecular graphics and modeling. *J. Mol. Graph. Mod*., in press.

(49) Allen, F. H.; Kennard, O.; Watson, D. G.; Brammer, L.; Orpen, A. G.; Taylor, R. Tables of Bond Lengths determined by X-ray and Neutron Diffraction. Part 1. Bond Lengths in Organic Compounds. *J. Chem. Soc., Perkin Trans. 2* **1987**, S1−S19.

(50) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Fields for the Simulation of Protein, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc*. **1995**, *117*, 5179−5197.

(51) Dewar, M. J. S.; Gleicher, G. J. Ground States of Conjugated Molecules. VII. Compounds Containing Nitrogen and Oxygen. *J. Chem. Phys*. **1966**, *44*, 759−773.

(52) Cieplak, P. Nucleic Acid Force Fields. In *Encyclopedia of Computational Chemistry*; von Ragué Schleyer, P., Ed.; Wiley: Chichester, UK, 1998; Vol. 3, pp 1922−1930.

(53) Hobza, P.; Kabelac, M.; Sponer, J.; Mejzlik, P.; Vondrasek, J. Performance of Empirical Potentials (AMBER, CFF95, CBF, CHARMM, OPLS, POLTEV), Semiempirical Quantum Chemical Methods (AM1, MNDO/M, PM3), and Ab Initio Hartree−Fock Method for Interaction of DNA Bases: Comparison with Nonempirical Beyond Hartree−Fock Results. *J. Comput. Chem*. **1997**, *18*, 1136−1150.

(54) Bludsky, O.; Sponer, J.; Leszczynski, J.; Spirko, V.; Hobza, P. Amino groups in nucleic acid bases, aniline, aminopyridines, and aminotriazines are nonplanar: Results of correlated ab initio quantum chemical calculations and anharmonic analysis of the aniline inversion calculation. *J. Chem. Phys*. **1996**, *105*, 11042−11050.

(55) Kiralj, R. RESON. A program for determination of weights of Kekulé structures by the method of Krygowski. Rudjer Boskovic Institute: Zagreb, Croatia, 1998.

(56) Halgren, T. A. Merck molecular force field. 1. Basis, form, scope, parametrization, and performance of MMFF94. *J. Comput. Chem*. **1996**, *17*, 490−519.

(57) Dewar, M. J. S.; Thiel, W. Ground-States of Molecules .38. MNDO Method: Approximations and Parameters. *J. Am. Chem. Soc*. **1977**, *99*, 4899−4907.

(58) Dewar, M. J. S.; Zoebisch, E. G.; Healy, G. F.; Stewart, J. J. P. The Development and Use of Quantum-Mechanical Molecular Models .76. AM1: A New General-Purpose Quantum-Mechanical Molecular Model. *J. Am. Chem. Soc*. **1985**, *107*, 3902−3909.

(59) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods .1. Methods. *J. Comput. Chem*. **1989**, *10*, 209−220.

(60) Allen, F. H.; Motherwell, W. D. S. Applications of the Cambridge Structural Database in organic chemistry and crystal chemistry. *Acta Crystallogr*. **2002**, *B58*, 407−422.

(61) Bobadova-Parvanova, P.; Galabov, B. Ab Initio Molecular-Orbital Study on Hydrogen-Bonded Complexes of Carbonyl Aliphatic Compounds and Hydrogen Fluoride. *J. Phys. Chem*. **1998**, *A102*, 1815−1819.

(62) Kiralj, R.; Ferreira, M. M. C. A Combined Computational and Chemometric Study of Indole-3-Acetic Acid. *Int. J. Quantum Chem*. submitted for publication.

(63) Kiralj, R.; Ferreira, M. M. C. A Molecular and Quantum Mechanical Study of Indole-3-Acetic Acid. XI Simpósio Brasileiro de Química Teórica; Caxambu, MG, Brazil, 18−21 November 2001. Poster P302.

(64) Bertolasi, V.; Gilli, P.; Ferreti, V.; Gilli, G. Intermolecular N−H···O Hydrogen Bonds Assisted by Resonance. Heteroconjugated Systems as Hydrogen-Bond-Strengthening Functional Groups. *Acta Crystallogr*. **1995**, *B51*, 1004−1015.

(65) Taylor, R.; Kennard, O. Accuracy of Crystal Structure Error Estimates. *Acta Crystallogr*. **1986**, *B42*, 112−120.

(66) Allen, F. H. A Systematic Pairwise Comparison of Geometric Parameters Obtained by X-ray and Neutron Diffraction. *Acta Crystallogr*. **1986**, *B42*, 515−522.

(67) Pauling, L. Atomic Radii and Interatomic Distances in Metals. *J. Am. Chem. Soc.* **1947**, *69*, 542−553.
(68) Gordy, W. Dependence of Bond Order and of Bond Energy Upon Bond Length. *J. Chem. Phys.* **1947**, *15*, 305−310.
(69) Pirouette 3.01. Infometrix, Inc.; Seattle: WA, 2001.
(70) Matlab 6.1. MathWorks, Inc.; Natick, MA, 2001.
(71) Ferreira, M. M. C.; Montanari, C. A.; Gaudio, A. C. Variable Selection in QSAR. *Quím. Nova* **2002**, *25*, 439−448.
(72) Ferreira, M. M. C.; Antunes, A. M.; Melgo, M. S.; Volpe, P. L. O. Chemometrics I: Multivariate calibration, a tutorial. *Quím. Nova* **1999**, *22*, 724−731.
(73) Kubinyi, H. Evolutionary Variable Selection in Regression and PLS Analyses. *J. Chemometrics* **1996**, *10*, 119−133.
(74) Cho, S. J. *The QSAR Server: Partial Least Squares (PLS)*, Laboratory for Molecular Modeling, School of Pharmacy, University of North Carolina, Chapel Hill, NC. http://mmlin1.pha.unc.edu/∼jin/QSAR/PLS/pls.html [last accessed on 23 July 2002].
(75) Ferreira, M. M. C.; R. Kiralj. Chemomeric and QSPR Analysis of a Set of Hydropathy Scales. 24ª Reunião Anual da Sociedade Brasileira de Química; Poços de Caldas, MG, Brazil, 20−23 May 2002. Poster QT040.
(76) Krygowski, T. M. Crystallographic Studies of Inter- and Intramolecular Interactions Reflected in Aromatic Character of π-Electron Systems. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 70−78.
(77) Frisch, M. J.; G. W. Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. Gaussian 98 − Revision A.11, Gaussian, Inc., Pittsburgh, PA, 2001.
(78) Kielkopf, C. L.; Ding, S.; Kuhn, P.; Rees, D. C. Conformational Flexibility of B-DNA at 0.74 Å Resolution: d(CCAGTACTGG)$_2$. *J. Mol. Biol.* **2000**, *296*, 787−801.
(79) Glusker, J. P.; Lewis, M.; Rossi, M. *Crystal Structure Analysis for Chemists and Biologists*; VCH Publishers: New York, 1994; pp 429−431.
(80) Trinajstic, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.
(81) Gavezzotti, A. Are Crystal Structures Predictable? *Acc. Chem. Res.* **1994**, *27*, 309−314.
(82) Slowikowska, J. M.; Wozniak, K. Influence of hydrogen bonding on the geometry of the adenine fragment. *J. Mol. Struct.-Theochem* **1996**, *374*, 327−337.

CI0200523