Modeling Purines and Pyrimidines with the Linear Combination of Connectivity Indices−Molecular Connectivity "LCCI-MC" Method

Lionello Pogliani†

Dipartimento di Chimica, Università della Calabria, 87030 Rende (CS), Italy

Received February 22, 1996⊗

A series of five experimental properties of DNA−RNA bases (U, T, A, G, and C): singlet excitation energies ΔE_1 and ΔE_2, oscillator strengths f_1 and f_2, and molar absorption coefficient ϵ_{260} plus three experimental properties of a wider set of purine and pyrimidine bases: average $\langle pK \rangle = {}^1/_2(pK_a + pK_b)$, molecular weights MW, and solubility have been simulated in two different ways with linear combinations of connectivity indices (LCCI) chosen from a medium sized molecular connectivity $\{\chi\} = \{D,D^v,{}^0\chi,{}^0\chi^v,{}^1\chi,{}^1\chi^v,\chi_t,\chi_t{}^v\}$ set. A forward selection technique and a full combinatorial space technique have been used to choose the best linear combination of connectivity indices for an optimal modeling. The given properties are very well modeled with the only exception being $\langle pK \rangle$, whose modeling could be improved with the introduction of fragment reciprocal connectivity indices, that take into account the number of basic and acid groups of the given molecules. The (easier to perform) forward selection technique is in many occasions a good alternative to the more cumbersome full space selection technique and can normally be used to restrict the dimension of the full combinatorial space, thus, facilitating the computation. Limits in the forward selection method can frequently be overcome with the introduction of orthogonal indices. While the simulation of the molecular weights cast some light on the modeling of hydrogen-rich or -poor molecules, the simulation of the solubility shows (i) how far a satisfactory modeling of a small number of compounds can be extrapolated by the aid of the same indices to a wider set, (ii) the importance of linear combinations of squared connectivity indices used in the absolute value mode, and (iii) the contribution of supramolecular connectivity indices for an improved modeling of the solubility. The positive role of the χ_t and $\chi_t{}^v$ indices, all along the modeling of the different properties, seems to be due to the rather low collinearity of these indices relative to the other indices of the connectivity set, a fact that underlines their use in molecular modeling with linear combinations of connectivity indices. In an Appendix, the notation of δ cardinal number is extended to the triplet code words to generate the different families and subfamilies of the genetic code.

INTRODUCTION

The main objective of the method of linear combinations of connectivity indices (LCCI) is to find an answer to the following three questions: (i) which physicochemical property can be modeled, (ii) of which kind of molecules, and (iii) with which type and number of molecular connectivity indices can a modeling be optimal. This method uses the well-known molecular connectivity indices,[1−16] that are graph theoretical topological indices defined within the frame of molecular connectivity MC theory.[1−5] One of the drawbacks of the method is the quest for the experimental values of physicochemical properties of a large set of molecules. Normally, due to the lack of a sufficiently large database one resorts in general to studying as many experimental properties as possible of small molecular sets with the hope to derive some interesting properties, limits, and hints for further molecular modeling studies.

Recently, a linear combination of connectivity indices succeeded in modeling many different physicochemical experimental properties of natural amino acids,[17−19] of a set of alkanes, caffeine derivatives, organophosphorus compounds,[20] and of unsaturated organic compounds (inclusive of the cis−trans isomerism).[21] Even a set of inorganic salts[22] has recently been succesfully modeled. Some interesting results and hints have been obtained from these last studies. As a rule, the valence molecular connectivity indices[3−5] play a central role in every modeling where different values for the connectivity and valence connectivity indices can be derived. The modeling of the solubility of caffeine homologs allowed to introduce and test the supramolecular connectivity indices that together with linear combination of squared connectivity indices (LCSCI) offered an impressive modeling for this property.[20] Linear combinations of orthogonal connectivity indices (LCOCI), introduced recently by Randić,[23−26] could sometimes improve the estimation of a property and allow one to find out, in many cases, the dominant descriptor (or descriptors) of an experimental property, whenever normal connectivity indices rate poorly. Furthermore, while special ordering-dependent orthogonal indices have recently been defined and tested with success on different properties of amino acids[27] it was also detected that linear combinations of reciprocal connectivity indices (LCRCI)[28,29] were better descriptors of the solubility of the natural amino acids than linear combinations of fragment connectivity indices introduced to estimate the pH at the isoelectric point of natural amino acids.[19,30] Along the last two works[28,29] it has been shown that a modeling matrix equation based on the absolute values normally improved the corresponding estimation getting rid, furthermore, of negative simulated values for the experimental properties.

† From November 1995 till October 1996 on sabbatical leave at the Centro de Quimica Fisica Molecular IST, Av. Rovisco Pais, 1096 Lisboa-Codex, Portugal.

⊗ Abstract published in *Advance ACS Abstracts,* October 1, 1996.

S0095-2338(96)00020-0 CCC: $12.00

The aim of the present paper is to widen the descriptive power of the LCCI-MC method applying it to the modeling of a set of molecules that, like the α-amino acids, are at the very core of the attention of biochemists, molecular biologists and pharmacologists, that is, purine and pyrimidine bases among which there are the well-known RNA and DNA bases: U (RNA only), T (DNA only), A, G, and C, that is, uracil, thymine, adenine, guanine, and cytosine. In fact, a special attention will be devoted to some experimental properties of these five bases that have also been studied by quantum theoretical methods:[31] first ΔE_1 singlet excitation energy, second ΔE_2 singlet excitation energy, first f_1 oscillator strength and second f_2 oscillator strength of the first singlet excitation energy and molar absorption ϵ_{260} coefficient at 260 nm and pH = 7. The following experimental properties of a set of $n = 12-25$ purines and pyrimidines will also be estimated: average $\langle pK \rangle$ value, molecular weight MW, and solubility. The modeling of the molecular weight of 25 purine and pyrimidine bases will, here, be used to unravel the effect of hydrogens in molecular modeling. In connection with this modeling the importance of the total connectivity index[10] will be checked as this index played an important role in modeling the properties of alkanes.[20]

METHOD

The set of connectivity indices used to encode the given experimental physicochemical properties of purines and pyrimidines is the following set of eight molecular connectivity indices: $\{\chi\} = \{D, D^v, {}^0\chi, {}^0\chi^v, {}^1\chi, {}^1\chi^v, \chi_t, \chi_t{}^v\}$. While ${}^i\chi$, ${}^i\chi^v$ indices were defined 20 years ago by Randić, Kier, and Hall[1-5] (RKH) and constitute the basic graph theoretical descriptors of molecular connectivity theory (a kind of topological molecular orbitals). Indices D, D^v, χ_t, and $\chi_t{}^v$ recently defined[16,10] are also rooted on δ and δ^v numbers, that are the raw data used for every kind of MC calculation (see Acknowledgment for another definition of D).

$$D = \sum_i \delta_i \tag{1}$$

$${}^m\chi = \sum_p (\delta_1 \delta_2 ... \delta_{m+1})^{-1/2} \tag{2}$$

In eq 1 summation runs over the number of non-hydrogen atoms (i varies from 1 to n). The summation in eq 2 runs over the m-order p paths ($m = 0, 1, 2, ..., n$, where n is the number of non-hydrogen atoms of a molecule) and subscripts 1, 2, 3, ..., etc., stand for vertex degrees (δ values) of successive adjacent non-hydrogen atoms. Thus, while for $m = 0$ we obtain the zeroth-order connectivity ${}^0\chi$ index and index p of the sum becomes i, for $m = 1$ the first-order connectivity ${}^1\chi$ index is obtained, where summation runs over the number of first-order paths, that is, the number of σ bonds. When m equals the total molecular path of the chemical graph, the total molecular connectivity χ_t index is obtained, and the summation boils down to a single term. The corresponding valence D^v and ${}^m\chi^v$ connectivity indices are obtained when the δ branching values are substituted by the corresponding δ^v valence values that include information on the valence shell.

To keep under control the collinearity among these indices it is a good choice to follow a collinearity criterion[32] that states that two χ indices show a strong interrelation if the regression coefficient R of their linear relationship is $0.98 < R < 1$, even if exclusion of an index is justified only when $R = 1$.[23-25,32] All along this study the mean interrelation coefficient of the interrelation matrix for a specific property P coded by a set of χ indices $\langle R_{IM}(P: \chi) \rangle$ will be used[18] to test the overall collinearity of a set of χ indices of a given property. The recently introduced molecular orthogonal connectivity ${}^i\Omega$ indices[23-26] will also be used to detect possible dominant components whenever ${}^i\chi$ indices are poor simulators.

Two standard procedures for searching the best LCCI based on spanning the combinatorial space of the $\{\chi\}$ set of connectivity indices will be applied: the forward selection method and the full selection method. The forward selection method is a sequential method for index selection based on the notion that connectivity indices should be inserted one at time until a satisfactory Q-LCCI ($Q = r/s$, where is r is the correlation coefficient and s the standard deviation of estimates[18]) is obtained. This method spans a subspace (36 combinations for eight χ indices) of the full combinatorial space. The method is as follows: (1) choose the χ index that gives the largest value of Q and we call this the best single LCCI, then (2) choose the next χ index of the $\{\chi\}$ set that when inserted in the model gives the largest increase in Q, in the presence of the previous index, and we call this the best 2-χ indices LCCI (or 2-χ indices LCCI), and so on till Q starts to decrease constanly with the introduction of the next χ index of the set. The more elaborated and precise full selection method, instead, searches the full combinatorial space (255 combinations for eight χ indices) spanned by the indices of the set. Combinations are also sorted following their F value ($F = fr^2/[(1-r^2)v]$, where f and v are the degrees of freedom and the number of variables, respectively[5]), that can be a valuable aid in discriminating among different Q-LCCI with different number of indices.[18-20] Former experience[20] has shown that the first method offers an adequate alternative to the more precise but more time consuming full combinatorial method.

The following dot product is the modeling equation for the experimental properties of a class of compounds

$$\mathbf{P} = \mathbf{C} \cdot \chi \tag{3}$$

but sometimes it is better to choose the following modeling equation that guarantees simulated positive properties

$$\mathbf{P} = |\mathbf{C} \cdot \chi| \tag{4}$$

where bars stand for absolute value. Normally, this equation not only gets rid of negative calculated properties but guarantees also a better Q/F modeling.[28,29] Vector χ includes the unitary $\chi^0 \equiv 1$ connectivity index, that renders **C** matrix and χ column vector formally conformable (for a single property **C** becomes a row vector and P a scalar) and converts the nonhomogeneous estimate of P into a homogeneous one. Values of **C** matrix are calculated by the aid of a multiple regression analysis.

While the different values of the connectivity indices were calculated following relations 1 and 2, $\langle pK \rangle = {}^1/_2(pK_a + pK_b)$ values, where pK_a and pK_b are the negative of the logarithm of the dissociation constant for the first two ionizable groups in purines and pyrimidines, experimental molecular weights MW, and solubility, Sol, of purines and pyrimidines were taken from refs 20 and 33. The experimental values for ΔE_1, ΔE_2, f_1, f_2, and ϵ_{260} of the DNA/RNA bases were taken from refs 31 and 34, respectively.

Table 1. Experimental (exp) Molar Absorption Coefficient $\epsilon_{260,exp}$ at 260 nm and pH = 7.0,[33] First $\Delta E_{1,exp}$ and Second $\Delta E_{2,exp}$ Singlet Excitation Energies in eV and First $f_{1,exp}$ and Second $f_{2,exp}$ Oscillator Strength Values (of the First Singlet Excitation Energies) of the Nucleotide DNA Bases[a] [31]

DNA bases	$\epsilon_{260,exp}/1000$	$\Delta E_{1,exp}$	$\Delta E_{2,exp}$	$f_{1,exp}$	$f_{2,exp}$
A	15.4	4.75	5.99	0.28	0.54
G	11.7	4.49	5.03	0.20	0.27
U	9.9	4.81	6.11	0.18	0.30
T	9.2	4.67	5.94	0.18	0.37
C	7.5	4.61	6.26	0.13	0.72

[a] A = adenine, G = guanine, U = uracil, T = thymine, C = cytosine.

Table 2. Experimental $\langle pK \rangle = {}^1/_2(pK_a + pK_b)$ Values, Molecular MW Weights, and Aqueous Solubility Sol (Units of Grams per 100 mL of Water) at the Shown Temperature (T, °C) of Purine and Pyrimidine Bases[a] [33]

PP bases	MW	Sol (T, °C)	PP bases	MW	Sol (T, °C)	$\langle pK \rangle$
7I8MTp	250.28	0.63 (20)	UA	168.11	0.002 (20)	8.35
7B8MTp	250.28	0.45 (20)	7MG	165.16		6.7
7ITp	236.28	2.7 (20)	OA	156.1	0.18 (18)	5.95
7BTp	236.25	0.37 (30)	X	152.11	0.05 (20)	4.1
1BTb	236.25	0.56 (30)	IsoG	151.13	0.006 (25)	6.75
7PTp	222.23	23.11 (30)	G	151.13	0.004 (40)	6.25
1PTb	222.23	1.38 (30)	5hyMC	141.13		8.65
7ETp	208.21	3.66 (30)	HypoX	136.11	0.07 (19)	5.45
1ETb	208.21	3.98 (30)	A	135.14	0.09 (25)	<2.5
Cf	194.19	2.58 (30)	T	126.11	0.40 (25)	>12
Tp	180.17	0.81 (30)	5MC	125.13	0.45 (25)	8.5
Tb	180.17	0.054 (30)	U	112.09	0.36 (25)	5
			C	111.1	0.77 (25)	8.35

[a] Abbreviations as in Table 1: B = butyl, Cf = caffeine = 137MMMX = 7MTp, hyM = M-OH, I = isobutyl, M = methyl, P = propyl, Tb = theobromine = 37MMX, Tp = theophylline = 13MMX, OA = orotic acid, UA = uric acid, X = xanthine.

RESULTS

The experimental data values of the DNA/RNA bases have been collected in Table 1. Abbreviations in this and other tables follow the method used by Kier and Hall.[6] In Table 2 the experimental $\langle pK \rangle$, MW, and Sol of purines and pyrimidines have been collected. The δ values of δ matrices are given in Table 3A−C. Figure 1 exemplifies for the case of adenine (A: Table 3B) and cytosine (C: Table 3A) how the positions of the δ values can be evaluated. A coding of the different bases by the aid of the δ and valence δ^v values of their non-hydrogen atoms has been chosen. These δ values have been collected into the form of δ and δ^v matrices throughout Table 3A−C, respectively. This kind of coding for molecules, that has already been used for amino acids,[17] (i) offers the possibility to avoid the use of time and space consuming figures, where each non-hydrogen atom has to be characterized by its own δ and valence δ value, (ii) gives a direct view of the δ values and of the connections among them, and (iii) avoids also the writing of the corresponding bulkier connectivity matrices, where δ values are indirectly coded. Clearly, not every molecule offers the possibility to formulate such a kind of δ matrices. At the top of these matrices in italics are given the positions of the δ values, that allow one to reconstruct the molecule and to detect eventual connections: normal (when no rings are present) connections take place horizontally, in the first row only, and vertically at each position. When rings are present, as is here the case, the first nonprimed (*1* for pyrimidines, *4* for purines, *1* in Tb derivatives, and *5* in Tp derivatives) position always connects with a position before a ";" sign (in a δ matrix this sign is reported when it is not in the last position), connections between positions on the same side or on both sides of ";" but once not directly near it, are in italic form in the δ matrix; e.g., in A there is connection between *4* and *5* and between *4* and *9*, while in 7ETp there is connection between *5* and *7* and between *5* and *9* positions.

Table 3

A: δ and δ^v Matrices of Pyrimidine Bases

Pyr - bases	δ matrices	δ^v matrices
positions	*1,2,3,4,5,6;*	*1,2,3,4,5,6;*
C	2,3,2,3,2,2 0,1,0,1,0,0	4,4,5,4,3,3 0,6,0,3,0,0
U	2,3,2,3,2,2 0,1,0,1,0,0	4,4,4,4,3,3 0,6,0,6,0,0
T	2,3,2,3,3,2 0,1,0,1,1,0	4,4,4,4,5,3 0,6,0,6,1,0
5MC	2,3,2,3,3,2 0,1,0,1,1,0	4,4,5,4,4,3 0,6,0,3,1,0
positions	*1,2,3,4,5,6;7,8*	*1,2,3,4,5,6;7,8*
5hyMC	2,3,2,3,3,2;2,1 0,1,0,1,*0*,0;*0*,0	4,4,5,4,4,3;2,5 0,6,0,3,*0*,0;*0*,0
OA	2,3,2,3,2,3;3,1 0,1,0,1,0,0;1,0	4,4,4,4,3,4;4,5 0,6,0,6,0,0;6,0

B: δ and δ^v Matrices of Purine Bases

Pur - bases	δ matrices	δ^v matrices
positions	*4,3,2,1,6,5;7,8,9;*	*4,3,2,1,6,5;7,8,9;*
A	3,2,2,2,3,3;2,2,2 0,0,0,0,1,0;0,0,0	4,5,3,5,4,4;5,3,4 0,0,0,0,3,0;0,0,0
G	3,2,3,2,3,3;2,2,2 0,0,1,0,1,0;0,0,0	4,5,4,4,4,4;5,3,4 0,0,3,0,6,0;0,0,0
7MG	3,2,3,2,3,3;3,2,2 0,0,1,0,1,0;1,0,0	4,5,4,4,4,4;5,3,5 0,0,3,0,6,0;1,0,0
IsoG	3,2,3,2,3,3;2,2,2 0,0,1,0,1,0;0,0,0	4,4,4,5,4,4;5,3,4 0,0,6,0,3,0;0,0,0
X	3,2,3,2,3,3;2,2,2 0,0,1,0,1,0;0,0,0	4,4,4,4,4,4;5,3,4 0,0,6,0,6,0;0,0,0
HypoX	3,2,2,2,3,3;2,2,2 0,0,0,0,1,0;0,0,0	4,5,3,4,4,4;5,3,4 0,0,0,0,6,0;0,0,0
UA	3,2,3,2,3,3;2,3,2 0,0,1,0,1,0;0,1,0	4,4,4,4,4,4;4,4,4 0,0,6,0,6,0;0,6,0
Tb	3,3,3,2,3,3;3,2,2 0,1,1,0,1,0;1,0,0	4,5,4,4,4,4;5,3,5 0,1,6,0,6,0;1,0,0
Tp	3,3,3,3,3,3;2,2,2 0,1,1,1,1,0;0,0,0	4,5,4,5,4,4;4,3,5 0,1,6,1,6,0;0,0,0
Cf	3,3,3,3,3,3;3,2,2 0,1,1,1,1,0;1,0,0	4,5,4,5,4,4;5,3,5 0,1,6,1,6,0;1,0,0

C: δ and δ^v Matrices of Purines (Theobromine and Theophylline Derivatives)

Pur - bases	δ matrices	δ^v matrices
position - Tb	*3′,2′,1′,1,2,3,4,9,8,7,5,6;*	*3′,2′,1′,1,2,3,4,9,8,7,5,6;*
1ETb	1,2,3,3,3,3,2,2,3,3,3 0,0,0,1,1,*0*,0,0,1,*0*,1	1,2,5,4,5,4,5,3,5,4,4 0,0,0,6,1,*0*,0,0,1,*0*,6
1PTb	1,2,2,3,3,3,3,2,2,3,3,3 0,0,0,0,1,1,*0*,0,0,1,*0*,1	1,2,2,5,4,5,4,5,3,5,4,4 0,0,0,0,6,1,*0*,0,0,1,*0*,6
1BTb	2,2,2,3,3,3,3,2,2,3,3,3 1,0,0,0,1,1,*0*,0,0,1,*0*,1	2,2,2,5,4,5,4,5,3,5,4,4 1,0,0,0,6,1,*0*,0,0,1,*0*,6
position - Tp	*5,6,1,2,3,4,9,8,7;7′,8′,9′*	*5,6,1,2,3,4,9,8,7;7′,8′,9′*
7ETp	3,3,3,3,3,3,2,2,3; 2, 1 *0*,1,1,1,1,*0*,0,0,0; 0, 0	4,4,5,4,5,4,5,3,5; 2, 1 *0*,6,1,6,1,*0*,0,0,0; 0, 0
7PTp	4,3,3,3,3,3,2,2,3; 2, 2, 1 *0*,1,1,1,1,*0*,0,0,0; 0, 0, 0	4,4,5,4,5,4,5,3,5; 2, 2, 1 *0*,6,1,6,1,*0*,0,0,0; 0, 0, 0
7BTp	3,3,3,3,3,3,2,2,3; 2, 2, 2 *0*,1,1,1,1,*0*,0,0,0; 0, 0, 1	4,4,5,4,5,4,5,3,5; 2, 2, 2 *0*,6,1,6,1,*0*,0,0,0; 0, 0, 1
7ITp	3,3,3,3,3,3,2,2,3; 2, 3, 1 *0*,1,1,1,1,*0*,0,0,0; 0, 1, 0	4,4,5,4,5,4,5,3,5; 2, 3, 1 *0*,6,1,6,1,*0*,0,0,0; 0, 1, 0
7I8MTp	3,3,3,3,3,3,2,3,3; 2, 3, 1 *0*,1,1,1,1,*0*,0,1,0; 0, 1, 0	4,4,5,4,5,4,5,4,5; 2, 3, 1 *0*,6,1,6,1,*0*,0,1,0; 0, 1, 0
7B8MTp	3,3,3,3,3,3,2,3,3; 2, 2, 2 *0*,1,1,1,1,*0*,0,1,0; 0, 0, 1	4,4,5,4,5,4,5,4,5; 2, 2, 2 *0*,6,1,6,1,*0*,0,1,0; 0, 0, 1

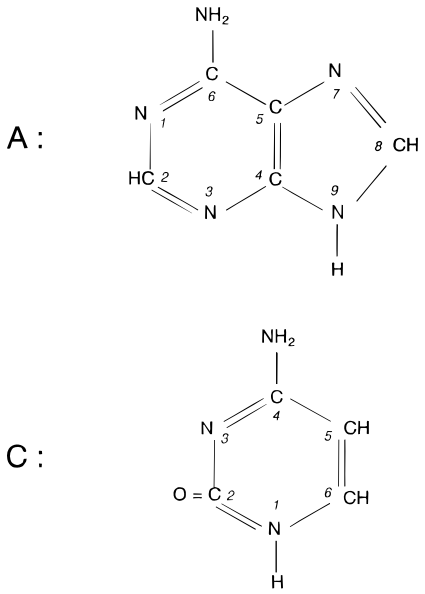

Figure 1. The numbering of a purine (adenine: A) and pyrimidine (cytosine: C).

The calculated connectivity values of the studied purines and pyrimidines have been collected throughout Table 4.

DISCUSSION

Let us first model the five experimental properties of nucleic acid bases, U, T, A, G, and C of Table 1 and compare this modeling with quantum theoretical results[31] and precisely with semiempirical SCF LCAO MO-CI calculations. Results from these calculations, obtained with two different sets of values for the overlap integrals $\beta_{C,N}(1,2)$ and $\beta_{C,O}(1,2)$, are shown in Figure 2. The two sets of overlap integrals show very similar plots with minor differences in their highest portion.

Simulation of these five experimental properties for the five nucleic acid bases ($n = 5$) has been achieved with LCCI made up of 1-, 2-, and 3-χ indices (this lasts for the forward selection method only). Clearly, the use of three indices for $n = 5$ experimental points is exorbitant, but it should not be forgotten that our aim is also (i) to derive the best single orthogonal (the best 1-Ω-LCOCI) index from the corresponding best multi-χ-LCCI and (ii) to compare the simulation power of the two selection techniques for the best LCCI. In fact, for the final simulation only a single index LCCI will be used. Connectivity indices for these five properties are not excessively collinear as $\langle R_{IM}(P(n = 5): \{\chi\})\rangle = 0.956$, while the strongest (s) and weakest (w) correlations are $R_s(D,{}^1\chi) = 0.995$ and $R_w({}^0\chi,\chi_t) = 0.891$. The combinatorial space was chosen by the aid of the forward selection method, that encompasses only a subspace of the full combinatorial space, which is the second methof of choice.

First $\Delta E_{1,exp}$ Singlet Excitation Energy. The best 1-, 2-, and 3-χ combinations extracted by the aid of the forward selection method are

$\{{}^0\chi\}$ $Q = 5.508$ $F = 1.879$ $R = 0.621$ $S = 0.11$

$\{{}^0\chi,\chi_t^v\}$ $Q = 8.389$ $F = 2.179$ $R = 0.823$ $S = 0.010$

$\{D,{}^0\chi,\chi_t^v\}$ $Q = 11.18$ $F = 2.58$ $R = 0.941$ $S = 0.08$

Simulation achieved by the single index seems rather deceiving, while the one achieved by the 3 index is clearly inflated as it simulates five points with three indices. Let us now inspect for a better 2 index combination with the full combinatorial space. This second method gives the following interesting result

$\{{}^1\chi,\chi_t^v\}$ $Q = 11.99$ $F = 4.453$ $R = 0.904$ $S = 0.08$

Improvement achieved by the full combinatorial space seems excellent (the single-χ combination is the same in both spaces) and the contribution, along the two methods, of the total valence connectivity and sum-δ index seems decisive. To avoid inflated LCCI and to improve the simulation power of the single index, orthogonal descriptors will now be introduced. The good score of the first 3-χ-LCCI points to the probable existence of a dominant orthogonal descriptor. In fact, index ${}^2\Omega$ of the orthogonal $\{{}^1\Omega,{}^2\Omega,{}^3\Omega\}$ set of indices derived from the $\{D \equiv {}^1\Omega,\chi_t^v,{}^0\chi\}$ connectivity set for this property (see Table 5) shows an enhanced quality

$\{{}^2\Omega\}$ $Q = 7.400$ $F = 3.391$ $R = 0.728$ $S = 0.10$

Second $\Delta E_{2,exp}$ Singlet Excitation Energy. The 1- and 2-χ best combinations derived by the aid of the forward selection method for this property are

$\{D^v\}$ $Q = 4.370$ $F = 17.85$ $R = 0.925$ $S = 0.21$

$\{D^v,\chi_t^v\}$ $Q = 14.18$ $F = 93.96$ $R = 0.995$ $S = 0.07$

The single-index combination is already satisfactory and can be used for the simulation of this property. The second LCCI is excellent, and it will now be interesting to see if the full combinatorial method can do better. In fact, this latter method of choice finds the following noteworthy combination

$\{{}^0\chi,{}^0\chi^v\}$ $Q = 90.96$ $F = 3866$ $R = 0.99987$ $S = 0.01$

To notice is the totally different kind of combination derived by this method relatively to the preceding one: no total connectivity indices are here relevant for the exceptional simulation.

First Oscillator f_1 Strength of the First Singlet Excitation Energies. While the subspace of the forward selection method selects for this property the following best 1, 2, and 3 index LCCI

$\{\chi_t\}$ $Q = 18.81$ $F = 4.22$ $R = 0.764$ $S = 0.04$

$\{D,\chi_t\}$ $Q = 34.07$ $F = 6.919$ $R = 0.935$ $S = 0.03$

$\{D,\chi_t,\chi_t^v\}$ $Q = 74.70$ $F = 22.17$ $R = 0.9926$ $S = 0.01$

The second and third combinations, where the total connectivity index seem to play an important role, are very fine, and even if the third is clearly inflated it will be used to derive the corresponding orthogonal indices as the good score of both 2- and 3-χ combinations seem to indicate the presence of a dominant orthogonal descriptor for this property. The full combinatorial space prefers the following excellent 2-χ

Table 4. Calculated χ Values of Purine and Pyrimidine Bases (PP Bases)

PP bases	D	D^v	$^0\chi$	$^0\chi^v$	$^1\chi$	$^1\chi^v$	χ_t	χ_t^v
7I8MTp	38	62	13.61036	11.38981	8.34111	5.97071	0.003564	8.51E-05
7B8MTp	38	62	13.44723	11.22667	8.48527	6.11486	0.003086	7.37E-05
7ITp	36	60	12.74012	10.46716	7.93043	5.53989	0.004365	9.82E-05
7BTp	36	60	12.57699	10.30402	8.07459	5.68405	0.00378	8.51E-05
1BTb	36	60	12.57699	10.30402	8.07459	5.68405	0.00378	8.51E-05
7PTp	34	58	11.86988	9.59691	7.57459	5.18405	0.005346	0.00012
1PTb	34	58	11.86988	9.59692	7.57459	5.18405	0.005346	0.00012
7ETp	32	56	11.16277	8.88981	7.07459	4.68405	0.00756	0.00017
1ETb	32	56	11.16277	8.88981	7.07459	4.68405	0.00756	0.00017
Cf	30	54	10.45567	8.1827	6.53658	4.10793	0.01069	0.00024
Tp	28	52	9.58542	7.23549	6.1259	3.71758	0.013095	0.000269
Tb	28	52	9.58542	7.23549	6.10906	3.7135	0.013095	0.000269
UA	26	54	8.71518	5.72474	5.6647	3.11237	0.01604	0.00013
7MG	26	48	8.71518	6.40459	5.68154	3.35084	0.01604	0.00038
OA	22	50	8.43072	5.24931	5.09222	2.66333	0.03928	0.00027
X	24	48	7.84493	5.34106	5.27086	2.92873	0.01964	0.00034
IsoG	24	46	7.84493	5.45738	5.27086	2.96049	0.01964	0.00043
G	24	46	7.84493	5.45738	5.27086	2.96049	0.01964	0.00043
5hyMC	20	40	7.56048	5.16448	4.73638	2.68714	0.04811	0.0012
HypoX	22	42	6.97469	4.95738	4.87701	2.74509	0.02406	0.00085
A	22	40	6.97469	5.07369	4.87701	2.77277	0.02406	0.00108
T	18	36	6.85337	4.89385	4.19838	2.4856	0.06804	0.00301
5MC	18	34	6.85337	5.01016	4.19838	2.51736	0.06804	0.0038
U	16	34	5.98313	3.9712	3.78769	2.06893	0.08333	0.00347
C	16	32	5.98313	4.08751	3.78769	2.1007	0.08333	0.00439

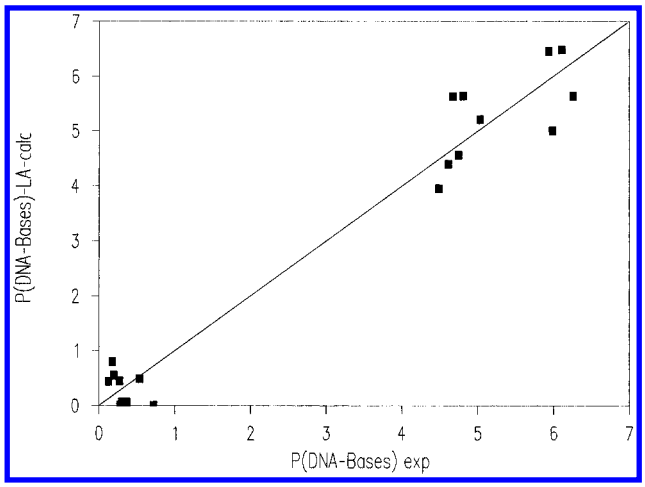

Figure 2. Plot of the calculated (with Ladik−Appel quantum theoretical calculations) versus the experimental ΔE_1, ΔE_2, f_1, and f_2 properties of the DNA/RNA bases.

index LCCI where the total connectivity index and a valence sum-δ index are decisive for the outstanding description

$$\{D^v,\chi_t\} \quad Q = 411.6 \quad F = 1009.6 \quad R = 0.9995 \quad S = 0.002$$

Orthogonalizing the forward selection $\{\chi_t \equiv {}^1\Omega, D, \chi_t^v\}$ set (Table 5) the following dominant orthogonal index is obtained

$$\{{}^3\Omega\}: \quad Q = 22.27 \quad F = 5.912 \quad R = 0.814 \quad S = 0.04$$

while the following 2-Ω LCOCI scores better than the 2-χ forward selection LCCI

$$\{{}^2\Omega, {}^3\Omega\}: \quad Q = 58.94 \quad F = 20.71 \quad R = 0.977 \quad S = 0.02$$

Second Oscillator f_2 Strength of the First Singlet Excitation Energies. The forward selection method selects the following 1, 2, and 3 index combinations with a very poor single descriptor and a good 3-index descriptor, that uncovers, thus, the existence of a dominant orthogonal descriptor

$$\{D^v\} \quad Q = 2.74 \quad F = 1.066 \quad R = 0.512 \quad S = 0.19$$

$$\{D, D^v\} \quad Q = 7.05 \quad F = 3.53 \quad R = 0.883 \quad S = 0.13$$

$$\{D, D^v, \chi_t^v\} \quad Q = 11.98 \quad F = 6.78 \quad R = 0.976 \quad S = 0.08$$

To notice is the fact that, here again, D^v and χ_t^v indices play a central role. The full combinatorial space achieves a nice result with two total connectivity indices

$$\{\chi_t, \chi_t^v\} \quad Q = 22.17 \quad F = 34.85 \quad R = 0.986 \quad S = 0.044$$

The orthogonal indices (Table 5) for this property derived from the $\{D \equiv {}^1\Omega, D^v, \chi_t^v\}$ set yield an interesting ${}^2\Omega$ dominant orthogonal descriptor, that scores much better than D^v and a 2-Ω LCOCI which is better than the corresponding 2-χ forward selection LCCI (orthogonal indices based on a different order of χ indices did not achieve the same good score)

$$\{{}^2\Omega\}: \quad Q = 6.568 \quad F = 6.118 \quad R = 0.819 \quad S = 0.13$$

$$\{{}^2\Omega, {}^3\Omega\}: \quad Q = 8.605 \quad F = 5.150 \quad R = 0.917 \quad S = 0.11$$

Thus, simulation with dominant orthogonal indices is not at all to underestimate and could surely be improved by orthogonalizing the best multi-χ LCCI derived from a full combinatorial space. The forward selection LCCI have here been chosen for orthogonalization to detect how far a good dominant descriptor can be obtained by the aid of this easier combinatorial method. In Figure 3 the given properties estimated by the aid of the best single index LCCI (D^v for $\Delta E_{2,\mathrm{exp}}$) or LCOCI (${}^2\Omega$ for $\Delta E_{1,\mathrm{exp}}$ and f_2 and ${}^3\Omega$ for f_1) are plotted versus the corresponding experimental values. The χ or Ω and **C** vectors for the single index modeling of the

Table 5. Calculated $^2\Omega$ and $^3\Omega$ Indexes from $\{D,\chi_t^v,^0\chi\}$, $\{\chi_t,D,\chi_t^v\}$, and $\{D,D^v,\chi_t^v\}$ Sets for the Modeling of Experimental ΔE_1, f_1, and f_2 of the Nucleotide NB Bases U, G, U, T, and C (E-05 = 10^{-5})

NB	$^2\Omega(\Delta E_1)$	$^3\Omega(\Delta E_1)$	$^2\Omega(f_1)$	$^3\Omega(f_1)$	$^2\Omega(f_2)$	$^3\Omega(f_2)$
A	−0.00011	−0.30243	−0.80102	−0.00163	−1.75756	−0.00027
G	0.000145	0.133373	0.695611	−0.00185	1.272739	0.000213
U	−0.00041	−0.04404	−0.05105	0.000979	1.151523	−0.00036
T	2.5E-05	0.368721	0.207645	0.000594	0.181827	8.1E-06
C	0.000505	−0.15562	−0.05105	0.001899	−0.84848	0.000408

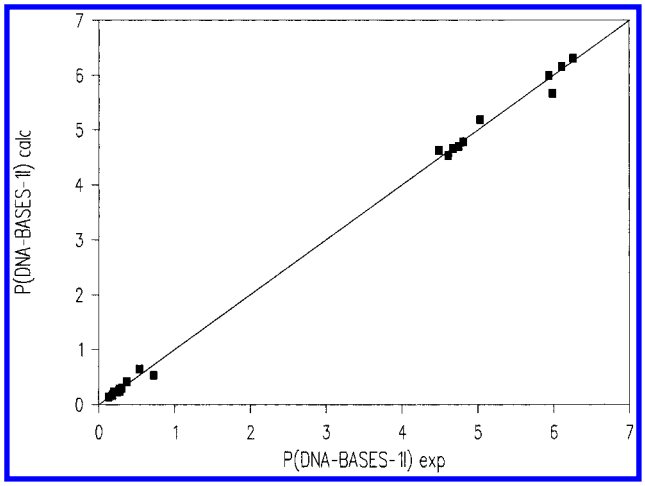

Figure 3. Plot of the calculated (with χ or Ω and **C** vectors 5−8) versus the experimental ΔE_1, ΔE_2, f_1, and f_2 properties of the DNA/RNA bases.

four properties, that result in a better modeling than the quantum theoretical modeling are

$$\Delta E_{1,\text{exp}}: \quad \Omega = (^2\Omega,\Omega^0) \quad \mathbf{C} = (-268.676,\ 4.67433) \quad (5)$$

$$\Delta E_{2,\text{exp}}: \quad \chi = (D^v,\chi^0) \quad \mathbf{C} = (-0.08058,\ 8.89597) \quad (6)$$

$$f_{1,\text{exp}}: \quad \Omega = (^3\Omega,\Omega^0) \quad \mathbf{C} = (-26.8465,\ 0.1940) \quad (7)$$

$$f_{2,\text{exp}}: \quad \Omega = (^2\Omega,\Omega^0) \quad \mathbf{C} = (-0.11839,\ 0.4400) \quad (8)$$

It is to notice that we are comparing the MC modeling with 1966 quantum theoretical results,[31] such a comparison is not so anomalous if we think that MC theory is today nearly 20 years old while in 1966 quantum theory was more than 30 years old.

Molar Absorption $\epsilon_{260,\text{exp}}$ Coefficients. These are in fact the molar absorption coefficients at 260 nm and pH = 7 of nucleotides UMP, TMP, AMP, GMP, and CMP (the spectra of the corresponding ribo- and deoxynucleotides as well as the nucleosides are essentially identical[34]) but as the only noncommon part of these nucleotides are U, T, A, G, and C bases, the simulation of this property is done by the aid of the $\{\chi\}$ values of U, T, A, G, and C only.

The forward selection method chooses the following 1, 2, and 3 index combinations

$$\{\chi_t\} \quad Q = 0.423 \quad F = 6.47 \quad R = 0.827 \quad S = 1.86$$

$$\{D,\chi_t\} \quad Q = 0.800 \quad F = 11.57 \quad R = 0.959 \quad S = 1.20$$

$$\{D,\chi_t,\chi_t^v\} \quad Q = 7.599 \quad F = 696.9 \quad R = 0.9998 \quad S = 0.13$$

The second and last combination could have a better counterpart in the full combinatorial method, furthermore, their good score relative to the interesting single index combination could be a hint for the presence of a better dominant orthogonal descriptor. The full combinatorial space constructed with two connectivity indices displays the presence of the following first rate 2-χ index combination

$$\{D^v,\chi_t^v\} \quad Q = 2.027 \quad F = 74.4 \quad R = 0.9933 \quad S = 0.49$$

where the valence D and total connectivity indices play the decisive role.

These results about these five properties of DNA/RNA molecules seem very promising and they underline the plausibility of "searching and finding" a special single index LCCI for every experimental property of these molecules. It should be noticed here that every calculated value is positive.

Modeling ⟨p*K*⟩, MW, and Sol of the Expanded Set of Purine and Pyrimidine Bases. The set of connectivity indices that simulate the average ⟨p*K*⟩ value of $n = 13$ purines and pyrimidines show no striking collinearity with $\langle R_{IM}(pK\text{: } \{\chi\}\rangle = 0.904$, while the strongest and weakest interrelation are $R_s(D,^1\chi) = 0.993$ and $R_w(^0\chi^v,\chi_t^v) = 0.779$. The most remarkable LCCI, starting with the single index simulation, derived by the aid of the forward selection method are rather poor

$$\{\chi_t\} \quad Q = 0.184 \quad F = 2.37 \quad R = 0.421 \quad S = 2.29$$

$$\{^1\chi^v,\chi_t\} \quad Q = 0.394 \quad F = 5.43 \quad R = 0.722 \quad S = 1.83$$

$$\{^0\chi,^0\chi^v,^1\chi,^1\chi^v,\chi_t\} \quad Q = 0.488 \quad F = 3.33 \quad R = 0.839 \quad S = 1.72$$

Comparison among the results of the 1-χ LCCI and the multi-χ LCCI suggests the possibility that a single orthogonal index could score better. The full combinatorial space method here is not able to find any new interesting combination. If the outliner U is excluded from the modeling then the following LCCI can be detected with both the forward and full selection method.

$$\{^0\chi^v,^1\chi^v,\chi_t,\chi_t^v\} \quad Q = 0.62 \quad F = 6.38 \quad R = 0.839 \quad S = 1.61$$

Even if this combination shows a somewhat better score than the preceding LCCI it is not an optimal descriptor, and furthermore, like the preceding one, it is too inflated. This fact should not surprise as p*K* values are highly dependent on the number of acid or basic groups and only weakly dependent on the whole molecular structure as it has been demonstrated with amino acids.[17,28] The search, once U has been excluded from the modeling, for a new kind of connectivity indices grounded on the number of acid and basic groups ends up with the finding of these two new fragment R_{1f} and R_{2f} reciprocal connectivity indices

$$R_{1f} = \Delta n_1/D^v \quad (9)$$

$$R_{2f} = \Delta n_2/D^v \quad (10)$$

where $[\Delta n_1 = n_{CO} + n_{OH} - n_{COOH}]$ is the algebraic sum of

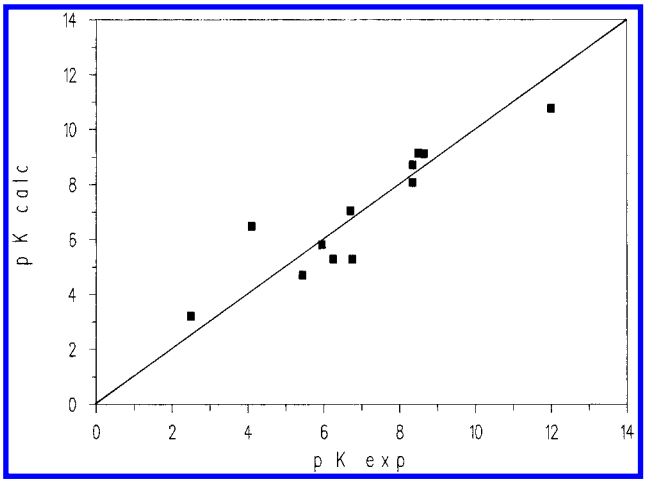

Figure 4. Plot of the calculated (with χ and **C** vectors 11) versus the experimental $\langle pK \rangle$ of $n = 12$ purine and pyrimidine bases.

the number of CO, OH, and COOH groups in each molecule (3 in UA, 2 in T, 5 in hyMC and X, 0 in A, and 1 in the remnants) and $[\Delta n_2 = n_{CO} + n_{OH} + n_{NH} - n_{COOH}]$ (7 in UA, 5 in X, 4 in T, 2 in 5MC and 7MG, 1 in A, and 3 in the remnants). These two indices offer an interesting solution for an improved modeling of this property with less indices (3 instead of 4 or 5) with the forward selection method

$\{R_{1f}\}$ $Q = 0.408$ $F = 11.09$ $R = 0.725$ $S = 1.78$

$\{R_{1f},\chi_t\}$ $Q = 0.587$ $F = 11.44$ $R = 0.847$ $S = 1.44$

$\{R_{1f},{}^1\chi^v,\chi_t\}$ $Q = 0.728$ $F = 11.73$ $R = 0.903$ $S = 1.24$

$\{R_{2f},{}^1\chi^v,\chi_t\}$ $Q = 0.742$ $F = 12.20$ $R = 0.906$ $S = 1.22$

Improvement in Q and F values with R_{1f} fragment index relative to the preceding modeling of $\langle pK \rangle$ is more than interesting: single- and 2-index LCCI double their quality nearly and the 2-index LCCI is as powerful as the 4-index LCCI. The combination with R_{2f} index shows a good score only with ${}^1\chi^v$ and χ_t indices and in this case they show a somewhat better quality than the R_{1f} combination.

A better search and description with less indices (less than 3) is here not possible as (i) two $\langle pK \rangle$ values are not exact (see Table 2, T and A), (ii) four bases show a third pK_c value <13,[33] and (iii) tautomeric equilibria are not considered. The simulating vectors used to draw Figure 4 are

$$\chi = ({}^1\chi^v,\chi_t,R_{1f},\chi^0)$$
$$\mathbf{C} = (5.80354,\ 121.778,\ 69.7117,\ -15.8002) \quad (11)$$

Figure 4 confirms the promising simulation of $\langle pK \rangle$ with the mixed set of χ and R indices. This set of indices are not only poorly interrelated but also the R_{1f} index displays (nearly) the property of an orthogonal index, in fact $R({}^1\chi^v,\chi_t) = 0.901$, $R({}^1\chi^v,R_{1f}) = 0.115$, and $R(\chi_t,R_{1f}) = 0.280$.

While orthogonal indices derived from the given set with $R_{1,2f}$ indices are no better descriptors, the orthogonal $\{{}^1\Omega,{}^2\Omega,{}^3\Omega,{}^4\Omega\}$ indices (see Table 6) derived from the $\{{}^1\chi^v \equiv {}^1\Omega,{}^0\chi,{}^1\chi,{}^0\chi^v\}$ connectivity set for $n = 13$ compounds offer an interesting alternative for this $\langle pK \rangle$ simulation. The set used for the orthogonalization has been detected by the aid of the full combinatorial technique and shows the following statistical score:

$\{{}^0\chi,{}^0\chi^v,{}^1\chi,{}^1\chi^v\}$:
$Q = 0.522$ $F = 4.76$ $R = 0.839$ $S = 1.61$

The single LCOCI is, in fact, consistently better than the corresponding LCCI for $n = 13$

$\{{}^3\Omega\}$: $Q = 0.375$ $F = 9.872$ $R = 0.688$ $S = 1.83$

The simulation of the molecular MW weights, which is in itself rather trivial, will be used here to derive some information about the simulation of hydrogen-rich and hydrogen-poor molecules. Molecular connectivity, that is based on hydrogen-suppressed graph and whose descriptors do not include information on hydrogens, normally underestimates this problem, that can easily be detected with the modeling of molecular weights of different classes of compounds, that have different hydrogen density, like, f.e., bases and amino acids. The mean correlation coefficient of the interrelation matrix and the strongest and weakest correlation for the $n = 25$ bases is

$\langle R_{IM}(\text{MW: } \{\chi\}) \rangle = 0.883$ $R_s(D,{}^1\chi) = 0.997$
$R_w({}^0\chi^v,\chi_t^v) = 0.621$

The interrelation is here rather low while the χ_t^v index is rather poorly interrelated with the other indices. The best LCCI for $n = 25$, starting with the single index combination, with the forward selection method are the following

$\{{}^0\chi\}$ $Q = 0.379$ $F = 7002$ $R = 0.998$ $S = 2.63$

$\{{}^0\chi,{}^1\chi\}$ $Q = 1.010$ $F = 24885$ $R = 0.99978$ $S = 0.99$

$\{{}^0\chi,{}^0\chi^v,{}^1\chi\}$ $Q = 2.316$ $F = 87190$ $R = 0.99996$ $S = 0.43$

Comparing these results with the corresponding results on the molecular weights of amino acids[18] whose best 3-χ indices LCCI for $n = 21$ amino acids is

$\{D^v,{}^0\chi,{}^1\chi^v\}$ with
$Q = 0.259$ $F = 401$ $R = 0.993$ $S = 3.84$

we notice that (i) this combination is also a very fine descriptor for the $n = 25$ bases with, $Q = 1.809$, $F = 53168$, $R = 0.99993$, and $S = 0.55$, while (ii) the $\{{}^0\chi,{}^0\chi^v,{}^1\chi\}$ combination for the same $n = 21$ amino acids shows $Q = 0.918$, $F = 233$, $R = 0.988$, and $S = 5.00$. In both cases total χ_t and χ_t^v indices play no role. Thus, while connectivity indices are exceptional descriptors of the MW of nucleotide bases, they are not as good descriptors of the same property for amino acids. The three index LCCI for the $n = 25$ bases shows Q and F values that are 9 and 200 times better (respectively) than the three index LCCI for the 21 amino acids. Furthermore, both classes of molecules are described by the same single index $\{{}^0\chi\}$, but the resulting Q, F, and R values for the bases are 2.5, 18, and 1.02 times better than the corresponding Q, F, and R values for the amino acids. Now, the different hydrogen density along the two classes of compounds could explain the difference in the simulation power for this property. In fact, the calculated ratio, f, of hydrogen atoms to C, O, and N heteroatoms in purine/pyrimidine bases and amino acids is $f(PP) = 0.85$ and $f(AA) = 1.06$, respectively. Thus, hydrogen-poor compounds seem

to be more satisfactorily described by the well-known molecular connectivity χ indices, that are grounded on hydrogen-suppressed graphs. The simulation of molecular weights of different classes of molecules should not be underestimated as it can then be a good test for an eventual definition of connectivity descriptors of hydrogen-non-suppressed graphs.

The last property considered in this study is the solubility of a first set of $n = 12$ bases (from Tb to C, see Table 2) and of a second $n = 23$ set of bases (including Tb and Tp derivatives). Tp has been excluded from the first set as it is in this case a strong outliner. Use of two different subsets of the same class of compounds will allow one not only to check in a more detailed way the descriptive power of the LCCI method but also to detect changes in χ-index combinations introduced by a flexible molecular data set. It is interesting to notice that even here the value of $\langle R_{IM} \rangle$ lowers with growing n and that χ_t^v and χ_t behave like orthogonal indices nearly

$\langle R_{IM}$(Sol(12):

$$\{\chi\})\rangle = 0.894 \quad R_s(D,{}^1\chi) = 0.997 \quad \text{and} \quad R_w({}^1\chi^v,\chi_t^v) = 0.70$$

$\langle R_{IM}$(Sol(23):

$$\{\chi\})\rangle = 0.711 \quad R_s(D,{}^1\chi) = 0.9998 \quad \text{and} \quad R_w(D^v,\chi_t) = 0.31$$

The subspace of the forward selection method, starting with the single-index extracts the following best combinations for $n = 12$

$\{\chi_t^v\}$ $Q = 10.59$ $F = 71.94$ $R = 0.937$ $S = 0.09$

$\{D^v,\chi_t^v\}$ $Q = 11.20$ $F = 40.21$ $R = 0.948$ $S = 0.08$

$\{D,D^v,{}^0\chi,{}^1\chi,\chi_t^v\}$ $Q = 13.55$ $F = 23.56$ $R = 0.975$ $S = 0.07$

The single index has the characteristics of a dominant descriptor. The description is not bad, especially the F-best LCCI, and it could surely be improved if the temperature of the solubility set was more homogeneous (see Table 2 in parentheses).

The full combinatorial space, instead, shows the following interesting Q-combination (1-χ- and 2-χ-index combinations are the same for both combinatorial spaces)

$$\{D,{}^0\chi,{}^0\chi^v,{}^1\chi,\chi_t^v\}$$
$$Q = 14.05 \quad F = 25.33 \quad R = 0.977 \quad S = 0.07$$

The last two LCCI obtained with both methods are clearly too inflated (five indices for 12 points), but they will serve not only for comparing the two methods but also, later on, for an unexpectedly good *expansion test*.

Simulation of the full $n = 23$ (Tp and and Tb derivatives inclusive) set of purines and pyrimidines, that shows a greater temperature homogeneity, is only possible by the aid of supra-χ-indices for 7PTp ($a = 4$), 1Etb ($a = 2$), Cf ($a = 2$), and 7ITp ($a = 1.5$, where a is an association parameter) that have already been successfully used for these molecules in a preceding work.[20] But let us first simulate the $n = 23$ purine and pyrimidine bases by the aid of the same LCCI used for the restricted set $n = 12$ purines and pyrimidines. This kind of *expansion test* provides an interesting simulation with a very bad single-index LCCI followed by rather fine multi-χ LCCI that uncover, thus, the existence of a dominant orthogonal descriptor.

Table 6. Calculated $^2\Omega$, $^3\Omega$, and $^4\Omega$ Indexes from $\{{}^1\chi^v,{}^0\chi,{}^1\chi,{}^0\chi^v\}$ Set for the Modeling of $\langle pK \rangle$ of $n = 13$ Purine and Pyrimidine PP Bases

PP	$^2\Omega$	$^3\Omega$	$^4\Omega$
T	−0.05741	−0.21288	−0.0399
5hyMC	0.20311	−0.08488	−0.03463
5MC	−0.12779	−0.24883	0.003179
UA	0.415547	0.065152	−0.08346
C	−0.07475	0.0359	0.027554
IsoG	−0.11815	0.067066	0.001219
7MG	−0.11288	−0.1879	0.062054
G	−0.11815	0.067066	0.001219
OA	1.126111	0.074053	0.068241
HypoX	−0.51109	0.140856	−0.00029
U	−0.00435	0.071857	−0.0155
X	−0.04778	0.10301	−0.04187
A	−0.57243	0.109529	0.052197

$\{\chi_t^v\}$ $Q = 0.022$ $F = 0.25$ $R = 0.109$ $S = 4.86$

$\{D^v,\chi_t^v\}$ $Q = 0.636$ $F = 101.6$ $R = 0.954$ $S = 1.50$

$\{D,{}^0\chi,{}^0\chi^v,{}^1\chi,\chi_t^v\}$ $Q = 0.850$ $F = 72.6$ $R = 0.977$ $S = 1.15$

The forward selection method with supramolecular connectivity indices for the $n = 23$ bases shows a better single-index LCCI, that has the property of a dominant descriptor, the same 2-χ-index LCCI of the expansion test and a rather fine 4-χ-index LCCI

$\{D^v\}$ $Q = 0.514$ $F = 132.8$ $R = 0.929$ $S = 1.81$

$\{D^v,\chi_t^v\}$ $Q = 0.636$ $F = 101.6$ $R = 0.954$ $S = 1.50$

$\{D^v,{}^0\chi^v,{}^1\chi^v,\chi_t^v\}$ $Q = 0.896$ $F = 100.8$ $R = 0.978$ $S = 1.09$

Introduction of the following $\{\chi^2\}$ combination results in an astonishing improvement of the modeling

$$\{(D^v)^2,({}^0\chi^v)^2,({}^1\chi^v)^2,(\chi_t^v)^2\}$$
$$Q = 2.781 \quad F = 972.2 \quad R = 0.9977 \quad S = 0.36$$

The best combinations derived by the aid of the full combinatorial space show no interesting improvement over the best combinations of the preceding subspace

$\{D,{}^1\chi\}$ $Q = 0.659$ $F = 109.2$ $R = 0.957$ $S = 1.45$

$\{D,{}^0\chi^v,{}^1\chi\}$ $Q = 0.827$ $F = 114.6$ $R = 0.973$ $S = 1.18$

$\{{}^0\chi^v,{}^1\chi,{}^1\chi^v,\chi_t\}$ $Q = 0.919$ $F = 106.2$ $R = 0.979$ $S = 1.07$

Even here, the linear combination of squared connectivity indices LCSCI achieves an impressive improvement of the modeling

$\{({}^0\chi^v)^2,({}^1\chi)^2,({}^1\chi^v)^2,(\chi_t)^2\}$:
$$Q = 2.79 \quad F = 978.4 \quad R = 0.9977 \quad S = 0.36$$

Thus, the χ^2 and **C** simulating vectors are

$$\chi^2 = (({}^0\chi^v)^2,({}^1\chi)^2,({}^1\chi^v)^2,(\chi_t)^2,\chi^0)$$

$$\mathbf{C} = (-0.03418,\ 0.04046,\ 0.08613,\ 163.579,\ -0.87390) \tag{12}$$

Because two calculated solubility values are negative, the

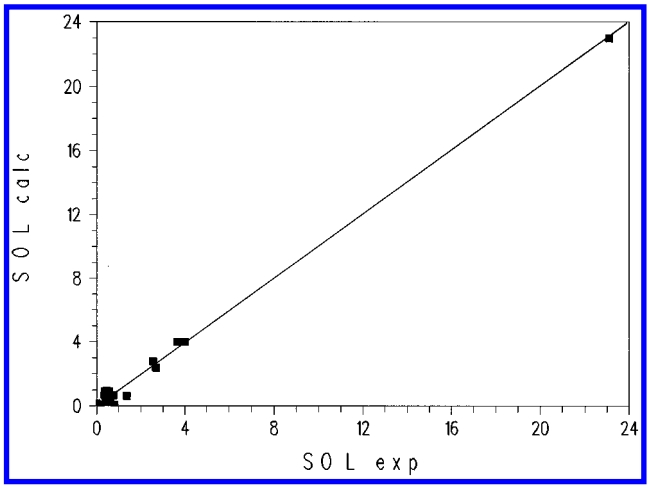

Figure 5. Plot of the calculated (with χ and **C** vectors 12 and modeling equation 4) versus the experimental solubility of $n = 23$ purine and pyrimidine bases.

absolute value equation is here used for the simulation without loss of descriptive power, in fact

$$\mathbf{Sol} = \mathbf{C}\cdot\chi^2: \quad Q = 3.01 \quad F = 4566$$

$$\mathbf{Sol} = |\mathbf{C}\cdot\chi^2|: \quad Q = 3.02 \quad F = 4596$$

Figure 5 has been, in fact, obtained by the aid of vector 12 and of this last equation and it shows a fine quality indeed.

CONCLUSIONS

From the considered simulations of the different experimental properties of purine and pyrimidine bases we notice the importance of the D and D^v and of the total χ_t and χ_t^v connectivity indices for the simulation of the DNA−RNA properties while for the other properties for $n > 5$ purines and pyrimidines, ${}^0\chi$, ${}^0\chi^v$, ${}^1\chi$, and ${}^1\chi^v$ connectivity indices are more relevant for a good simulation. Noticeable is the forward selection description of the solubility space by the aid of valence connectivity indices only. While orthogonal connectivity indices are the dominant descriptors for ΔE_1, f_1, f_2, and $\langle pK \rangle$ ($n = 13$), the fragment reciprocal, R_{1f} and R_{2f}, indices, that take into account the number of acid and basic groups (COOH, CO, OH, and NH), seem to be interesting descriptors for this property. Nonetheless, the relatively poor simulation of this last property, when compared with the other simulations, could be ascribed to inaccurate pK_i values and to the exclusion from the modeling of tautomeric effects. Simulation of the molecular weights of purine plus pyrimidine bases and amino acids offers the possibility to detect effects that are due to the different hydrogen density of these two different classes of compounds. This simulation offers, thus, some interesting hints about the possibility to describe hydrogen-rich and hydrogen-poor molecules with χ indices that normally do not encode information about the hydrogen atoms. All along the given simulations it can be seen that the forward selection method normally offers a good alternative to the full combinatorial method for ΔE_2, $\langle pK \rangle$, and for Sol, while the estimation of MW with the forward selection method is enough for an exceptional modeling. An interesting characteristic of the forward selection method consists of the possibility to restrict the dimension of the full combinatorial space to be explored. It is, then, always worth starting the search for the best LCCI with this easier method and then employ the full combinatorial space that should model the same property with less or at least the same number of indices, excluding, thus, higher multi-index combinations. The importance of the fact that the forward selection method is a positive method for deriving good combinations can be hardly underlined at the light of the combinatorial problem caused, f.e., by 10 or 20 different indices. In this case the full selection method generates 1032 and 1 045 875 combinations, respectively, while the forward selection method generates just 55 and 210 combinations, respectively. The *expansion test* method applied to Sol reveals how far simulations on restricted sets of compounds can be extrapolated to larger sets and offers, thus, the possibility to short-circuit, f.e., the problem of redefining each time the orthogonal connectivity indices when one or more compounds are added to the set of n compounds.

The two simulation of the two sets of solubility values together with preceding results[20] maintain that linear combinations of squared connectivity indices (LCSCI) are the best choice for this kind of property with this class of molecules. Interesting is the fact that the solubility of amino acids is instead simulated by linear combinations of reciprocal connectivity indices (LCRCI).[28,29]

The good score of many simulations is certainly due to the rather low collinearity (defined as $R < 0.98$[32]) of χ_t and χ_t^v indices relative to the other χ indices of the set as can be seen from the following correlation values

ΔE_1, ΔE_2, f_1, f_2, ϵ_{260}:	$0.891 < R(\chi/\chi_t,\chi_t^v) < 0.989$
$\langle pK \rangle$ ($n = 13$):	$0.779 < R(\chi/\chi_t,\chi_t^v) < 0.968$
MW ($n = 25$):	$0.621 < R(\chi/\chi_t,\chi_t^v) < 0.902$
Sol ($n = 12$):	$0.700 < R(\chi/\chi_t,\chi_t^v) < 0.952$
Sol ($n = 23$):	$0.310 < R(\chi/\chi_t,\chi_t^v) < 0.352$

APPENDIX. ENCODING BASE TRIPLETS WITH THE AID OF AN EFFECTIVE VALENCE OF THE MIDDLE BASE

In the following the original notion of δ valence will be used to derive the different families and subfamilies of the genetic code. It is well-known that amino acids can be coded by a single base triplet, a subfamily, or a family of base triplets[34] formed by U, C, A, and G and that the coding relationship between 64 triplets and 20 natural amino acids are normally summarized by grouping codons with similar first (5′-OH terminal base) and middle base into a grid where the third terminal base (3′-OH terminal) changes from triplet to triplet. The individual boxes (16 boxes generated by the intersection of the first and second base) of this grid are code word families (eight of them are divided into subfamilies), that differ only in their third base, that is not always decisive for the encoding of an amino acid. Now, topologically speaking, the bases of these triplets can be identified by their neighborhood relations, that is, each base of these codons can be identified by a Δ supravalence number or supraconnectivity degree. Clearly, while for the first B_1 and last B_3 base, $\Delta = 1$ or ${}^1B$ (B symbolizes a δ^v matrix that is unique

Table 7. Effective Valence of the Middle Base and Its Family Partitioning Power

$^{\Delta eff}B_m$	*when* $\Delta_{eff} = 2$; Fml ptg: (*type of* B_3)
^{1}C	never; no partitioning: (*meaningless)*
^{1,2}U	$B_1 = \{$A: (*G/A, pyr*); U: (*pur/pyr*)$\}$
^{1,2}G	$B_1 = \{$A: (*pur/pyr*); U: (*A/G/pyr*)$\}$
^{2}A	always; double partitioning: (*pur/pyr*)

for each of the four genetic bases), for the middle B_m base, $\Delta = 2$. For a specific encoding of the genetic code it is worthy to introduce for the middle base an effective Δ, that can assume the values 1 and 2, that is, $\Delta_{eff} = 1$ or 2. This Δ_{eff} number takes into account the effectiveness of the third B_3 base in partitioning a family into subfamilies and thus the capability of one or more code words in characterizing an amino acid. Thus, while a family partitioning follows only when $\Delta_{eff} = 2$, a $\Delta_{eff} = 1$ means that the encoding of an amino acid is not determined by a specific B_3 but only by B_1 and B_m (or B_2). In this way, families and subfamilies of the genetic code can be encoded by $^{\Delta eff}B_m$. In Table 7 the Δ_{eff} valence value of $B_m = \{$A, G, U, C$\}$, the conditions for $\Delta_{eff} = 2$ (for U and G controlled by B_1) and the corresponding family partitioning (Fml ptg) due to the type of the third base (*type of* B_3) have been collected. In this table pur $= \{$A, G$\}$ and pyr $= \{$U, C$\}$.

This table shows that for $B_m = {}^{2}$A, two two-membered subfamilies can be generated when B_3 equals either a purine or a pyrimidine, while for $B_m = {}^{1}$C no family partitioning follows, with the result that, each of the four four-membered (due to four B_1 and four B_3) families encode an amino acid. The partitioning due to B_3 when $B_m = {}^{2}$G and ^{2}U depends on B_1 only if $B_1 =$ A or U. For example, for $B_1 =$ A and $B_m = {}^{2}$U, the four-membered family is split by B_3 into (i) a subfamily AUG, that encodes an amino acid (Met), and (ii) a three-member subfamily AUA, AUU, and AUC, that encodes another amino acid (Ile). For $B_1 =$ U (and $B_m = {}^{2}$U), instead, we have two two-membered pur/pyr families due to B_3 base encoding two different amino acids. With $B_1 =$ C and G and $B_m = {}^{1}$U two four-membered families are obtained, each family encoding an amino acid. The same reasoning follows for ^{1,2}G, with the only difference, that, now, the partitioning (type of B_3) due to A and U are nearly inverted (in fact, A do not mix with pyr).

ACKNOWLEDGMENT

I would like to thank the audience of PACIFICHEM '95 (Honolulu), Professor Milan Randić of the Drake University, and both referees for their helpful suggestions. A referee suggested for the index D (see eq 1) the following relation

$$D = 2(n + \mu - 1)$$

where n is the number of vertices of a molecular graph G (or non-hydrogen atoms in a molecule) and μ is the cyclomatic number of a polycyclic graph G[35] that is equal to the minimum number of edges necessary to be erased from G in order to transform it into the related acyclic subgraph (where M is the number of edges in G, and $N = n$).[36,8]

$$\mu = M - N + 1$$

REFERENCES AND NOTES

(1) Randić, M. *J. Am. Chem. Soc*. **1975**, *97*, 6609.
(2) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randić, M. *J. Pharm. Sci*. **1975**, *64*, 1971.
(3) Kier, L. B.; Hall, L. H. *J. Pharm. Sci*. **1976**, *65*, 1806.
(4) Kier, L. B.; Hall, L. H. *J. Pharm. Sci*. **1981**, *70*, 583.
(5) Kier, L. B.; Hall, L. H. In *Molecular Connectivity in Structure−Activity Analysis*; Wiley: New York, 1986.
(6) Kier, L. B.; Hall, L. H. In *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York 1976.
(7) Kier, L. B.; Hall, L. H. In *Advances in Drug Research*; Testa, B., Ed.; Academic: New York, 1992.
(8) Trinajstić, N. In *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.
(9) Basak, S. C.; Magnuson, V. R.; Niemi, G. C.; Regal, R. R. *Discrete Appl. Math*. **1988**, *19*, 17.
(10) Needham, D. E.; Wei, I. C.; Seybold, P. G. *J. Am. Chem. Soc*. **1988**, *110*, 4186.
(11) Kier, L. B.; Hall, L. H. *Pharm. Res*. **1990**, *7*, 801.
(12) Basak, S. C.; Niemi, G. V.; Veith, G. D. *J. Math. Chem*. **1991**, *7*, 243.
(13) Balaban, A. T.; Kier, L. B.; Joshi, N. *MATCH* **1992**, *28*, 13.
(14) Hall, L. H.; Mohney, B. K.; Kier, L. H. *Quant. Struct.-Act. Relat*. **1993**, *12*, 44.
(15) Nikolić, S.; Medić-Sarić, M.; Rendić, S.; Trinajstić, N. *Drug Metab. Rev*. **1994**, *26*, 717.
(16) Lucić, B.; Nikolić, S.; Trinajstić, N.; Jurić, A.; Mihalić, Z. *Croat. Chem. Acta* **1995**, *68*, 417.
(17) Pogliani, L. *J. Phys. Chem*. **1993**, *97*, 6731.
(18) Pogliani, L. *J. Phys. Chem*. **1994**, *98*, 1494.
(19) Pogliani, L. *Curr. Top. Pept. Prot. Res*. **1994**, *1*, 119.
(20) Pogliani, L. *J. Phys. Chem*. **1995**, *99*, 925.
(21) Pogliani, L. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 801.
(22) Pogliani, L. Proceedings of MATH/CHEM/COMP '96, *Croat. Chem. Acta*, to be published.
(23) Randić, M. *New J. Chem*. **1991**, *15*, 517.
(24) Randić, M. *J. Chem. Inf. Comput. Sci*. **1991**, *31*, 311.
(25) Randić, M. *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45.
(26) Randić, M. *Int. J. Quantum Chem.: Quantum Biol. Symp*. **1994**, *21*, 215.
(27) Lucić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D.; Jurić, A. *Croat. Chim. Acta* **1995**, *68*, 435.
(28) Pogliani, L. *Amino Acids* **1995**, *9*, 217.
(29) Pogliani, L. *Croat. Chem. Acta* **1996**, *69*, 95.
(30) Pogliani, L. *J. Pharm. Sci*. **1992**, *81*, 334.
(31) Ladik, J.; Appel, K. *Theoret. Chim. Acta* **1966**, *4*, 132.
(32) Mihalić, Z.; Nikolić, S.; Trinajstić, N. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 28.
(33) *CRC Handbook of Chemistry and Physics*; David, R. L., Ed.-in-Chief; CRC Press: Boca Raton, FL, 1991−1992; p 7-3.
(34) Lehninger, A. In *Biochemistry*; Worth: New York, 1977; pp 314−315.
(35) Balaban, A. T. *Chem. Phys. Lett*. **1982**, *89*, 399.
(36) Wilson, R. J. In *Introduction to Graph Theory*; Oliver & Boyd: Edinburgh, 1972.

CI960020D