Amino Acids (1997) 13: 237–255

Amino
Acids
© Springer-Verlag 1997
Printed in Austria

Modeling properties of biochemical compounds with connectivity terms

L. Pogliani

Dipartimento di Chimica, Universita' della Calabria, Rende (CS), Italy

Accepted May 5, 1997

Yardsticks are good for measuring if you have yards to measure (Benét SV, from Barr, 1979)

Summary. The descriptive and utility power of linear combinations of connectivity terms (LCCT) derived by a trial-and-error procedure from a medium-sized set of 8 connectivity indices: $\{\chi\} = \{D, D^v, {}^0\chi, {}^0\chi^v, {}^1\chi, {}^1\chi^v, \chi_t, \chi_t^v\}$ or from a subset of it has been tested on properties of heterogeneous classes of biochemical compounds centered on the homogeneous class of natural L-amino acids. To choose the appropriate combination of indices the forward selection and the complete combinatorial technique have been used, whenever more than a single term was necessary for the description. The forward selection technique searches only a subspace of the complete combinatorial space, but nevertheless has many advantages among which to be a good tool for an elementary and direct test for newly defined indices. The modeling has been followed centering the attention not only on the predictive power of the proposed linear equations but also on their utility. The modeling of the solubility of the entire heterogeneous class of n = 43 amino acids, purines and pyrimidines could satisfactorily be achieved with a set of supraconnectivity terms based on the χ_t^v index mainly. The unfrozen water content of a mixed class of inorganic salts and natural amino acids has satisfactorily been modeled with two connectivity terms and the modeling shows a remarkable utility. The utility of the given LCCT can nevertheless be enhanced, especially when the modeling requires 2 or more terms, with the introduction of the corresponding orthogonal indices, as can be seen for S(AA + PP) and UWC.

Further, the δ cardinal number is used as starting point for the definition of a supravalence index Δ to be used for a topological codification of the genetic code and the amino acids in proteins. In fact, the notion of supravalence can be extended to the triplet code words to generate the different families and subfamilies of the genetic code and to visualize the connections of amino acids in proteins. Three properties of the DNA-RNA bases (U, T, A, G and C), the singlet excitation energies ΔE_1 and ΔE_2, and the molar absorption coefficient ε_{260} have been simulated with a single connectivity term chosen from the same medium-sized set of 8 molecular connectivity indices.

Keywords: Amino acids – Modeling – Connectivity terms – Properties – Purines and pyrimidines

Background

Many structure-property studies use graph theoretical indices that are based on the topological properties of a molecule viewed as a graph. The main goal of topology is always towards the general, that is, towards relations and theorems that apply to any space, without reference to measurements or any kind of metrics. Thus, atoms embedded in a graph will no longer be Euclidean points but any unspecified thing to which we can apply these relationships meaningfully. This kind of generalization is natural to mathematics; six pairs are a dozen, whether loaves or atoms or days. A graph, in a topological context becomes, thus, the abstracted essence of the properties of traversing and joining, and conversely a molecule is a concrete manifestation of an abstracted graph where the Euclidean metric together with the notions of congruence (two or more geometric figures are congruent if they differ only in location in space) and similarity (two geometric figures are similar if one is an enlargement of the other, that is, two similar polygons have corresponding angles equal and proportional corresponding sides) go by the board. A graph G can be defined as a set of V vertices with a set of E edges that connect these vertices, that is, G = (V,E). Thus a graph is determined by the set of vertices and by the set of edges joining the vertices and not by the particular appearance of the configuration. A chemical graph is a graph where atoms and bonds are represented by vertices and edges respectively, clearly double bonds or lone-pair electrons cannot be fitted by a graph, for this reason sometimes pseudographs are used to represent organic molecules. A pseudograph G = (V,E) is a more general form of graph which consists of vertices and edges between these vertices, and that allows multiple edges between pairs of vertices and loops, which are edges from a vertex to itself (Rosen, 1995). Every graph is also a pseudograph. However, not all pseudographs are simple graphs, since in a pseudograph two or more edges may connect the same pair of vertices and multiple edges may be associated to the same pair of vertices. A central characteristic of a graph or pseudograph is the degree of a vertex, which can be defined as the number of edges incident with it, except that a loop at a vertex contributes twice to the degree of that vertex. The degree of a vertex reminds, thus, strongly the chemical concept of valence and, in fact, in chemical graph theory it is often used with this meaning (Balaban, 1976; Kier and Hall, 1986; Turro, 1986; Rouvray, 1989; Trinajstić, 1992; Randić and Trinajstić, 1994; references therein). But, while the degree of a vertex in a simple chemical graph denotes directly the connections of the chemical vertices the degree of a vertex in a chemical pseudograph (with loops simulating lone-pair electrons and multiple edges simulating π-type of bonds) is directly

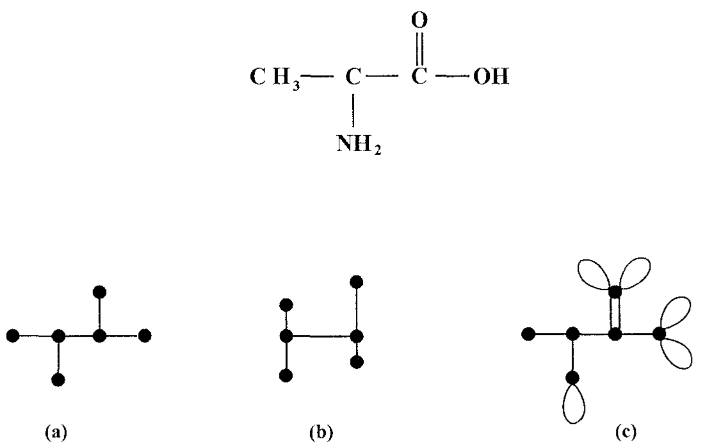

Fig. 1. The amino acid Ala, its corresponding hydrogen suppressed graph (**a**), an equivalent form of the same graph (**b**) and the corresponding pseudograph of Ala (**c**)

related to the chemical concept of valence. From what has been said till now we can easily understand that organic molecules are well suited to be represented by chemical graph or pseudographs whose elementary mathematical properties can be used in (QSAR) quantitative structure-activity and (QSPR) structure-property studies. In Fig. 1 the molecule of the amino acid Ala and its corresponding hydrogen suppressed graph (a) and pseudograph (c) are reported. To clarify how the appearance of a configuration is not important in graph theory, in this same figure an equivalent form of graph (a) is also reported (graph b).

Use of topological concepts is not at all new in modern chemistry, f.e., the topology of a hydrogen suppressed graph, rather than its geometry is the basis of the Hückel molecular orbital theory (Atkins, 1990). Chemical graphs are thus characterized by topological indices related to the degree of a vertex, that are single numbers like most properties they describe and not by bond lengths and angles that are totally disregarded, as topology does not take into account such quantities.

Introduction

A great deal of QSPR and QSAR studies are based on hydrogen suppressed chemical graphs for which graph theoretical indices, the molecular connectivity χ indices, have been defined and further refined all along these last twenty years by Randić (1975, 1988, 1991a, 1991b, 1991c, 1994; Randić et al., 1988) and by Kier and Hall (1977, 1981, 1986 and references therein; Kier et al., 1993; Hall et al., 1993) into a self-consistent theoretical frame known as the molecular MC connectivity theory. All along these years interesting contributions to this theory have also been given by the Zagreb group (Trinajstić, 1992; Mihalić and Trinajstić, 1992; Mihalić et al., 1992; Lucić et al., 1995), by the University of Minnesota group (Basak et al., 1988; Basak et al., 1991;

Basak and Grunwals, 1994) by the Wright State University group (Seybold et al., 1986; Needham et al., 1988) and by the Penn State University group (Hansen and Jurs, 1988; Stanton and Jurs, 1992) as well by other authors (Balaban, 1992; Maier, 1992; Pogliani, 1992–1997). Given citations are certainly not exhaustive but nevertheless they give an idea of the interesting development undergone by the molecular RKH connectivity theory (RKH = Randić, Kier and Hall). Aim of the MC theory can be phrased into the following way: *All predictions can be reached using nothing more than pencil and paper* (Trinajstić, 1983). Clearly, more than pencil and paper are needed to model physicochemical properties of compound with the MC theory, but by the aid of this theory it is possible to perform such modeling in a rather easy and straightforward way handling with few elementary, and easily understandable mathematical tools.

An interesting aspect of the MC-RKH theory being that it allows to model physicochemical properties of classes of organic or biochemical compounds as well as association phenomena in solution, cis-trans isomerism, and even physicochemical properties of inorganic compounds (Pogliani, 1992–1997) by the aid of linear combinations of molecular connectivity indices (LCCI) or of special $X = f(\chi)$ molecular connectivity indices (LCXCI). Aim of the present paper are: i) to present the modeling of the solubility S of the heterogeneous class of amino acids and purine and pyrimidine bases with the minimum number of X indices, removing contemporarily the connectivity degeneracy of three pairs of bases, ii) to model the unfrozen UWC water content of a mixed set of amino acids and inorganic compounds by the aid of connectivity X terms, iii) to refine further the obtained topological model of the genetic code (Pogliani, 1996b) and iv) to model three properties of the DNA-RNA bases by the aid of X terms.

The recently introduced connectivity X terms, can be derived with the aid of a trial-and-error procedure coupled with the forward selection technique from a small- or medium-sized set of connectivity indices (Pogliani, 1996a). Normally, they are good descriptors of many properties of different classes of compounds, and linear combinations of such terms (linear combinations of connectivity terms, LCCT or LCXCI) show both good predictive power and utility. Furthermore, orthogonal indices derived from X connectivity terms show an even improved utility (Pogliani, 1996a).

Method

The chosen medium-sized $\{\chi\}$ set of molecular and valence (denoted by the uppercase v) molecular connectivity indices for numerical encoding the different properties of the different classes of compounds is given by the following 8 indices,

$$\{\chi\} = \{D, D^v, {}^0\chi, {}^0\chi^v, {}^1\chi, {}^1\chi^v, \chi_t, \chi_t^v\}$$

The indices, that can be considered the basis functions of a molecular connectivity MC calculation are based on the degree $\deg(i) = \delta_i$ or δ_i^v (for pseudographs) of each i vertex of a molecular graph, and can be computed as follows. D, sum-delta index is given by (Pogliani, 1992):

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

This kind of index is strictly related to a famous theorem in graph theory, the handshaking theorem (Rosen, 1995) that says: let G = (V,E) be a graph or a pseudograph with E edges and V vertices then: $2e = \Sigma\delta_i$ where e is the number of edges. Since an edge (inclusive loops which contribute twice to the degree of a vertex) is incident with exactly two vertices then it contributes twice to the sum of the degrees of the vertices. Handshaking because of the analogy between an edge having two end points and a handshake involving two hands. It is to notice that for stereoisomeric compounds like amino acids or sugars: D = 2e represents the sum of the vertex degrees of both L- and D-forms, that is, $D = e_L + e_D$, white $e_L = e_D$. This underlines the fact that connectivity indices for the D- and L-forms are equivalent, with $\delta i_L = \delta_{iD}$, and they are, actually, unable to distinguish between these two forms.

The zeroth- and first-order indices are defined as (Kier and Hall, 1986):

$${}^0\chi = \Sigma(\delta i)^{-0.5} \tag{2}$$

$${}^1\chi = \Sigma(\delta_i\delta_j)^{-0.5} \tag{3}$$

χ_t is the total structure connectivity index over the N non-hydrogen atoms of the molecule (Needham et al., 1988):

$$\chi_t = (\delta_1\delta_2. \ldots \delta_N)^{-0.5} \tag{4}$$

where δ_i is the delta cardinal number which represents the count on non-hydrogen σ bond electrons contributed by atom i (Kier and Hall, 1986). The sum in eqs. 1, 2 and 3 is taken over all N vertices and all edges of the molecular graph (corresponding to non-hydrogen atoms and σ bonds) respectively. Replacing δ with valence δ^v (which represents the count of all non-hydrogen electrons contributed by atom i) in eqs. 1–4, the corresponding four valence molecular connectivity χ^v indices are obtained. The δ^v(S) values in amino acids Cys and Met (0.56 and 0.67 respectively) have been taken from Kier and Hall (1986). Recently, supraconnectivity indices have been introduced to improve the modeling of experimental properties (Pogliani, 1993a, 1995, 1996). These supra indices are obtained multiplying the normal molecular connectivity indices by an association constant a, that can be derived from experimental evidence of associative phenomena or inferred from anomalous values of the physicochemical constants.

The modeling of the UWC of a mixed set of amino acids and metal chlorides involves the definition of a MC model for inorganic salts. Now, a clear MC theory for inorganic salts has not yet been defined but it can easily been obtained by extrapolating the concepts already developed for organic molecules. Such an extrapolation even if theoretically somewhat venturesome not only shows practical advantages but has some loose theoretical ground, because he transition between covalent and the ionic bonding type is not an abrupt transition. In fact, between the two extremes of ionic and covalent bonding (NaF and diamond) there is a wide region of intermediate cases (Laing, 1993), that is ionic compounds with some covalent character. This partial covalent character enables to represent the inorganic molecule by a chemical graph or pseudograph (e.g., NaCl = •—•). This is equivalent to consider the gas phase structure of the molecule where bonds are localized. The δ value of a vertex of an inorganic graph equals the number of edges incident with it, and the δ^v value of a vertex of an inorganic pseudograph is obtained by the aid of the following relation (Kier and Hall, 1986):

$$\delta^v = Z^v/(Z - Z^v - 1) \tag{5}$$

where Z^v is the number of valence electrons and Z is the atomic number of the corresponding atom. While δ value of the atoms of the studied inorganic compounds are just 1 or 2 (Me in $MeCl_2$) the valence δ^v values of some interesting inorganic vertices computed by the aid of eq. 4 are [δ^v(Cu) = 2/26]

Li	Na	K	Rb	Cs	Be	Mg	Ca	Sr	Ba	F	Cl	Br	I	O	S
1	1/9	1/17	1/35	1/53	2	2/9	2/17	2/35	2/53	7	7/9	7/27	7/45	6	6/9

The linear estimation problem of a physicochemical property P with $X = f(\chi)$ terms is the estimation of the random variable P_{exp} (experimental properties) in terms of a linear function of X (being $X = \chi$ a special case), that is, $P = |\Sigma_k c_k X_k|$ or in matrix notation

$$P = |C \cdot X| \tag{6}$$

where P is the calculated property of a compound, row vector C is the vector of the coefficients c_k that are determined by the linear least-square procedure and column vector X is the vector of the best connectivity X_k descriptors. The descriptor of the constant c_0 term being the unitary index $X^0 \equiv 1$. If X is a $m \cdot n$ matrix (where, n = number of compounds) then P is a property column vector of the entire class of compounds. Bars in equation 6 stand for absolute value to get rid of negative P values (whenever they are detected) with no physical meaning. Normally, this operation enhances the description of the property (Pogliani, 1996). As the presence of biased measurements of the experimental properties may impair the validity of the results of least squares estimation as the least squares method is fair from robust with respect to biased observations, it will be assumed that our experimental properties are not (consistently) biased. Normally, even though P is not a linear function of χ, it is a linear function of c_i parameters, even if equation involve transcendental functions of χ.

For every index of a LCCT equation the fractional utility, $u_k = |c_k/s_k|$, as well as the average fractional utility $\langle u \rangle = \Sigma u_k/m$ will be given. Connectivity terms can also be orthogonalized generating thus, orthogonal connectivity terms that short-circuits the collinearity problem due to mutual interrelation among X indices, i) improves $\langle u \rangle$, ii) generate coefficients that are stable upon introduction of a new orthogonal index, and iii) create dominant descriptors when X indices are poor descriptors. For details about the technique to derive these indices and their importance the interested reader is referred to the original Randić's works (1991, 1994).

Two procedures for index selection will here be used: the forward selection (fs) and the complete combinatorial (cc) procedures. The forward selection technique is a sequential procedure based on the notion that indices should be inserted one at time until a satisfactory Q-LCCI ($Q = r/s$, where r = correlation coefficient and s = standard deviation of estimates) is obtained, this technique spans a subspace of the complete combinatorial space (Pogliani, 1996). Shortly, this procedure does not alter the already found best indices while is choosing the next best new index. Instead, the complete combinatorial technique is a procedure that searches the entire combinatorial C space spanned by the indices of the {X} set, and extracts the best combinations ($C = \Sigma_r C_{vr}$, $v = m - 1$, number of X variables, X^0 excluded, and $r = 1$ to v). Combinations are sorted following their Q and F values ($F = fr^2/[(1 - r^2)v]$, where f = degrees of freedom = n − m, v here is the number of parameters of a LCCT, from 1 to m − 1). The F statistics can be a valuable aid in discriminating among different LCCT with rather similar Q values and in detecting which descriptor endangers the quality of the modeling. The difference between the forward selection and complete combinatorial procedures is clearly shown in Table 1 where the overall number of possible combinations with the two procedures with growing number of indices has been collected.

Before leaving this paragraph let us spend a word about plotting methods that not always are taken in due consideration in modeling studies. Plotting methods can illustrate and detect violation of assumptions: values should show random fluctuations around the main diagonal of the figure, that is equivalent to saying that residuals should show random fluctuation around a value of zero. Clusters of positive and negative values might suggest, that a curvilinear trend in the data should be investigated. In a set of values obtained in sequence, there should not be long runs of values on the same side of the main diagonal of the figure. This is equivalent to saying that there should not be systematic trends in the sequence of residuals. Unfortunately, it is difficult to quantify what consti-

Table 1. Number (N°) of possible combinations for m indices with the forward selection (fs) and complete combinatorial (cc) technique

M	N° of fs combinations	N° of cc combnations
2	3	3
4	10	15
6	21	63
8	36	255
10	55	1,023
20	210	1,048,575
30	465	1,073,741,823

Table 2. The molecular connectivity indices for 21 amino acids (AA)

AA	D	D^v	$^0\chi$	$^0\chi^v$	$^1\chi$	$^1\chi^v$	χ_t	χ_t^v
Gly	8	20	4.28446	2.63992	2.27006	1.18953	0.40825	0.03727
Ala	10	22	5.15470	3.51016	2.64273	1.62709	0.33333	0.03043
Cys	12	23.56	5.86181	4.55358	3.18074	2.40290	0.23570	0.02875
Ser	12	28	5.86181	3.66448	3.18074	1.77422	0.23570	0.00962
Val*	14	26	6.73205	5.08751	3.55342	2.53777	0.19245	0.01757
Thr	14	30	6.73205	4.53473	3.55342	2.21862	0.19245	0.00786
Met	16	26.67	7.27602	6.14607	4.18074	4.04355	0.11785	0.01859
Pro	16	28	5.98313	4.55413	3.80453	2.76688	0.08333	0.00932
Leu	16	28	7.43916	5.79462	4.03658	3.02094	0.13608	0.01242
Ile	16	28	7.43916	5.79462	4.09142	3.07578	0.13608	0.01242
Asn	16	36	7.43916	4.70278	4.03658	2.30434	0.13608	0.00254
Asp	16	38	7.43916	4.57273	4.03658	2.23927	0.13608	0.00196
Lys	18	32	7.98313	5.91594	4.68074	3.36624	0.08333	0.00439
Hyp	18	34	6.85337	4.87159	4.19838	2.84158	0.06804	0.00340
Gln	18	38	8.14627	5.40997	4.53658	2.80434	0.09623	0.00179
Glu	18	40	8.14627	5.27984	4.53658	2.73927	0.09623	0.00139
His	22	42	8.26758	5.81918	5.19838	3.15529	0.03402	0.00080
Arg	22	42	9.56048	6.70883	5.53658	3.60022	0.04811	0.00078
Phe	24	42	8.97469	6.60402	5.69838	3.72222	0.02406	0.00069
Tyr	26	48	9.84493	6.97388	6.09222	3.85651	0.01964	0.00027
Trp	32	54	10.83650	8.10402	7.18154	4.71624	0.00567	0.00009

*In Pogliani, 1993–1994, D and D^v values of Val are incorrectly quoted (14 and 26) and χ_t/χ_t^v are not considered.

tutes a 'long' run. Furthermore, employing plotting methods it is easier to detect the presence of outliers in the data set, a presence that often leads to an inflated standard deviation.

Results

In Table 2 and 3 the values for the molecular connectivity indices of 21 amino acids (AA) and 23 purines and pyrimidines (PP) with five significant digits

Table 3. Calculated χ values for 23 purine and pyrimidien bases (PP)

PP	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
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
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

A Adenine, *G* Guanine, *U* Uracil, *T* Thymine, *C* Cytosine, *OA* orotic acid, *UA* uric acid, *X* Xanthine, *M* methyl, *P* propyl, *B* butyl, *I* isobutyl, *Cf* Caffein = 137MMMX = 7MTp, *Tb* Theobromine = 37MMX, *Tp* theophylline = 13MMX.
Compounds with (+), (●) and (□) have similar {χ} values.
Tp and Tb derivatives together with A, G, Cf, IsoG, UA, X, and HypoX are purines, the remnants are pyrimidines.

have been collected. While natural amino acids show no degeneracy in their {χ} values, purines and pyrimidines of Table 3 show three pairs of compounds with the same {χ} values (see explanation at the bottom of the Table). In Table 4 the experimental solubility S for 20 amino acids, the experimental solubility S for 23 purine and pyrimidine bases taken from the available literature (Pogliani, 1995; Lide, 1991–1992), and the unfrozen water content UWC for 8 amino acids (Nakashima and Suzuki, 1984) have been collected. In Table 5 the unfrozen water content UWC of 5 metal chlorides MeCl (Nakashima and Suzuki, 1984) with their corresponding molecular connectivity index values are reported. Table 6 and 7, which describe the encoding of the genetic code, will be discussed in the text. Table 8 shows the experimental molar absorption coefficient ε_{260}, together with the experimental first ΔE_1 and second ΔE_2 singlet excitation energies of the nucleotide DNA/RNA bases taken from Ladik and Appel (1966).

Table 4. Experimental solubility, S (at 25°C in units of grams per kg of water) for 20 L-amino acids AA, unfrozen water content UWC (g H_2O/g AA) for 8 AA, and experimental solubility, S (at the indicated T°C, in units of grams per kg of water) for 23 purines and pyrimidines (PP)

AA	UWC	S	PP	S (T°C)
Gly		251	7I8MTp	6.3 (20)
Ala		167	7B8MTp	4.5 (20)
Cys		–	7ITp	27 (20)
Ser	0.48	422	7BTp	3.7 (30)
Val		58	1BTb	5.6 (30)
Thr	0.72	97	7PTp	231.1 (30)
Met		56	1PTb	13.8 (30)
Pro	1.07	1,622	7ETp	36.6 (30)
Leu		23	1ETb	39.8 (30)
Ile		34	Cf	25.8 (30)
Asn		25	Tp	08.1 (30)
Asp		5	Tb	0.54 (30)
Lys	0.93	6	UA	0.02 (20)
Hyp	0.70	361	OA	1.8 (18)
Gln		42	X	0.5 (20)
Glu	0.97	8.6	IsoG	0.06 (25)
His	0.66	43	G	0.04 (40)
Arg	0.46	181	HypoX	0.7 (19)
Phe		29	A	0.9 (25)
Tyr		0.5	T	4.0 (25)
Trp		12	5MC	4.5 (25)
			U	3.6 (25)
			C	7.7 (25)

Table 5. Unfrozen water content UWC (g H_2O/g MeCl' of 5 metal chlorides (MeCl) and their corresponding molecular connectivity index values

MeCl	UWC	D	$^0\chi$	$^1\chi$	D^v	$^0\chi^v$	$^1\chi^v$	χ_t^v
LiCl	6.5	2	2	1	1.7778	2.1339	1.1339	1.1339
NaCl	3.0	2	2	1	0.8889	4.1339	3.4016	3.4017
KCl	1.8	2	2	1	0.8366	5.2570	4.6752	4.6752
$CaCl_2$	4.0	4	2.7071	1.4142	1.6732	5.1832	6.6117	3.7485
$CuCl_2$	4.0	4	2.7071	1.4142	1.6325	5.8733	8.1777	4.6357

Discussion

Modeling the solubility S of amino acids, purine and pyrimidine bases

The simulation of the solubility (grams per Kg of water) of n = 43 amino acids (AA) plus purines and pyrimidines (PP) with supraconnectivity indices for Pro (a = 8 and 1/8), and Ser, Arg, Hyp (a = 2 and 1/2), for the {D, D^v, $^0\chi$, $^0\chi^v$, $^1\chi$, $^1\chi^v$} and {χ, χ_t^v} subsets respectively, and for 7Ptp (a = 4), 1ETb (a = 2), Cf (a = 2), and 7ITp (a = 1.5) for the entire connectivity set can be achieved by the following set of connectivity terms:

$$\{X\} = \{{}^{D}X, {}^{D}X^{v}, {}^{0}X, {}^{0}X^{v}, {}^{1}X, {}^{1}X^{v}, X_{t}, X_{t}^{v}\}$$

These terms are composite connectivity indices derived by a trial-and-error procedure from the normal set of connectivity indices. Each of these terms corresponds to the following composite indices of the following set

$$\{cD\chi_{t}^{v}, cD^{v}\chi_{t}^{v}, c^{0}\chi\chi_{t}^{v}, c^{0}\chi^{v}\chi_{t}^{v}, c^{1}\chi\chi_{t}^{v}, c^{1}\chi^{v}\chi_{t}^{v}, a\chi_{t}, b\chi_{t}^{v}\}$$

where the total valence χ_t^v index multiplies every other index with the exception of χ_t and itself and where c = a·b. While value a for the supraindices of AA and PP has already been defined (see preceding paragraph), b = 1 for every amino acid to avoid that a = k and a = 1/k (for χ_t and χ_t^v) of Pro, Ser, Arg and Hyp cancel each other and b = a for purines and pyrimidines. The degeneracy in χ values of the three couples of purines (see Table 3) for the couples 7PTp/1PTb and 1ETb/7ETp is here removed due to the different association values, in fact, for 7PTp, a = 4, and for 1PTb, a = 1 while for 1ETb, a = 2, and for 7ETp, a = 1. Instead, the degeneracy of the couple IsoG/G is experimentally well grounded as the experimental solubility values of these two compounds are rather similar.

The best 1-X and 2-X-index LCCT are common both to fs and cc techniques

$$\{{}^{D}X\}: Q = 0.008, F = 196, r = 0.909, s = 108.5, \langle u\rangle = 7.6$$

$$\{{}^{D}X, {}^{0}X^{v}\}: Q = 0.016, F = 341, r = 0.972, s = 62.3, \langle u\rangle = 7.2$$

The simultaneous modeling of both classes of compounds with X terms seems optimal, statistical Q and s values can further be improved with the following cc LCCT

$$\{{}^{D}X, {}^{D}X^{v}, X_{t}^{v}\}: Q = 0.019, F = 326, r = 0.981, s = 52.4, \langle u\rangle = 6.0$$

The best combination seems to be the third one, where improvement in Q, r and s nicely compensates the small worsening in ⟨u⟩ and F values. The cc technique does not discover any better LCCT even if r, s and Q improve a little bit with 4 terms, but after this combination only r value improves a little when more indices are included. Noteworthy is the fact that combinations without ${}^{D}X$ term show a drastic worsening of the modeling, a fact that can advantageously be used to shrink the number of combinations to be searched: only those combinations with the ${}^{D}X$ term should be selected. Modeling vectors that best fit the data shown in Fig. 2 are

$$X = ({}^{D}X, {}^{D}X^{v}, X_{t}^{v}, X^{0}), C = (-2366.21, 2222.50, -20336, 3.87615)$$

$$u = (4.68, 7.52, 11.45, 0.42)$$

The poor utility, u_4, of the C_4 constant term (the coefficient of the unitary connectivity term) can be enhanced with the introduction of orthogonal terms, that enhance also the utility of the whole LCCT. In fact, orthogonalizing, ${}^{D}X \equiv {}^{1}\Omega$, $X_t^v \rightarrow {}^{2}\Omega$, and ${}^{D}X \rightarrow {}^{3}\Omega$, we obtain an orthogonal set, which has an enhanced ⟨u⟩ = 12

$$\Omega = ({}^1\Omega, {}^2\Omega, {}^3\Omega, \Omega^0),\ C = (1196.98, -9821.57, 2222.50, -22.5853),$$
$$u = (29, 8.97, 7.52, 2.54)$$

This LCOCT has not only an enhanced u_4 value but also an exceptional improvement in utility of the main (first) connectivity term: from 4.69 to 29, while the overall utility of the full LCOCT now is $\langle u \rangle = 12$, that is, it has doubled relatively to the non orthogonal LCCT. This is the only statistics that has consistently improved; Q, F, r and s statistics of the 3-X LCCT and of the corresponding 3-Ω LCOCT are exactly the same, and, in fact, both series of vectors can be used to obtain the same Fig. 2. Thus, as a rule, as soon as a good LCCT is found, the corresponding LCOCT will show a better overall utility with an improved utility of its main terms.

The following squared X index gives rise to a remarkable single-index LCCT description of S(AA + PP) with no other squared index or combinations of squared indices showing such an interesting improvement in Q, F and ⟨u⟩ values relatively to the found {X} combinations

$$\{({}^D X^v)^2\}\colon Q = 0.011,\ F = 365,\ r = 0.948,\ s = 83.0,\ \langle u \rangle = 10.8$$

The model fails to describe the extremely low solubility values of IsoG, G and UA and the rather low values of A, HypoX, Tb and Tyr, even if the trend is well modeled. This is the origin of the rather large s value of this description. From this model we could infer that the relatively high hydrofobicity of these compounds might better be modeled if more details on the interaction of these molecules with the solvent were known.

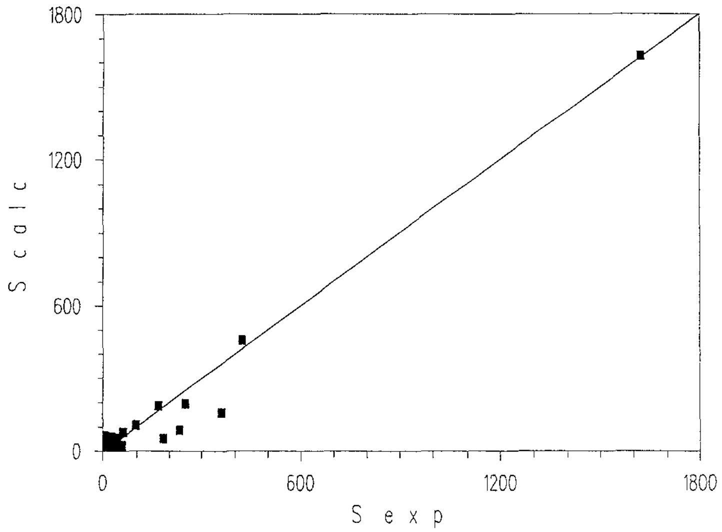

Fig. 2. Plot of the calculated versus the experimental solubility S of 43 amino acids, purines and pyrimidines

Modeling the unfrozen water content UWC of amino acids and inorganic salts

The modeling of the UWC of the heterogeneous class of [AA + MeCl] compounds with the cc selection technique applied to a maximum of 4 χ indices of the given connectivity set (up to now this modeling has been achieved with a subset of the given $\{\chi\}$ set, that does not include χ_t and χ_t^v indices (Pogliani, 1997); show the following results

$$\{D^v\}: Q = 0.71, F = 21, r = 0.81, s = 1.15, \langle u\rangle = 5.7$$

$$\{D^v, \chi_t^v\}: Q = 0.82, F = 14, r = 0.86, s = 1.05, \langle u\rangle = 3.5$$

$$\{{}^0\chi^v, {}^1\chi^v, \chi_t, \chi_t^v\}: Q = 2.41, F = 61, r = 0.984, s = 0.41, \langle u\rangle = 5.4$$

Combinations with 3-χ indices are not as good as the given 2-χ and 3-χ combinations. Clearly, linear combinations of 4 indices to model 13 values is exceeding, thus, a search for better and less X descriptors should be started. We will here use the complete combinatorial technique as only combinations with 2 terms or less will here be considered, and we will also use the same type of terms already used to model S(AA + PP) (but not the same values! Values are now derived from Tables 2 and 5), but without association constants, that is, with c = 1. The best X combinations are, (where: $D\chi_t^v = {}^DX$ and $D^v\chi_t^v = {}^DX^v$)

$$\{{}^DX^v\}: Q = 0.46, F = 8.7, r = 0.67, s = 1.46, \langle u\rangle = 2.5$$

$$\{{}^DX, {}^DX^v\}: Q = 2.12, F = 95, r = 0.975, s = 0.46, \langle u\rangle = 7.3$$

While the single term is a worse descriptor than the single D^v index, the two term LCCT is noteworthy and its vectors will be used to simulate the calculated values of Fig. 3. The single term ${}^DX^v$ is the leading term for the successive

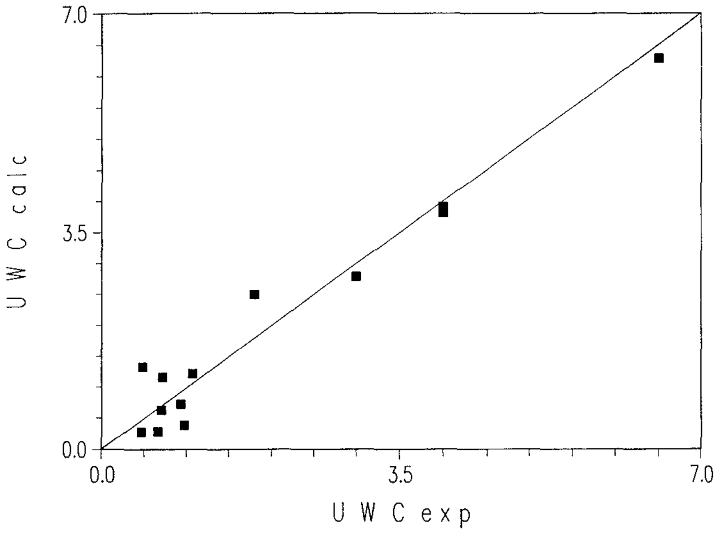

Fig. 3. Plot of the calculated versus the experimental unfrozen water content UWC of amino acids and inorganic salts

combinations, that is, only combinations with this term are statistically interesting. The vectors used to simulate the calculated values of Fig. 3 are

$$X = ({}^{D}X, {}^{D}X^{v}, X^{0}),\ C = (-1.93494, 5.22724, 0.14039),\ u = (10.1, 11.0, 0.77)$$

The low rating of the single term and the high rating of the two term combination unmasks the presence of a dominant orthogonal term, in fact, with ${}^{D}X \equiv {}^{1}\Omega$ and ${}^{D}X^{v} \rightarrow {}^{2}\Omega$ we obtain a rather good single orthogonal (but not as good as the single χ index) descriptor and an improved utility factor of the two Ω terms LCOCT. These linear combination of orthogonal connectivity terms should have the same Q and F score of the corresponding LCCT but a better $\langle u \rangle$ and an improved utility of the constant term (the coefficient of Ω^{0}):

$$\{{}^{2}\Omega\}\text{: } Q = 0.64,\ F = 17.1,\ r = 0.78,\ s = 1.22,\ \langle u \rangle = 4.9$$

$$\Omega = ({}^{1}\Omega, {}^{2}\Omega, \Omega^{0}),\ C = (0.17038, 5.22724, 1.25689),\ u = (8.25, 11.0, 8.25),\ \langle u \rangle = 9.2$$

Encoding the genetic code and amino acids with a supravalence delta number

A modified version of the original notion of degree of a vertex δ can be used to encode the different families and subfamilies of the genetic code following a line of reasoning briefly presented elsewhere (Pogliani, 1996b). Natural amino acids can be coded by a single base triplet, by a subfamily or by a family of base triplets (Lehninger, 1977, see Table 6). The relationship between 64

Table 6. The genetic code

	U		C		A		G	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
	UUA	Leu	UCA	Ser	UAA	ton	UGA	ton
	UUG	Leu	UCG	Ser	UAG	ton	UGG	Trp
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
	CUG	Pro	CCG	Pro	CAG	Gln	CGG	Arg
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

The third nucleotide is seen to be less important than the first two. "*ton*" indicates termination or nonsense codon.

Table 7. The effective valence of the middle B_2 base and its family partitioning power

$^{\Delta\mathrm{eff}}B_2$	$\Delta_{\mathrm{eff}} = 2$ for $B_1 = \{X$: (Fml ptg due to type of B_3)$\}$
^{1}C	$B_1 = \{0$: (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	$B_1 = \{$C, U, G, A: (pur / pyr)$\}$

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 resulting 16 boxes generated by the intersection of the first and second base of this grid are code word families. Eight of these families are divided into subfamilies, that differ only in their third base, that not always is 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 delta number or degree. Clearly, while for the first B_1 and last B_3 base, $\Delta = 1$ or ^1B, for the middle B_2 base, $\Delta = 2$. Now, for an effective encoding of the genetic code we introduce an effective Δ_{cff} number for the middle base, that can assume both 1 and 2 values, i.e., $\Delta_{\mathrm{eff}} = 1$ or 2. This Δ_{eff} number associated with B_2 describes when B_2 needs a nearby B_3 to divide a family into subfamilies and thus characterize an amino acid with four or less base triplets. Thus, while a family partitioning follows only when $\Delta_{\mathrm{eff}} = 2$, a $\Delta_{\mathrm{eff}} = 1$ means that for the encoding of an amino acid B_3 is unimportant. In this way, families and subfamilies of the genetic code can be encoded by $^{\Delta\mathrm{cff}}B_2$. In Table 7, $^{\Delta\mathrm{eff}}B_2$, the conditions for $\Delta_{\mathrm{eff}} = 2$ when $B_2 = \{$A, G, U, C$\}$, that for U and G are controlled by the type of B_1 and the corresponding family partitioning (Fml ptg) due to the type of the third base (pur = $\{$A, G$\}$ and pyr = $\{$U, C$\}$) have been collected.

Table 7 shows that $^{\Delta\mathrm{cff}}B_2 = {}^2$A, means that A needs always a third base, unrewarding of the type of B_1, and generates two 2-member subfamilies (see Table 6, 3rd column) when B_3 is a purine or a pyrimidine. For $^{\Delta\mathrm{cff}}B_2 = {}^1$C no family partitioning is generated by B_3, with the result that, each of the four 4-member families encode an amino acid (see Table 6, 2nd column). The partitioning due to B_3 when $^{\Delta\mathrm{eff}}B_2 = \{^2$G, ^{2}U$\}$ depends on B_1 only if $B_1 = \{$A, U$\}$. For example, for B_1 = A and $^{\Delta\mathrm{eff}}B_2 = {}^2$U, the 4-member family (Table 6, 3rd row, 1st column) is split by B_3 into i) a subfamily AUG, for B_3 = G that encodes the amino acid Met and ii) a 3-member subfamily AUA, AUU and AUC due to $B_3 = \{$A, U, C$\}$, that encode Ile. Instead, for B_1 = U, and $^{\Delta\mathrm{eff}}B_2 = {}^2$U (Table 6, 1st column and row) we have two 2-member pur/pyr families due to $B_3 = \{$pur / pyr$\}$ base encoding two different amino acids (Phe and Leu). With $B_1 = \{$C, G$\}$, and $^{\Delta\mathrm{cff}}B_2 = {}^1$U (Table 6, 1st column, 3rd and 4th rows) two 4-member full families are obtained, each family encoding an amino acid. The same reasoning being valid for ^{1,2}G, with the only difference, that, now, the partitioning (type of B_3) due to $B_1 = \{A, U\}$ are nearly inverted in fact, for $B_1 = A$, $B_3 = \{pur / pyr\}$ and for $B_1 = U$, $B_3 = \{A / G / pyr\}$. The information content of this table is straightforward: C and A as middle bases have two completely different behaviours while middle bases U and G have a behaviour relatively to B_3 that is dependent on B_1.

Now, following the same topological line of reasoning, every natural amino acids in a protein can be viewed as a vertex with $\Delta = 2$ (exceptuating the amino and carboxyl endings with $\Delta = 1$). Furthermore, some amino acids have another functional group responsible of a further binding, like, f.e., in glycoproteins, which contain carbohydrate groups attached covalently to the polypeptide chain and which represent a large group of wide distribution and biological significance (Lehninger, 1977). In these kind of proteins normally Ser, Thr, Gln, Lys have $\Delta = 3$. But, amino acids can also bind with other type of bonds (hydrogen bonds and disulfidic bridge at Cys), to form higher-order structures. If hydrogen bonds and disulfide bridges are also viewed as further connections between two points in the protein graph, then each amino acid undergoing these kind of bonds can have at least $\Delta = 3$ or more, if other bonds are present.

Modeling three properties of DNA/RNA bases

Let us now model the three experimental properties of the nucleic acid bases, U, T, A, G and C of Table 8, that are a preferred subject for quantum theoretical calculations (Ladik and Appel, 1966; Pogliani 1996b).

Simulation of these experimental properties for the five nucleic acid bases has already been achieved with LCCI with encouraging results. Let us now try to find satisfactory connectivity terms for the modeling of these properties and to compare their simulation power with the well-known connectivity χ descriptor. Due to the very low number ($n = 5$) of experimental points to be modeled we will consider only single term combinations, rendering thus the choice of the procedure for combinatorial selection trivial. The best single connectivity χ descriptors of the first $\Delta E_{1,exp}$ singlet excitation energy and of the second $\Delta E_{2,exp}$ singlet excitation energy are

Table 8. Experimental (exp) molar absorption coefficient $\varepsilon_{260,exp}$ at 260 nm and pH = 7.0, first $\Delta E_{1,exp}$ and second $\Delta E_{2,exp}$ singlet excitation energies in eV of the nucleotide DNA bases

DNA Bases	$\varepsilon_{260,exp}$	$\Delta E_{1,exp}$	$\Delta E_{2,exp}$
A	15,400	4.75	5.99
G	11,700	4.49	5.03
U	9,900	4.81	6.11
T	9,200	4.67	5.94
C	7,500	4.61	6.26

A Adenine, *G* Guanine, *U* Uracil, *T* Thymine, *C* Cytosine.

$\Delta E_{1,exp}$:	$\{^0\chi\}$	$Q = 5.51$,	$F = 1.88$,	$r = 0.62$,	$s = 0.1$	$\langle u\rangle = 6.1$
$\Delta E_{2,exp}$:	$\{D^v\}$	$Q = 4.37$,	$F = 17.9$,	$r = 0.93$,	$s = 0.2$	$\langle u\rangle = 8.3$

Simulation achieved by the single index for $\Delta E_{1,exp}$ seems rather deceiving, but orthogonalizing the subset $\{D(\equiv{}^1\Omega), \chi_t^v({}^2\Omega), {}^0\chi({}^3\Omega)\}$ it was possible to find a dominant orthogonal descriptor (${}^2\Omega$) with a better quality: $Q = 7.40$ and $F = 3.39$ (Pogliani, 1996b). Instead, for $\Delta E_{2,exp}$ the single-index combination seems adequate.

The molar absorption $\varepsilon_{260,exp}$ coefficient at 260 nm and pH = 7 reported in Table 8 refers to nucleotides UMP, TMP, AMP, GMP and CMP (the spectra of the corresponding ribo- and deoxynucleotides as well as the nucleosides are essentially identical; Lehninger, 1977), but as the only non-common part of these nucleotides are the 5 bases, U, T, A, G and C, the simulation of this property is performed by the aid of the $\{\chi\}$ and $\{X\}$ values of these bases only. The best single connectivity index for this property is

$\varepsilon_{260,exp}$:	$\{\chi_t\}$	$Q = 0.42$,	$F = 6.47$,	$r = 0.83$,	$s = 2.0$	$\langle u\rangle = 5.2$

The best single trial-and-error connectivity terms for these three properties are,

$\Delta E_{1,exp}$:	$\{^1X\}$	$Q = 8.61$,	$F = 4.60$,	$r = 0.78$,	$s = 0.1$	$\langle u\rangle = 52$
$\Delta E_{2,exp}$:	$\{^2X\}$	$Q = 7.69$,	$F = 55.2$,	$r = 0.97$,	$s = 0.13$	$\langle u\rangle = 46$
$\varepsilon_{260,exp}$:	$\{^3X\}$	$Q = 0.58$,	$F = 12.1$,	$r = 0.90$,	$s = 1.6$	$\langle u\rangle = 8.1$

where ${}^1X = (1/\chi_t^v)^3$, ${}^2X = ({}^0\chi)^3/\chi_t^v$, and ${}^3X = ({}^0\chi^v)^{4.5}(\chi_t\chi_t^v)^2$.

Comparison with the corresponding best single connectivity indices shows the improved quality of the single connectivity terms. The term describing $\Delta E_{1,exp}$ is even better than the orthogonal ${}^2\Omega$ connectivity index. These promising results underline the capacity of the connectivity terms to offer an interesting description of properties, whose quantum chemical description is far from being optimal. The vectors that best fit the data collected in Fig. 4, where the three properties have been collected all together are (in this figure $\varepsilon_{260,exp}$ is reported as $\varepsilon_{260,exp}/1{,}000$).

$X = ({}^1X, X^0)$,	$C = (-1.74767\cdot 10^{-11}, 4.71299)$,	$u = (2.14, 102)$
$X = ({}^2X, X^0)$,	$C = (-1.03367\cdot 10^{-6}, 6.20802)$,	$u = (7.43, 85.1)$
$X = ({}^3X, X^0)$,	$C = (-8.13621\cdot 10^{7}, 13528.2)$,	$u = (3.48, 12.8)$

Conclusion

The connectivity X terms derived by a trial-and-error composition procedure from a medium-sized set of connectivity indices and used to model different properties of heterogeneous classes of compounds including amino acids show an improved quality in every statistic, Q, F, r, s and $\langle u\rangle$. Their superior quality is underlined by the rather satisfactory modeling of the three properties of the five DNA-RNA bases. This quality can be further enhanced through an orthogonalization procedure as it is exemplified by the orthogonalization of the X terms for S(AA + PP) and for the UWC of amino

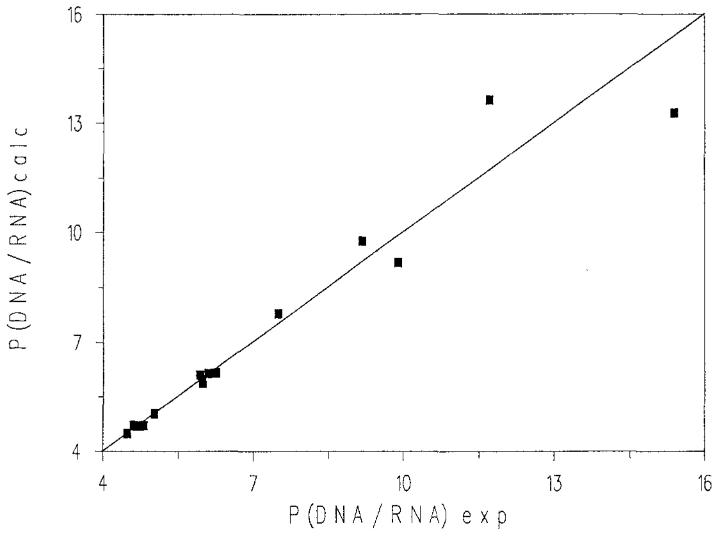

Fig. 4. Plot of the calculated versus the experimental properties P(DNA/RNA) of DNA and RNA bases (first and second singlet excitation energies ΔE_1, ΔE_2, and molar absorption coefficient $\varepsilon_{260}/1{,}000$)

acids and inorganic salts. The linear combination of orthogonal connectivity terms (LCOCT) has in both cases an improved $\langle u\rangle$ and u_k (of the main terms) statistic. These connectivity terms offer another interesting advantage: normally the best single term is a dominant term, and every combination of terms needs its presence to achieve a good quality. This fact reduces considerably the search of the combinatorial space in the complete combinatorial technique as only those combination with the leading X term have to be searched.

The topological description of the genetic code offers a way to synthesize the information content of a normal coding table, underlining the characteristic of the second base, and showing straightforwardly which base B_2 undergoes further connections, i.e., which needs a third base to code an amino acid, and which not and under which conditions, dictated by B_1. Clearly, a detailed understanding of protein properties and of the coordinated processes involved in protein biosynthesis by the aid of topological indices is still elusive, still, these short topological considerations on code words and on protein graphs might possibly help to simplify a picture of the biomolecular world, whose huge wealth of parameter renders, even today, the most simple calculations time- and cost-consuming and the corresponding theoretical results rather poor and questionable.

Acknowledgements

I am indebted to Professors L. B. Kier (Virginia Commonwealth University), M. Randić (Drake University), N. Trinajstić (Zagreb University), L. H. Hall (Eastern Nazarene College), and A. T. Balaban for encouragement, support and concern. I am also indebted to unknown reviewers for critical comments and suggestions and to Dr. C. S. Basak (University of Minnesota) for his constance in sending me his works.

References

Atkins PW (1990) Physical chemistry. Oxford, Oxford

Balaban AT (ed) (1976) Chemical applications of graph theory. Academic Press, London

Balaban AT (1992) Using real numbers as vertex invariants for third-generation topological indices. J Chem Inf Comput Sc 32: 23–28

Barr S (1979) Experiments in topology. Dover, New York

Basak SC, Magnuson VR, Niemi GJ, Regal RR (1988) Determining structural similarity of chemicals using graph-theoretical indices. Discr Appl Math 19: 17–44

Basak SC, Grunwald GD (1994) Molecular similarity and risk assessment: analog selection and property estimation using graph invariants. SAR & QSAR Environ Res 2: 289–307

Basak SC, Niemi GJ, Veith GD (1991) Predicting properties of molecules using graph invariants. J Math Chem 7: 243–272

Hall LH, Kier LB (1977) A molecular connectivity study of electron density in alkanes. Tetrahedron 33: 1953–1957

Hall LH, Kier LB, Frazer JW (1993) Design of molecules from quantitative structure-activity relationship models. 2. Derivation and proof of information transfer relating equations. J Chem Inf Comput Sci 33: 148–152

Hansen PJ, Jurs PC (1988) Chemical applications of graph theory. J Chem Ed 65: 574–580

Kier LB, Hall LH (1981) Derivation and significance of valence molecular connectivity indices. J Pharm Sci 70: 583–589

Kier LB, Hall LH (1986) Molecular connectivity in structure-activity analysis. Wiley, New York

Kier LB, Hall LH, Frazer JW (1993) Design of molecules from quantitative structure-activity relationship models. 1. Information transfer between path and vertex degree counts. J Chem Inf Comput Sci 33: 143–147

Ladik J, Appel K (1966) Pariser-Parr-Pople calculations on different DNA constituents. Theor Chim Acta 4: 132–144

Laing M (1993) A tetrahedron of bonding. Educ in Chem 30: 160–163

Lehninger A (1977) Biochemistry. Worth, New York

Lide DR (ed) (1991–1992) CRC Handbook of chemistry and physics. 72nd edn. CRC Press, Boca Raton

Maier BJ (1992) Wiener and Randić topological indices for graphs. J Chem Inf Comput Sci 32: 87–90

Mihalić Z, Trinajstić N (1992) A graph-theoretical approach to structure-property relationships. J Chem Ed 69: 701–712

Mihalić Z, Nikolić S, Trinajstić N (1992) Comparative study of molecular descriptors derived from the distance matrix. J Chem Inf Comput Sci 32: 28–37

Lucić B, Nikolić S, Trinajstić N, Juretić D, Jurić A (1995) A novel QSPR approach to physicochemical properties of the α-amino acids. Croat Chim Acta 68: 435–450

Nakashima N, Suzuki E-I (1984) Studies of hydration by broad-line pulsed nmr. Appl Spectr Rev 20: 1–53

Needham DE, Wei I-C, Seybold PG (1988) Molecular modeling of the physical properties of the alkanes. J Am Chem Soc 110: 4186–4194

Pogliani L (1992) Molecular connectivity model for determination of isoelectric points of amino acids. J Pharm Sci 81: 334–336

Pogliani L (1993a) Molecular connectivity model for determination of T_1 relaxation times of α-carbons of amino acids and cyclic dipeptides. Comput Chem 17: 283–286

Pogliani L (1993b) Molecular connectivity model for determination of physicochemical properties of α-amino acids. J Phys Chem 97: 6731–6736

Pogliani L (1994a) On a graph theoretical characterization of cis/trans isomers. J Chem Inf Comput Sci 34: 801–804

Pogliani L (1994b) Structure property relationships of amino acids and some dipeptides. Amino Acids 6: 141–153

Pogliani L (1994c) Molecular connectivity descriptors of the physicochemical properties of the α-amino acids. J Phys Chem 98: 1494–1499

Pogliani L (1995a) Modeling the solubility and activity of amino acids with the LCCI method. Amino Acids 9: 217–228

Pogliani L (1995b) Molecular modeling by linear combinations of connectivity indices. J Phys Chem 99: 925–937

Pogliani L (1996a) Modeling with special descriptors derived from a medium-sized set of connectivity indices. J Phys Chem 100: 18065–18077

Pogliani L (1996b) Modeling purines and pyrimidines with the linear combination of connectivity indices – molecular connectivity "LCCI-MC" method. J Chem Inf Comput Sci 36: 1082–1091

Pogliani L (1997) Modeling enthalpy and hydration processes of inorganic compounds. MATH/CHEM/COMP '96: Croat Chem Acta 70 (in press)

Randić M (1975) On characterization of molecular branching. J Am Chem Soc 97: 6609–6615

Randić M (1988) On characterization of three-dimensional structures. Int J Quant Chem: Quant Biol Symp 15: 201–208

Randić M (1991a) Orthogonal molecular descriptors. N J Chem 15: 517–525

Randić M (1991b) Resolution of ambiguities in structure-property studies by use of orthogonal descriptors. J Chem Inf Comput Sci 31: 311–320

Randić M (1991c) Search for optimal molecular descriptors. Croat Chim Acta 64: 43–54

Randić M (1994) Curve-Fitting paradox. Int J Quant Chem: Quant Biol Symp 21: 215–225

Randić M, Trinajstic N (1994) Notes on some less known early contributions to Chemical Graph theory. Croat Chem Acta 67: 1–35

Randić M, Hansen PJ, Jurs PC (1988) Search for useful graph theoretical invariants of molecular structure. J Chem Inf Comput Sci 28: 60–68

Rosen KH (1995) Discrete mathematics and its applications. McGraw-Hill, New York

Rouvray DH (1989) The limits of applicability of topological indices. J Mol Struct (Theochem) 185: 187–201

Seybold PG, May M, Bagal AU (1987) Molecular structure-property relationships. J Chem Ed 64: 575–581

Stanton DT, Jurs PC (1992) Computer-assisted study of the relationship between molecular structure and surface tension of organic compounds. J Chem Inf Comput Sci 32: 109–115

Trinajstić N (1982) Chemical graph theory, 1st edn. CRC Press, Boca Raton

Turro NJ (1986) Geometric and topological thinking in organic chemistry. Angew Chem Int Edn Engl 25: 882–901

Authors' address: Prof. L. Pogliani, Dipartimento di Chimica, Universita' della Calabria, I-87030 Rende (CS), Italy.

Received January 1, 1997