GRAPH-THEORETICAL CONCEPTS AND PHYSIOCHEMICAL DATA

Lionello Pogliani

Dipartimento di Chimica, Università della Calabria, 87030 Rende (CS) Italy.
Email: lionp@unical.it

ABSTRACT

Graph theoretical concepts have been used to model the molecular polarizabilities of fifty-four organic derivatives, and the induced dipole moment of a set of fifty-seven organic compounds divided into three subsets. The starting point of these modeling strategies is the hydrogen-suppressed chemical graph and pseudograph of a molecule, which works very well for second row atoms. From these types of graphs a set of graph-theoretical basis indices, the molecular connectivity indices, can be derived and used to model properties and activities of molecules. With the aid of the molecular connectivity basis indices it is then possible to build higher-order descriptors. The problem of 'graph' encoding the contribution of the inner-core electrons of heteroatoms can here be solved with the aid of odd complete graphs, Kp-(p-odd). The use of these graph tools allow to draw an optimal modeling of the molecular polarizabilities and a satisfactory modeling of the induced dipole moment of a wide set of organic derivatives.

Keywords: Chemical Graphs, Complete Graphs, Graph-Theoretical Indices, Molecular Basis Indices, Terms, Polarizabilities, Induced dipole moments.

1 INTRODUCTION

The easiest way to keep us from drowning in the rising sea of experimental data is to compress these data into algorithms, the easier the algorithm, the better. The advantages of this procedure are several, e.g., it allows the development of software to enable drug testing *in silico*. A theory is needed to tell us how to compress data into algorithms. A very successful theory was introduced in chemistry more than hundred years ago, the chemical graph theory, which was further refined during the second half of the twentieth century, giving rise to several theoretical branches spread over different fields of physical and pharmaceutical chemistry (Balaban, 1985; Kier & Hall, 1986, 1992, and 1999; Hansen & Jurs, 1988; Trach, Devdariani & Zefirov, 1990; Randiæ, 1991; Trinajstiæ, 1992; Randiæ & Trinajstiæ, 1994; Basak & Grunwald, 1995; Temkin, Zeigarnik, & Bonchev, 1996; Gutman, Klavzar & Mohar, 1997; Seybold, 1999; Balaban & Devillers, 1999; Randiæ, Mills & Basak, 2000; Pogliani, 2000; Klein, & Brickman, 2000; Galvez, Garcia-Domenech & de Gregorio-Alapont, 2000; Estrada & Uriarte, 2001; Diudea, 2001; King & Rouvray, 2002). One branch of this theory, the molecular connectivity theory, developed more than half a century ago by Randiæ, Kier, and Hall (Randiæ, 1975; Kier & Hall, 1986), states that for every chemical graph (cg) that has a set of graph theoretical basis indices, $\{\beta\}$, then it also has the property P, i.e., $\{cg\}(\{\beta\} \rightarrow P)$, in short-hand notation. This property also encompasses activity, A. Clearly, such an assertion has a probabilistic character.

In the following sections we will explain (i) what a chemical graph is, (ii) what graph-theoretical indices are, and (iii) how they can be used to model the values of the properties, P. The properties that will be modeled with the given graph theoretical concepts are the polarizability and the induced dipole moment of two heterogeneous sets of organic compounds taken from a recent molecular mechanics study (Ma, Lii & Allinger, 2000).

2 GRAPH THEORETICAL CONCEPTS

The following graph theoretical concepts are needed in molecular connectivity theory for modeling properties and activities are the following (for a quick reference see Rosen, 1995).

Graph, G = {V, E}: a graph can be defined as set of vertices, V, and a set of edges, E that connect these vertices. The ***degree*** of a vertex of a graph is the number of edges that occur with it.

Pseudograph: a graph that allows for multiple connections and loops (or self-connections). These types of graphs allow a faithful encoding of a molecule, as it is possible to encode multiple bonds and non-bonding electrons with them. The loop at a vertex contributes twice to its degree.

Chemical (or molecular) Graph or Pseudograph: a graph or pseudograph representation of a chemical compound or molecule. Normally, but not always, in chemical graph theory use is made of hydrogen-suppressed (or depleted) graphs of pseudographs, i.e., chemical graphs of pseudographs whose hydrogen atoms, if there are some, have been deleted, leaving only the non-hydrogen atoms, i.e., second or higher-row atoms, whose principal quantum number is $n \geq 2$. Throughout the present paper we will be concerned with these types of chemical graphs and pseudographs (Figure 1), even if they are just cited as graphs.

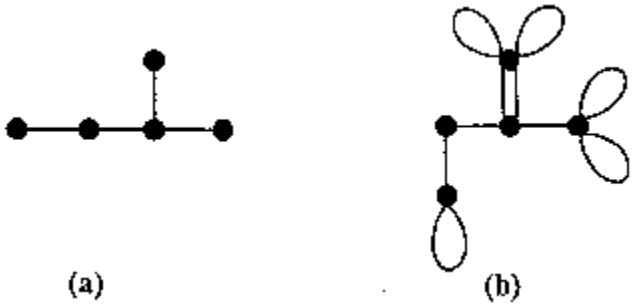

Figure 1. The chemical graph (a) and pseudograph (b) of the amino acid Glycine. The bent backbone in (b) indicates that in graph theory angles and geometric distances have no meaning.

Complete Graphs: A graph G is complete (Figure 2) if every pair of its vertices are adjacent. A complete graph of order p is denoted by K_p, (p-1=r) and is ***r*-regular**. A graph is r-regular if it has all vertices with the same degree r.

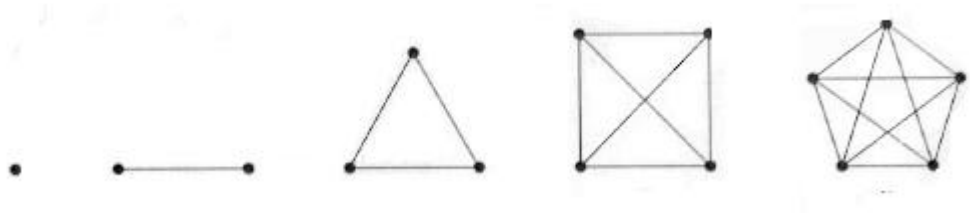

Figure 2. From left to right: the K_1, K_2, K_3, K_4 and K_5 complete graphs.

To encode the inner-core electrons of heteroatoms, odd complete graphs, K_p-(p-odd) where p = 1, 3, 5, and 7 will be used. In Figure 3 the pseudograph-odd-complete graph for the CH_3Cl molecule is shown.

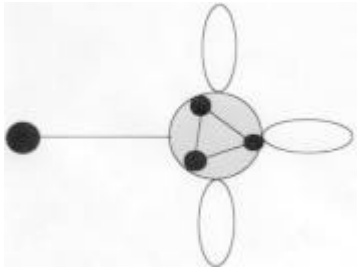

Figure 3. The pseudograph-odd-complete graph for the hydrogen-suppressed CH_3Cl molecule. The inner-core electrons of C and Cl are encoded with K_1 and K_3 complete graphs, respectively.

Adjacency Matrix of a graph: is a square and symmetrical matrix of order n, where n is the number of vertices (i.e., atoms) of the chemical graph, and whose elements g_{ij} are equal to ones if the vertices i and j of the graph are adjacent otherwise they are zero. Self-connections are not allowed in the

adjacency matrix of a graph matrix, i.e., $g_{ii} = 0$. A pseudograph adjacency matrix encodes not only the features of a graph adjacency matrix but has $g_{ii} = ps_{ii} \neq 0$, where $ps_{i,i}$ encodes the pseudograph characteristics of the adjacency matrix. In this matrix ps_{ii} equals the sum of the self-connections (or loops, which are counted twice) and multiple connections of vertex *i*. Thus, the hydrogen-suppressed pseudograph of a triatomic system, which also includes information about the odd complete graphs for the inner-core electrons (Pogliani, 2002a) yields

$$A = (p \cdot r + 1)^{-1}_{Kp} \begin{pmatrix} ps_{1,1} & g_{1,2} & g_{1,3} \\ g_{2,1} & ps_{2,2} & g_{2,3} \\ g_{3,1} & g_{3,2} & ps_{3,3} \end{pmatrix} \quad (1)$$

Factor $(p \cdot r + 1)^{-1}_{Kp}$ encodes the odd complete graph characteristics. This factor depends on the K_p of each vertex. Its contribution renders the A matrix asymmetrical, as it is evident from the following particular 3x3 matrix for CS_2 (K_1 for C and K_3 for S, atom superscripts denote the row),

$$A(S^1 = C^2 = S^3) = \begin{pmatrix} 5/7 & 1/7 & 0 \\ 1/1 & 2/1 & 1/1 \\ 0 & 1/7 & 5/7 \end{pmatrix} \quad (2)$$

The term 1/1≡ 1 has been written to allow an easier decoding of the formalism. The defined concept of a vertex degree or valency of atom i, δ_i, of a chemical graph is thus the number of simple connections (i.e., σ bonds) present in *i* in a hydrogen-suppressed chemical graph, and is the sum of the g_{ij} elements, in a row of the matrix, A. The vertex degree or valency of atom i, δ_i^v(ps) (valence delta) of a chemical pseudograph is instead the number of total connections, including self-connections (i.e., σ, π bonds and non-bonding electrons) present in *i* in a hydrogen-suppressed chemical graph. It is the sum of the $[g_{ij} + ps_{ii}]$ elements of a row of the A matrix. The vertex degree or valency of atom *i*, δ_i^v of a chemical pseudograph plus the odd-complete graph for the contribution of the inner-core electrons can, thus, be directly obtained by the aid of the following algorithm:

$$\delta_i^v = \delta_i^v(ps)/[p \cdot r + 1] \quad (3)$$

It is practically the sum of all the elements in a row of the full A matrix (see matrix 2). For p = 1, $\delta^v = \delta^v$(ps), and in alkanes, $\delta^v = \delta^v(ps) = \delta$. For the different halogens of the halo-compounds we have p = 1, 3, 5, 7, i.e., $\delta^v(F) = 7$, $\delta^v(Cl) = 7/7$, $\delta^v(Br) = 7/21$, $\delta^v(I) = 7/43$. Parameter $p \times r$ in Eq. (3) equals $\Sigma_i d_i$ for the complete graph, and is an interesting invariant in graph theory, as the *Handshaking theorem* of graph theory states that it equals twice the number of connections, since a connection occurs with two vertices, thus, it contributes twice to the sum of the degree of the vertices (Rosen, 1995).

3 THE GRAPH THEORETICAL BASIS INDICES

The chosen subset of graph theoretical basis indices, {β}, and the raw material of QSPR (Quantitative Structure-Property Relationships) studies is made up of the following medium-sized subset of eight molecular connectivity indices that are defined within the frame of molecular connectivity theory (Kier & Hall, 1986)

$$\{\beta\} = \{D, {}^0\chi, {}^1\chi, \chi_t, D^v, {}^0\chi^v, {}^1\chi^v, \chi^v_t\} \quad (4)$$

These indices are based on the δ and δ^v connectivity numbers of a hydrogen-suppressed graph and pseudograph plus K_p-(p-odd) graph (for the inner-core electrons) respectively, and their definitions are

$$D = \Sigma_i \delta_i \quad (5)$$

$${}^0\chi = \Sigma_i (\delta_i)^{-0.5} \quad (6)$$

$${}^1\chi = \Sigma (\delta_i \delta_j)^{-0.5} \quad (7)$$

$$\chi_t = (\delta_1 \cdot \delta_2 \cdot \delta_3 \cdot \ldots\ldots \cdot \delta_N)^{-0.5} \quad (8)$$

Index χ_t (and χ^v_t) is the total molecular connectivity index. The sums in Eqs. (5 and 6), as well as the product in Eq. (8), are taken over all *n* vertices (i.e., atoms) of the chemical graph (i.e., molecule). The sum in Eq. 7 is over all edges (σ bonds in a molecule) of the chemical graph. By replacing δ with δ^v (see Eq. 3) in all these Eqs. the subset of valence χ indices $\{D^v, {}^0\chi^v, {}^1\chi^v, \chi^v_t\}$ is obtained. With the aid of these basis indices it is then possible to construct, through trial-and-error, higher-order structural invariants, *S*, among which are the molecular connectivity terms, X = f(χ), (Pogliani, 2000). These terms have the general form of a rational function,

$$S = [a(\beta_1)^m + b(\beta_2)^n]^q / [c(\beta_3)^r + d(\beta_1)^p]^s \qquad (9)$$

Here β is a basis index, and S = X if $\beta = \chi$. Depending on the type of β basis indices other higher-order indices can be constructed (Pogliani, 2002a; 2002b). The optimization parameters, *a* - *d*, *m* - *s*, can either be negative, or zero or one. In these last two cases the rational function condenses into a much simpler form. As can be seen from Eq. (9) the power of each basis index is again optimized, which means that the original power (- 1/2, see Eqs. (5-8)) loses its restrictive meaning.

4 THE STRUCTURE-PROPERTY RELATION

Two types of Structure-Property relation will here be used: (i) the linear relation, $P = c_1S + c_0U_0$, and the multilinear relation, $P = \Sigma c_iS_i$, where c_i, are the regression coefficients, and c_0 the regression coefficient of the unitary index, $U_0 \equiv 1$. Normally, the linear type relation has: $S = X$, while the multilinear relation has: $S_i = \beta_i$, here: $\{\beta\} = \{\chi\}$. The multiple linear relations can normally be written as a dot product: $P = \mathbf{C}\cdot\boldsymbol{S}$, where $\mathbf{C} = (c_1, c_2, ..., c_0)$, and $\mathbf{S} = (b_1, b_2, .., U_0)$. To avoid negative P_{calc} values, with no physical meaning it is advantageous to use the modulus equation, i.e, $P = |\Sigma c_iS_i,|$, where the bars stand for an absolute value.

The statistical performance of a structural descriptor, S, be it a single descriptor or a combination of basis indices, is controlled by a quality factor, $Q = r / s$, and by the Fisher ratio $F = fr^2/[(1-r^2)\nu]$, where *r* and *s* are the correlation coefficient and the standard deviation of the estimates, respectively, *f* is the number of degrees of freedom = $N - (n + 1)$, n is the number of variables, and *N* is the number of data. Parameter Q has no absolute meaning as it is an 'intra' statistical parameter able only to compare the descriptive power of different descriptors for the same property; further, this property should always be given in the same scale. The F ratio, which has the character of an 'inter'-statistical parameter, tells us, even if Q improves, which additional descriptor endangers the statistical quality of the combination. For every index of a linear combination as well as for U_0 the fractional utility, $u_i = |c_i/s_i|$, where s_i is the confidence interval of c_i, and the average fractional utility $\langle u\rangle = \Sigma u_i/(n+1)$ is given. If the modeling relation is linear, then $\langle u\rangle = (u_1+u_0)/2$. The utility statistics gives indirect information about the role of the descriptor in the modeling equation, as it allows the detection of descriptors that give rise to unreliable coefficient values (c_i), whenever they have a high deviation interval (s_i). Recently (Pogliani, 2002b; 2002c), the critical importance of the standard deviation of the estimate *s* has been underlined, so that it is advantageous to know directly how much this statistic improves along a series of improved descriptors. For this reason we introduce here the ratio $s_R = s_0/s_i$, where s_0 is the *s* value of the best single-index description and s_i refers to the *s* values of the improved sequential descriptions. Thus, halving of s_i can be read as a doubling in s_R, which will allow a direct measure of the progress of *s* along a series of sequential descriptions. It should be stressed that, now, (i) all statistical parameters will grow with improved modeling (ii) every model is under the control of all these statistics, and (iii) nothing justifies using an improved Q as a sign of improved modeling. The richness in statistical parameters can also be used to detect possible printing errors, as redundancy is very useful in the construction of self-correcting codes. For an interesting discussion about the Q statistics see Todeschini (2001). To avoid bothering the reader with the dimensional problems of the modeling equation, every property P should be read as P/P° where P° is the unitary value of the property, so that this choice allows P to be read as a purely numerical number (Berberan-Santos & Pogliani, 1999).

5 MODELING INDUCED DIPOLE MOMENTS

5.1 Alcohols, Amines and Ethers

Table 1 shows the experimental induced dipole moment values, the corresponding calculated ones, and the residual modulus, $|\Delta E| = |\mu(E) - \mu(C)|$. Throughout the present modeling study we will follow the division in subclasses proposed by a recent molecular mechanics study, with different MM3 algorithms completed by quantum mechanical parameters (Ma, Lii & Allinger, 2000). At the present level of sophistication in molecular connectivity (MC) studies, and, also it seems, in the MM3 studies, it is practically impossible to model the induced dipole moments for an entire class of heterogeneous compounds without diminishing the quality of the overall modeling. The studied compounds have a simple and nearly constant topology. Nevertheless, if the modeling of the entire class of compounds is attempted the functional groups introduce consistent discontinuities in the quality of the model of this property. This can be clearly seen with the poor model of the subclass, aldehydes, ketones, acids and esters, which are made up of four different subclasses.

The reader should not forget that for the following subclasses of compounds the odd complete graph algorithm gives rise to $\delta^v = \delta^v(ps)$, a value that can also be obtained with other algorithms based on atomic concepts (Kier & Hall, 1986; Pogliani, 2002a). The best descriptor for this particular property in this set of compounds is the following combination of basis indices (here: $s_0 = 0.35$)

$$\{{}^0\chi,\ D^v,\ {}^0\chi^v\}\text{: } Q = 7.919,\ F = 110,\ r = 0.974,\ s = 0.12,\ s_R = 3.5,\ n = 22,\ \langle u\rangle = 7.1,$$
$$\mathbf{u} = (5.0,\ 6.9,\ 9.2,\ 7.5),\ \mathbf{C} = 0.84028,\ 0.10678,\ -1.17700,\ 0.99778)$$

The attentive reader should keep an eye on the ${}^0\chi^v$ index, as it seems important for this and the next property. The correlation vector, **C**, of the last description will be used to derive the calculated $\mu(C)$ values and the corresponding residual modulus, $|\Delta E| = |\mu(E) - \mu(C)|$, for this class of compounds. These last two sets of values, $\mu(C)$ and $|\Delta E|$, shown in Table 1, underline the good quality of the modeling of this property for this class of compounds. The following interesting X term could also be detected,

$$X = [\chi_t^{v}\cdot{}^1\chi^v - 0.009\cdot{}^0\chi]^{-0.01}\text{: } Q = 5.15,\ F = 140,\ s_R = 1.9,\ n = 22$$

5.2 Aldehydes, Ketones, Acids, and Esters

It is practically impossible to achieve a satisfactory description of the dipole moment of this class of compounds with the molecular connectivity indices alone. The best descriptor, which is a rather poor descriptor, is the following X term, (here, $s_0 = 0.56$)

$$X = [D^v - 1.4\cdot D]^{0.7}({}^0\chi)^{1.1}\text{: } Q = 2.03,\ F = 35,\ r = 0.804,\ s_R = 1.4,\ n = 21,\ \langle u\rangle = 11$$
$$\mathbf{u} = (5.9,\ 16),\ \mathbf{C} = (-\ 0.08643,\ 3.54828)$$

From the calculated and residual values of Table 1 we note that (i) the modeling is far from being optimal, but (ii) nevertheless these values are not at all absurd, and a large deviation can only be detected for formic acid.

5.3 Sulfides and Phosphines

Table 1 also shows the experimental and calculated dipole moment values, as well as the residual modulus for sulfides and phosphines. Now, due to S and P atoms, with n = 3, the K_p-(p-odd) algorithm (Eq. 3) allows the following d^v values for S and P in sulfides and phosphines: $d^v(\text{-}SH)=5/7$, $d^v(\text{-}S\text{-})=6/7$, $d^v(\text{-}PH_2)=3/7$, $d^v(\text{-}PH\text{-})=4/7$, $d^v(\text{-}P<)=5/7$ ($s_0 = 0.15$). The following combination of two basis indices has the best descriptive quality,

$$\{{}^0\chi,\ D^v\}\text{: } Q = 10.8;\ F = 28;\ r = 0.915;\ s_R = 1.9,\ n = 14$$

An optimal X term shows an interesting improvement over the preceding combination

$$X = [({}^0\chi^v - {}^0\chi)^{2.4} + 0.06\cdot\chi_t^v] \,/\, [({}^1\chi^v)^{0.5} - 0.1\,{}^0\chi^v]^{1.2}$$

Q = 23.72, F = 272, r = 0.979, s_R = 3.8, <u> = 45, **u** = (16, 73), n = 14
C = (- 1.96602, 1.73169)

The calculated and residual values of sulfides and phosphines shown in Table 1 demonstrate the good quality of the model.

Table 1. Experimental (E) and calculated (C) Dipole Moments, μ (D), and residual modulus, $|\Delta\mu| = |\mu(E) - \mu(C)|$, for (I) Alcohols, Amines, Ethers, (II) Aldehydes, Ketones, Acids, Esters, and (III) Sulfides and Phosphines.°

*Compound**	m(*E*)	m(*C*)	|Dm|	*Compound*	m(*E*)	m(*C*)	|Dm|
Methanol	1.700	1.616	0.084	3Pentanone	2.720	2.725	0.005
Ethanol	1.680	1.591	0.088	CyPentanone	3.250	3.046	0.204
Propanol	1.550	1.567	0.017	CyHexanone	3.250	3.274	0.024
2-Propanol	1.580	1.512	0.068	Formic acid(*t*)	3.790	2.402	1.388
tert-Butanol	1.670	1.414	0.256	Acetic acid	1.700	2.096	0.396
1,2Etdiol(*g*)	2.410	2.644	0.234	Propionic acid	1.550	1.910	0.360
1,2Prdiol(*g*)	2.568	2.565	0.003	2MePr-acid	1.790	1.708	0.082
DiMethylether	1.310	1.292	0.018	2,2diMePr-acid	1.700	1.553	0.147
Et,Me-ether	1.174	1.268	0.094	Me-Formate	1.770	2.168	0.398
DiEt-ether	1.061	1.243	0.182	Et-Formate(*t*)	1.980	1.978	0.002
Me,Pr-ether(*t-t*)	1.107	1.243	0.136	Me-Acetate	1.706	1.909	0.204
Di-IsoPr-ether	1.130	1.084	0.046	Et-Acetate	1.780	1.772	0.008
THF	1.750	1.654	0.096	Me-Propionate	1.750	1.772	0.022
THP	1.740	1.629	0.111	Et-Propionate	1.810	1.689	0.121
2MetoxyEtOH	2.360	2.321	0.039	Methane thiol	1.520	1.525	0.005
Methylamine	1.296	1.249	0.047	Ethane thiol (*t*)	1.580	1.590	0.010
Ethylamine	1.220	1.224	0.004	1-Propan thiol (*t*)	1.598	1.635	0.037
DiMethylamine	1.010	0.971	0.039	Dimethyl sulfide	1.500	1.473	0.027
n-Propylamine	1.180	1.200	0.020	Ethyl, Me sulfide (*t*)	1.560	1.517	0.043
Isopropylamine	1.190	1.145	0.045	2-Me-2-Pr thiol	1.660	1.624	0.036
TriMethylamine	0.612	0.801	0.189	Diethyl sulfide (*t*)	1.520	1.546	0.026
N-Me-aminoEt	0.880	0.946	0.066	Methyl phosphine	1.100	1.120	0.020
Formaldehyde	2.360	2.960	0.600	Ethyl phosphine (*t*)	1.226	1.194	0.032
Acetaldehyde	2.730	2.820	0.090	Dimethyl phosphine	1.230	1.164	0.066
Propanal	2.520	2.730	0.210	IsoPr phosphine (*t*)	1.230	1.218	0.011
Butanal	2.740	2.705	0.040	Trimethyl phosphine	1.192	1.266	0.074
2,2MePr-aldehyde	2.660	2.686	0.026	tert-Butyl phosphine	1.170	1.224	0.054
Acetone	2.930	2.687	0.242	Et-diMephosphine (*g*)	1.310	1.299	0.011
2Butanone	2.775	2.668	0.107				

° The different classes are divided by solid lines; * B-diene= Butadiene, Cy = cyclo, Et = Ethyl, Ethane, g = gauche, Me = Methyl, Pr = Propyl, Propion, Propan, *t = trans; t-t* = trans-trans, THF = tetrahydrofuran; THP = tetrahydropyran.

6 MODELING THE MOLECULAR POLARIZABILITIES

The experimental mean polarizability $<\alpha(E)> = (\alpha_1(E)+ \alpha_2(E) + \alpha_3(E))/3$, and total molecular principal polarizabilities, $\alpha_1(E)$, $\alpha_2(E)$, and $\alpha_3(E)$, of fifty-four and forty organic compounds, respectively, are collected in Table 2. Whenever $\alpha_i(E)$ values are absent, some $<\alpha(E)>$ values are the result of quantum mechanical calculations. This Table also shows the calculated polarizability values (i) $<\alpha(C)>$, and the residual modulus, $|\Delta\alpha| = |<\alpha(E)> - <\alpha(C)>|$. Here we will model the $<\alpha(E)>$ values, and we will check the quality of the best descriptor for this property when it is used to model the single $\alpha_i(E)$.

Throughout this case index $^0\chi^v$ is the best single-index descriptor.

P	(n)	{β}	Q	F	R	s_R	<u>	**u**
<α>	(54)	$\{^0\chi^v\}$	1.045	867	0.971	1.3	15	(29, 0.4)
α_1	(40)	$\{^0\chi^v\}$	0.576	232	0.927	1.1	8.8	(15,2.4)
α_2	(40)	$\{^0\chi^v\}$	0.888	609	0.970	1.1	13	(25, 1.4)
α_3	(40)	$\{^0\chi^v\}$	0.996	633	0.971	1.2	14	(25, 2.5)

The dominant character of this index allows us to use the forward combination procedure, or greedy algorithm (Pogliani, 2000), to derive the best combinations of indices, among which the following combination of four connectivity indices shows an exceptional modelling quality (here, $s_0 = 1.15$). It is also worth noting that the following optimal combination for <α(E)> is also a good descriptor of the single α_1(E), α_2(E), and α_3(E) properties. Thus, this linear combination is, practically, validated by the nice modeling of the α_i(E) polarizabilities.

P	(n)	$\{\beta_1, \beta_2, \beta_3, \beta_4\}$	Q	F	r	S_R	<u>	**U**
<α>	(54)	$\{^0\chi^v, {}^1\chi, D^v, \chi_t^v\}$	2.086	863	0.993	2.5	6.5	(12, 9.5, 7.7, 2.7, 3.4)
α_1	(40)	$\{^0\chi^v, {}^1\chi, D^v, \chi_t^v\}$	0.799	112	0.963	1.5	3.1	(2.3, 5.3, 4.5, 2.4, 1.0)
α_2	(40)	$\{^0\chi^v, {}^1\chi, D^v, \chi_t^v\}$	1.414	386	0.989	1.8	5.2	(7.4, 6.5, 5.2, 2.3, 4.7)
α_3	(40)	$\{^0\chi^v, {}^1\chi, D^v, \chi_t^v\}$	1.093	191	0.978	1.4	3.0	(8.1, 1.6, 1.6, 0.7, 3.3)

The following simple but interesting term, $X = [3\cdot{}^0\chi^v + {}^1\chi]$, could be detected, which has: Q = 1.414, F = 1587, r = 0.984, s_R = 1.6, n = 54, <u> = 21 for <α>. The relatively low utility value of the constant unitary index, u_0, is mainly due to the vanishingly small value of the corresponding regression parameter c_0. Small deviations near zero can have a dramatic effect on the utility value.

The calculated <α(C)> values in Table 2 have been obtained with the correlation vector of the last combination, **b** = ($^0\chi^v$, $^1\chi$, D^v, χ_t^v, U_0), i.e, **C** = (1.50028, 2.47483, -0.13929, 0.23527, -0.55904). The model is impressive as confirmed by the $|\Delta\alpha|$ = |<α(E)> - <α(C)>| values. The good quality of the modeling is confirmed by $<a_{loo}>$ values based on the leave-one-out method and the corresponding $|\Delta_{loo}|$ = |<α(E)> - $<\alpha_{loo}>$| values, shown in the same Table 2, as well as by Figure 4. In this figure the calculated <α(C)> values (Calc. Polarization) are plotted against the corresponding <α(E)> values (Exp Polarization). The algebraic values of the residuals are also shown around the zero line.

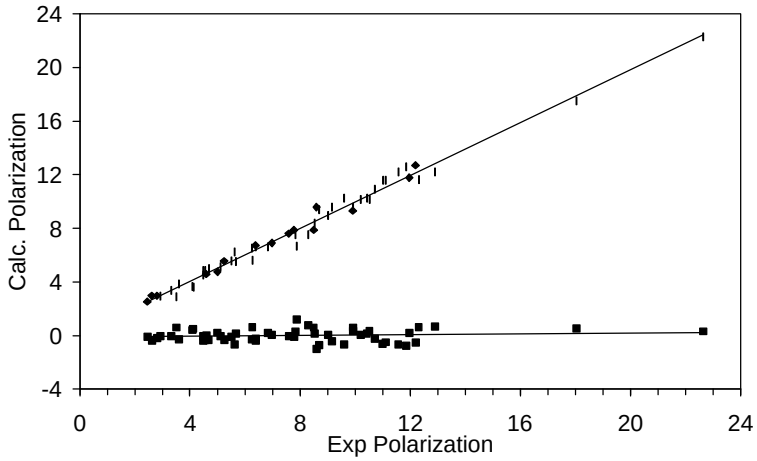

Figure 4. Plot of the calculated (Calc.) vs. the experimental polarization (Exp Polarization) together with the algebraic residual values.

Table 2. Experimental <α(E)>, α_i(E) (i = 1-3), computed <α(C)> molecular polarizabilities, and the corresponding residual modulus $|\Delta\alpha|$ of fifty four organic compounds in units of $Å^3$. $<\alpha_{loo}>$ is the predicted value based on the leave-one-out method and $|\Delta_{loo}|$ is the corresponding residual.*

Compound	$<a(E)>$	$<a(C)>$	$\|Da\|$	$<a_{loo}>$	$\|D_{loo}\|$	$a_1(E)$	$a_2(E)$	$a_3(E)$
Ethane	4.48	4.87	0.39	4.89	0.41	5.49	3.98	3.98
Propane	6.38	6.61	0.23	6.62	0.24	7.66	5.74	5.74
Neopentane	10.20	10.14	0.06	10.14	0.06	10.20	10.20	10.20
Cyclopropane	5.50	5.58	0.08	5.42	0.08	5.74	5.74	5.04
Cyclopentane	9.15	9.58	0.43	10.73	1.58	9.68	9.17	8.40
Cyclohexane	11.00	11.59	0.59	11.65	0.65	11.81	11.81	9.28
Ethylene	4.12	3.60	0.52	3.57	0.55	4.82	3.71	3.25
Propene	6.26	5.63	0.63	5.61	0.55			
2MePropene	8.29	7.51	0.78	7.48	0.81			
Trans-2-Butene	8.49	7.88	0.61	7.86	0.63			
Cyclohexene	10.70	10.93	0.23	10.95	0.25			
Butadiene	7.87	6.68	1.19	6.64	1.23	11.93	6.14	5.54
Benzene	9.92	9.56	0.36	9.53	0.39	11.20	11.20	7.36
Toluene	12.30	11.64	0.66	11.59	0.71			
HexaMeBenzene	22.63	22.29	0.34	22.06	0.57	22.63	22.63	22.63
Acetylene	3.50	2.89	0.61	2.85	0.65	4.79	2.85	2.85
Propyne	4.68	5.01	0.33	5.02	0.34	6.14	3.94	3.94
$C(C\equiv CH)_4$	12.19	12.70	0.51	12.83	0.64	12.19	12.19	12.19
Allene	5.00	4.76	0.24	4.75	0.25	8.97	4.43	4.43
Methanol	3.32	3.36	0.04	3.36	0.04	4.09	3.23	2.65
Ethanol	5.11	5.13	0.02	5.13	0.02	5.76	4.98	4.50
2-Propanol	6.97	6.93	0.04	6.93	0.04			
Cyclohexanol	11.56	12.19	0.63	12.24	0.68			
Dimethylether	5.24	5.54	0.30	5.54	0.30	6.38	4.94	4.39
p-Dioxane	8.60	9.56	0.96	9.62	1.01	9.40	9.40	7.00
Methylamine	3.59	3.86	0.27	3.87	0.28	3.94	3.40	3.38
Formaldehyde	2.45	2.54	0.09	2.55	0.10	2.76	1.70	1.83
Acetaldehyde	4.59	4.58	0.01	4.58	0.01			
Acetone	6.39	6.73	0.34	6.74	0.35	7.37	7.37	4.42
F-Methane	2.62	2.96	0.34	2.98	0.36	3.18	2.34	2.34
TriF-Methane	2.81	2.92	0.15	3.00	0.19	2.87	2.87	2.69
TetraF-Methane	2.92	2.95	0.03	2.99	0.07	2.92	2.92	2.92
Cl-Methane	4.55	4.87	0.32	4.89	0.34	5.68	3.99	3.98
DiCl-Methane	6.82	6.61	0.21	6.60	0.22	8.81	6.30	5.36
TriCl-Methane	8.53	8.39	0.14	8.39	0.14	9.42	9.42	6.74
TetraCl-Methane	10.51	10.15	0.36	10.23	0.28	10.51	10.51	10.51
Br-Methane	5.61	6.25	0.64	6.28	0.67	6.91	4.96	4.96
DiBr-Methane	8.68	9.36	0.68	9.40	0.72			
TriBr-Methane	11.84	12.59	0.75	12.72	0.88	13.00	13.00	9.53
I-Methane	7.59	7.59	0.00	7.59	0.00	9.02	6.87	6.87
DiI-Methane	12.90	12.22	0.68	12.11	0.79			
TriI-Methane	18.04	17.48	0.56	16.88	1.16	18.69	18.69	16.74
$CH_2=CCl_2$	7.83	7.51	0.32	7.50	0.33	8.96	8.79	5.75
Cis-CHCl=CHCl	7.78	7.88	0.10	7.80	0.10	9.46	7.80	6.08
DiSilane	11.10	11.58	0.48	11.82	0.72			
Formamide	4.08	3.65	0.43	3.62	0.46			
Acetamide	5.67	5.53	0.14	5.53	0.14			
Acetonitrile	4.48	4.52	0.04	4.52	0.04	5.74	3.85	3.85
Propionitrile	6.24	6.53	0.29	6.53	0.29			
Tert-BuCyanide	9.59	10.25	0.66	10.31	0.72	10.71	9.03	9.03
BenzylCyanide	11.97	11.78	0.19	11.74	0.23	16.16	11.60	8.15
TriCl-Acetonitrile	10.42	10.25	0.17	10.23	0.19	10.70	10.29	10.29
Pyridine	9.92	9.31	0.61	9.25	0.67	10.72	10.43	6.45
Thiophene	9.00	8.95	0.05	8.95	0.05	10.15	10.14	6.70

* $<\alpha(E)> = \Sigma_i\alpha_i(E)/3$, α_1, α_2, and α_3 are the principal molecular polarizabilities. Some $<\alpha(E)>$ values were computed with quantum methods (Ma, Lii & Allinger, 2000).

To avoid unreliable linear combinations due to collinearity among the basis indices, while maintaining their modeling power, it is advantageous to orthogonalize the corresponding basis indices. Or rather to obtain the orthogonal correlation coefficients of the correlation vector $\mathbf{C}(\Omega)$. For example, for $<\alpha(C)>$ there is no need to derive single Ω_i values as these correlation coefficients can be obtained with the aid of the coefficient of the sequential regressions (Randiæ, 1991; Pogliani, 2000). Thus, the orthogonal

correlation vector for $\boldsymbol{S}(\Omega) = ({}^{1}\Omega, {}^{2}\Omega, {}^{3}\Omega, {}^{4}\Omega, U_0) \leftarrow \boldsymbol{S} = ({}^{0}\chi^{v}, {}^{1}\chi, D^{v}, \chi_t^{v}, U_0)$, is: $\mathbf{C}(\Omega)$ = (2.36719, 0.83838, -0.12524, 0.23527, 0.11242).

7 CONCLUSION

Graph-theoretical tools based on concepts defined within the framework of the molecular connectivity theory are able to optimally model the mean polarizability of fifty-four organic compounds $<\alpha(E)>$, including forty values of the polarizability components, $\alpha_1(E)$, $\alpha_2(E)$, and $\alpha_3(E)$. These last values have been 'externally' modeled with the best descriptor for $<\alpha(E)>$. For the induced dipole moments the influence of the functional groups play a critical role in determining the quality of the modeling, as already suggested by a molecular mechanics study (Ma, Lii & Allinger, 2000). Following a suggestion from the cited MM study, four different subclasses were chosen for the modeling study. The resulting modeling is rather encouraging for the subclass of alcohols, amines, ethers, and, even more so for the subclass of sulfides, and phosphines, which include molecules with only two different functional groups. The modeling is unsatisfactory for the subclass {aldehydes, ketones, acids, esters}, which is made up of compounds with four different functional groups. The introduction of other types of connectivity indices, like semiempirical terms (Pogliani, 2000) might also help to improve the model. Actually, one of the main difficulties in molecular connectivity modeling, and in other modeling studies also, is mimicking the role played by the quantitatively unknown intermolecular interactions that in many cases shape the overall short- or long-lived supramolecular structures (Pogliani, 2002b; 2002c). Dipole moments play a bigger role than polarizability in shaping the overall topology of the supramolecular species, and this could be the main reason for the poor modeling of this property, especially for those subclasses of compounds made up of molecules with a large variety of functional groups.

A pivotal tool of the present modeling study is surely the introduction and use of odd complete graphs to encode the inner-core electrons of heteroatoms. This is in line with recent studies that have underlined the importance of these types of graphs in solving the problem of the inner-core electrons in chemical graph theory (Pogliani, 2002a), and especially in molecular connectivity theory. Other studies do not exclude the possibility of using sequential complete graphs, where p = 1, 2, 3, 4, ... (Pogliani, 2003a; 2003b)

8 ACKNOWLEDGMENTS

The author would like to thank the two anonymous referees for their valuable hints for improving the paper.

9 REFERENCES

Balaban, A.T. (1985) Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci.* 25, 334-343.

Balaban, A.T. & Devillers, J. (Eds.) (1999) *Topological Indices and Related Descriptors*, Amsterdam: Gordon & Breach.

Basak, S. & Grunwald, G.D. (1995) Estimation of lipophilicity from molecular structural similarity. *New J. Chem.* 19, 231-237.

Berberan-Santos, M.N. & Pogliani, L. (1999) Two alternative derivations of Bridgam's theorem. *J. Math. Chem.* 26, 255-261.

Diudea, M. V. (Ed.) (2001) *QSPR/QSAR Studies by Molecular Descriptors*, New York: Nova Science.

Estrada, E. & Uriarte, E. (2001) Recent advances on the role of topological indices in drug discovery. *Curr. Med. Chem.* 8, 1573-1588.

Galvez, J., Garcia-Domenech, R. & de Gregorio-Alapont, C. (2000) Indices of differences of path lengths: novel topological descriptors derived from electronic interferences in graphs. *J. Comput-Aided Mol.Design* 14, 679-687.

Gutman, I., Klavzar, S. & Mohar, B. (Eds,) (1997) *MATCH – Commun. Math. Comp. Chem*. 35, 1-259.

Hansen, P.J. & Jurs, P.C. (1988) Chemical applications of graph theory. *J.Chem.Ed*. 65, 574-580.

Kier, L.B. & Hall, L.H. (1986) *Molecular Connectivity in Structure-ActivityAnalysis*, New York: Wiley.

Kier, L.B. & Hall, L.H. (1992) Atom description in QSAR models: development and use of an atom level index, *Adv. Drug Res*. 22, 1-38.

Kier, L.B. & Hall, L.H. (1999) *Molecular Structure Description. The Electrotopological State*, New York: Academic Press.

King, R.B. & Rouvray, D. (Eds.) (2002) *Topology in Chemistry*, Chichester, UK: Horwood Pub.Lim.

Klein, D.J. & Brickman, J. (Eds.) (2000) *MATCH – Commun. Math. Comp. Chem*. 42, 1-290.

Ma B., Lii, J. H. & Allinger N. L. (2000) Molecular Polarizabilities and Induced Dipole Moments in Molecualr mechanics, *J. Comp. Chem.* 21, 813-825.

Pogliani, L. (2000) From molecular connectivity indices to semiempirical connectivity terms: recent trends in graph theoretical descriptors. *Chem. Rev*. 100, 3827-3858.

Pogliani, L. (2002a) Algorithmically compresses data and the topological conjecture for the inner-core elctrons. *J.Chem.Inf.Comput.Sci*. 42, 1028-1042..

Pogliani, L. (2002b) Topics in molecular modeling: dual indices, quality of modeling and missing information, truncation. *J.Mol.Struct. (THEOCHEM)* 581, 87-109.

Pogliani, L. (2002c) Limits with modeling data and modeling data with limits. *Data Science J.* 1, 76-88.

Pogliani, L. (2003a) Model with dual indices and complete graphs. The heterogeneous description of the dipole moments and polarizabilities. *New J. Chem*. in press.

Pogliani, L. (2003b) Introducing the complete graphs for the inner-core electrons. *In.J.Chem.* in press.

Randiæ, M. (1975) On characterization of molecular branching. *J. Am. Chem.Soc*. 97, 6609-6615.

Randiæ, M., (1991) Orthogonal molecular descriptors. *New J. Chem*. 15, 517-525.

Randiæ, M. & Trinajstiæ, N. (1994) Notes on less known early contributions to chemical graph theory. *Croat.Chem.Acta* 67, 1-35.

Randiæ, M., Mills, D. & Basak, S. (2000) On characterization of physical properties of amino acids. *Int.J.Quant.Chem.* 80, 1199-1209.

Rosen K. H. (1995) *Discrete Mathematics and its Applications*, New York: McGraw-Hill.

Seybold, P.G. (1999) Exploration of molecular structure-property relationships. *SAR & QSAR in Environ.Res.* 10, 101-115

Trinajstiæ, N. (1992) *Chemical graph Theory*, Boca Raton, Fl: CRC Press.

Temkin, D., Zeigarnik, A.V. & Bonchev, D. (1996) *Chemical Reaction Networks. A Graphical Theoretical Approach*, Boca Raton, Fl: CRC Press.

Todeschini, R. (2001) Some Observations about the Pogliani Q quality Index. *Chemometrics web news*, (Feb). Retrieved 28 January 2003, from UniMIB, Milano Chemometrics & QSAR Research Group website: http://www.disat.unimib.it/chm/CHMnews.htm.

Trach, S.S., Devdariani, R.O. & Zefirov, N.S. (1990) Combinatorial models and algorithms in chemistry. Topological-configurational analogs of the Wiener index. (translated from) *Zhurnal Organicheskoi Khimii* 26, 921-932.