STUDIA UNIVERSITATIS BABEŞ-BOLYAI, CHEMIA, XLIX, 2, 2004

Dedicated to Professor Valer Fărcăşan
at his 85th anniversary

TOPOLOGICAL DESCRIPTORS IN WEIGHTED MOLECULAR GRAPHS, APPLICATIONS IN QSPR MODELING

OLEG URSU[a], MIRCEA V. DIUDEA[b]

[a,b] Faculty of Chemistry and Chemical Engineering
Babes-Bolyai University, 400028 Cluj, Romania

ABSTRACT. Organic compounds containing heteroatoms or multiple bonds can be represented as vertex weighted and edge weighted molecular graphs. Different types of weighting schemes can be applied by computing parameter set containing each type of heteroatom. Topological descriptors derived from such weighting schemes are used to develop quantitative structure – property relationship (QSPR) models (property being the molar refraction) for a mixed set of alcohols, amines, and organic halides.

INTRODUCTION

Among a large variety of topological descriptors used today, many of them are defined only for simple graphs, representing eventually alkanes and cycloalkanes, ideal classes of compounds for investigating the molecular connectivity, size, branching, cyclicity, and shape on the variation of molecular properties. However, most of the organic compounds of interest are functional derivatives containing heteroatoms and/or multiple bonds. Usually, such a compound can be represented as a weighted molecular graph. Early applications of weighted molecular graphs are connected with computation of polynomials and spectra of heteroconjugated compounds. Several particular methods of computing topological descriptors from molecular graphs containing heteroatoms and/or multiple bonds were proposed: *Kier* and *Hall*[1] electrotopological state descriptors, *EATI* super-indices[2], Sanderson electronegativity valences (*SEV*)[3], walk matrix and walk operator derived descriptors[4,5]. In this study we discuss the use of following weighting schemes: formal charge, Sanderson electronegativity and covalent radius, within walk matrix and walk operator to provide weighted molecular descriptors for a QSPR study.

WEIGHTING SCHEMES, MOLECULAR MATRICES AND STRUCTURAL DESCRIPTORS

Sanderson electronegativity valences. *Diudea* and *Silaghi*[3] have proposed group electronegativity valences denoted *SEV* and defined by:

$$SEG_i = \left(SEA_i \cdot SEH^{hi}\right)^{1/(1+hi)} \tag{1}$$

$$SEV_i = \left(SEG_i\right)^{1/(1+v_i)} \tag{2}$$

SEA and *SEH* denote the Sanderson electronegativity for the atom *i* and hydrogen respectively, the number of hydrogen atoms attached to the group *i* is denoted by h_i while v_i stands for the degree of *i*. In the case of multiple bonds $v_i = \sum_j b_{ij}$ where b_{ij} is the conventional bond order for the bonds around *i*. Note that group electronegativities obey the electronegativity equalizing principle within the group and per molecule. The *SEV* values are used further in the construction of the *DS* index which showed good correlation with several physicochemical properties.

^{e}W$_M$ matrix. In its general form, the walk degree can be defined as:

$$^{e}w_{\boldsymbol{i}} = \sum_{j} [\mathbf{M}^{e}]_{ij} \qquad (3)$$

The raising at a power *e*, of a square matrix **M**, can be eluded by applying the algorithm of Diudea, Topan and Graovac.[4] It evaluates a (topological) property of a vertex *i*, by iterative summation of the first neighbors contributions. The algorithm, called **^{e}W$_M$** , is defined as:

$$\mathbf{M} + \mathbf{P} = {}^{0}\mathbf{W}_{M} \qquad (4)$$

$$[^{e+1}\mathbf{W}_{M}]_{ii} = \sum_{j} ([\mathbf{M}]_{ij} [^{e}\mathbf{W}_{M}]_{jj}) \qquad (5)$$

$$[^{e+1}\mathbf{W}_{M}]_{ij} = [^{e}\mathbf{W}_{M}]_{ij} = [\mathbf{M}]_{ij} \qquad (6)$$

where **^{e}W$_M$** is the matrix of walk degrees. The diagonal elements, [^{e}**W**$_M$]$_{ii}$ equal the row sum RS_i of **M**^{e}, or in other words, they are *walk degrees,* **$^e w_{M,i}$** (weighted with the property collected by the vertex property **P** diagonal matrix):[5]

$$[^{e}\mathbf{W}_{M}]_{ii} = \sum_{j} [\mathbf{M}^{e}]_{ij} = {}^{e}w_{\boldsymbol{M,i}} \qquad (7)$$

The half sum of the local invariants **^{e}w$_{M,i}$**, in a graph, defines a global invariant, called the *walk number*, **^{e}W$_M$**:

$$^{e}W_{M} = {}^{e}W_{M}(G) = \frac{1}{2} \sum_{i} {}^{e}W_{M,i} \qquad (8)$$

When **M = A; C**, the quantity **^{e}W$_M$** (or simply **^{e}W**) represents the so called *molecular walk count*; when **M = D**, (*i.e.*, the distance matrix) then **^{e}W$_M$** is the Wiener number of rank *e*.

Within *TOPOCLUJ* program[6] the formal charges are calculated as:

$$ch_{i,j} = \log\left[(SEG_j / SEG_i)^{1/(d_{i,j})^2} \right] \qquad (9)$$

$$ch_i = \sum_j ch_{i,j}$$

In the above relations, the log function provides the sign for the formal charge ch_{ij}, viewed as a distance decreasing perturbation of SEG_i produced by the atom *j* (see the exponent, d_{ij} being the Euclidean distance separating atoms *i* and *j*).

An *N*×*N* array collecting the entries ch_{ij} is called the charge matrix **CH**, whose row sums ch_i represent the total partial charge on hydride group/atom *i* in the molecule. This matrix can be processed by our program in various weighting schemes.

METHOD

Data input. All structures were sketched and optimized using *PM3* semiempirical parameterization with *HYPERCHEM* molecular modeling software package.[7] Molar refraction data reported by Diudea and Silaghi[3], and optimized geometries for the three sets of organic derivatives (see Tables 1, 2, 3) represent the input for molecular descriptors generation by *TOPOCLUJ* software package. The molecular descriptors thus generated are used as input for statistical analysis.

Statistical analysis. All three data sets were analyzed using simple linear regression, in STATISTICAL TOOLBOX, MATLAB.[8] The best models, showing excellent correlation with the chosen property (see eqs. 10-12) were validated by *LOO* (*Leave one Out*) method. Data are presented as follows: Tables 1 to 3 list the sets of compounds, the values for the experimental property, calculated property by the best estimation model (given in the QSPR eqs. Below each table) and the predicted property by *LOO*.

TABLE 1

Molar refractions of halogen derivatives

No.	Compound	Molar refraction		Predicted values
		Exp	Calc	by LOO
1	1-chloropropane	20.847	22.248	22.570
2	1-choloro-2-methypropane	33.940	31.628	31.462
3	3-cholorpentane	38.354	37.383	37.274
4	2-brompropane	38.314	37.485	37.390
5	1-brompropane	38.264	37.281	37.171
6	2-brombutane	42.891	43.024	43.062
7	1-bromo-2-methylpropane	47.610	48.794	49.611
8	1-brombutane	25.360	26.119	26.215
9	3-brompentane	30.161	30.809	30.858
10	2-iodobutane	23.935	22.599	22.308
11	3-iodopentane	23.679	24.129	24.206
12	2-iodopentane	28.651	28.405	28.383
13	1-iodopentane	28.537	27.897	27.834
14	1-iodohexane	28.347	29.975	30.103
15	1-iodoheptane	33.068	34.186	34.270

$$MR = 5.588 - 36.584 \cdot {}^{1}W_{[AD]}[CH] \qquad (10)$$

R^2=0.977; n = 15; *s* = 1.199; *F* = 564.062; R^2_{pred} = 9.969

TABLE 2

Molar refractions of amines

No.	Compound	Molar refraction		Predicted values
		Exp	Calc	after LOO
1	Trimethylamine	19.595	20.192	20.336
2	1-aminopropane	33.641	33.582	33.578
3	2-amino-2-methylpropane	33.816	34.025	34.037
4	1-aminobutane	33.794	34.025	34.038
5	1-amino-2,2-dimethylpropane	33.452	33.582	33.589
6	1-amino-3-methylbutane	33.290	33.139	33.131
7	3-aminopentane	38.281	38.636	38.672
8	Dipropylamine	38.038	37.750	37.724
9	1-aminopentane	38.003	37.750	37.727
10	Diisopropylamine	42.920	42.804	42.781
11	Butyldimethilamine	33.852	34.025	34.035
12	Triethylamine	19.401	19.305	19.279
13	Butylethylamine	47.783	47.858	47.891
14	1-aminohexane	24.257	23.917	23.871
15	Dimethylpentylamine	24.079	23.917	23.895
16	2-aminoheptane	28.471	28.528	28.531
17	1-aminoheptane	28.672	28.528	28.518
18	Diisobutylamine	28.617	28.528	28.522
19	Dimethylisobutylamine	33.515	33.582	33.585
20	Tripropylamine	28.728	28.528	28.514

$$MR = 5.030 + 1.803 \cdot {}^{1}W_{[AD]}\left[SEG\right] \qquad (11)$$

R^2=0.999; n = 20; s = 0,241; F = 16842; R^2_{pred} = 0.998

TABLE 3

Molar refractions of alcohols

No.	Compound	Molar refraction		Predicted values
		Exp	Calc	after LOO
1	Isopropanol	17.705	17.488	17.447
2	^{n}Propanol	26.618	26.719	26.725
3	2-methyl-1-propanol	31.211	31.335	31.339
4	^{n}Butanol	31.183	31.335	31.340
5	2-methyl-2-butanol	31.351	31.335	31.334
6	2-pentanol	31.138	31.335	31.342
7	3-methyl-1-butanol	31.489	31.335	31.330
8	2-methyl-1-butanol	31.164	31.335	31.341
9	^{n}Pentanol	31.180	31.335	31.340
10	3-pentanol	31.429	31.335	31.332
11	2-methyl-3-pentanol	35.675	35.951	35.962
12	3-methyl-3-pentanol	17.529	17.488	17.480
13	4-methyl-2-pentanol	35.822	35.951	35.956
14	4-methyl-3-pentanol	35.931	35.951	35.951
15	4-methyl-1-pentanol	36.094	35.951	35.945
16	2-methyl-1-pentanol	40.899	40.566	40.542

No.	Compound	Molar refraction		Predicted values
		Exp	Calc	after LOO
17	2-ethyl-1-butanol	40.447	40.566	40.575
18	^{n}Hexanol	40.439	40.566	40.575
19	2,4-dimethyl-3-pentanol	40.737	40.566	40.554
20	3-ehtyl-3-pentanol	40.625	40.566	40.562
21	2-methyl-1-haxanol	40.638	40.566	40.561
22	^{n}Heptanol	45.521	45.182	45.136
23	2-methyl-2-heptanol	22.103	22.104	22.104
24	3-methyl-3-heptanol	45.207	45.182	45.178
25	4-methyl-4-heptanol	44.920	45.182	45.217
26	6-methyl-1-heptanol	22.067	22.104	22.108
27	2-ehtyl-1-hexanol	26.722	26.719	26.719
28	^{n}Octanol	26.680	26.719	26.722
29	2,6-dimethyl-4-heptanol	26.904	26.719	26.709
30	2-methyl-2-octanol	26.697	26.719	26.721
31	4-ethyl-4-heptanol	26.801	26.719	26.715

$$MR = 3.971 + 5.994 \cdot {}^{1}W_{[AD]}\left[CR\right] \qquad (12)$$

R^2=0.999; n = 31; s = 0,16; F = 70162; R^2_{pred} = 0.999

Excellent results in simple linear regression, gave opportunity to test correlation ability of these descriptors in mixed sets of compounds. Thus we tried a global correlation model for all three above sets. The obtained model, with all three descriptors used before, has also shown a good correlation (see eq. 13), thus proving the weighting schemas and molecular descriptors are suitable and useful in QSPR studies. The calculated molar refractions by (13) were plotted against the experimental values (Fig. 1).

$$MR = 11.809 - 10.167 \cdot {}^{1}\mathrm{W}_{[AD]}\left[CH\right] + 22.914 \cdot {}^{1}\mathrm{W}_{[AD]}\left[CR\right] - 12.975 {}^{1}\mathrm{W}_{[AD]}\left[SEV\right]$$

R^2=0.987; n=66; F=1583.48 (13)

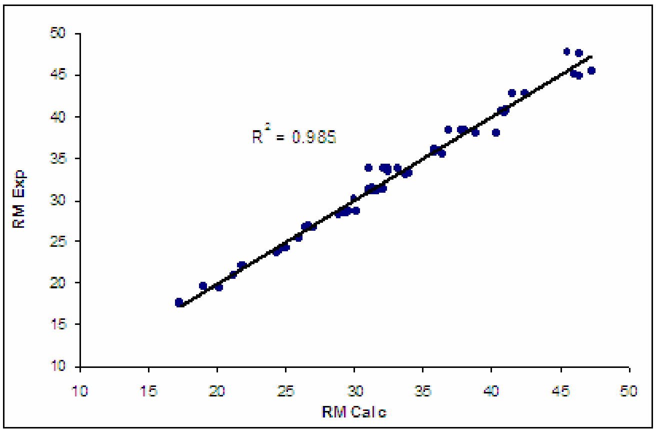

Figure 1. Plot of experimental vs calculated molar refractions for all three data sets.

CONCLUSIONS

Using weighting schemes and molecular descriptors derived from them showed that good models can be obtained even for mixed sets of organic compounds. The statistics of the obtained models appear excellent, both in estimation and prediction. The ability of TOPOCLUJ software program in providing various weighting schemes is a real promise.

REFERENCES

[1] Kier, L. B.; Hall, L. H. *Molecular Structure Description. The Electrotopological State*; Academic Press: New York, **1999**.

[2] Guo, M.; Xu, l.; Hu, C.Y.; Yu, S.M. Study on Structure-Activity Relationship of Organic Compounds - Applications of a New Highly Discriminating Topological Index, *Commun. Math. Comput. Chem. (MATCH),* **1997**, *35*, 185-197.

[3] Diudea, M.V.; Silaghi-Dumitrescu, I. Valence Group Electronegativity as a Vertex Discriminator, *Rev. Roum. Chim.* **1989**, *34*, 1175-1182.

[4] Diudea, M. V.; Topan, M.; Graovac, A. Layer Matrices of Walk Degrees, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1071 -1078.

[5] Diudea, M. V. Wiener and Hyper-Wiener Numbers in a Single Matrix, *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 833-836.

[6] Diudea, M. V.; Ursu O., Layer matrices and distance property descriptors. *Indian J. Chem.*, 42*A*, **2003**, 1283-1294.

[7] Hyperchem 7.03, Hypercube Inc. http://www.hyper.com

[8] STATISTICS TOOLBOX, MATLAB 6.5.1, Mathworks Inc., http://www.mathworks.com