STUDIA UNIVERSITATIS BABEŞ-BOLYAI, CHEMIA, XLV, 1-2, 2000

CORRELATING STUDIES BY CLUJ AND SZEGED INDICES

A.A. KISS, G. TURCU AND M.V. DIUDEA*

Faculty of Chemistry and Chemical Engineering,
"Babeş-Bolyai" University, 3400 Cluj, Romania

ABSTRACT. The correlating ability of Cluj and Szeged indices was tested on two sets of poly-chlorinated bipheniles (PCB) and barbiturates. The molecular property was the *vapor pressure,* at 25^oC and 100^oC, and *log P*, respectively. The results are discussed in comparison with some well known topological indices.

INTRODUCTION

QSPRs (Quantitative Structure-Property Relationships) link in a quantitative manner the physico-chemical properties of chemicals with the molecular structure.[1]

Some molecular properties (i.e. those of which numerical value vary with changes in the molecular structure) such as the normal boiling point, critical parameters, viscosity, solubility, retention chromatographic index, are often used for characterizing chemicals in databases. However, a certain property is not always available in tables or other reference sources. It is just the case of newly synthesized compounds. As a consequence, methods of evaluating physico-chemical properties from the structural features of organic molecules become very important.

Monitoring the environmental pollution needs the prediction of toxicity of chemicals in air, waste waters and sole. **QSAR**s (Quantitative Structure-Property Relationships) can be used to predict the toxicity accurately, without using more expensive experimental methods.[2] Drug research and production is also related to the **QSAR** techniques.[3]

In this work new correlating results by using Cluj and Szeged topological indices are reported, with the aim to demonstrate the capability of our indices to model the molecular properties of organic compounds.

* E-mail: diudea@chem.ubbcluj.ro

The above mentioned indices are calculated on the ground of the Cluj and Szeged matrices,[4-10] respectively. Before defining these matrices, some graph-theoretical background is needed.

Definitions

Let $G = (V, E)$ be a connected graph, with V being the set of vertices and $E \subset V \times V$ the set of edges.

A *walk* w is[11] an alternating string of vertices and edges, $w_{1,n} = (v_1, e_1, v_2, e_2, ..., v_{n-1}, e_m, v_n)$, $v_i \in V(G)$, $e_i \in E(G)$, $m \geq n - 1$, such that any subsequent pair of vertices $(v_{i-1}, v_i) \in E(G)$. Revisiting of vertices and edges is allowed. Then $V(w_{1,n}) = \{v_1, v_2, ..., v_{n-1}, v_n\}$ is the set of vertices of $w_{1,n}$. Similarly, $E(w_{1,n}) = \{e_1, e_2, ..., e_{m-1}, e_m\}$ is the set of edges of $w_{1,n}$. The *length* of a walk, $l(w_{1,n}) = |E(w_{1,n})| \geq |V(w_{1,n})| - 1$, equals to the number of its traversed edges. The walk is *closed* if $v_1 = v_n$ (i.e. its *endpoints* coincide) and is *open* otherwise. The *set of all walks* in G is denoted by $W(G)$.

A *path* p is a walk having all its vertices and edges distinct: $v_i \neq v_j$, $(v_{i-1}, v_i) \neq (v_{j-1}, v_j)$ for any $1 \leq i < j \leq n$. As a consequence, the revisiting of vertices and edges, as well as branching, is prohibited. The *length* of a path is $l(p_{1,n}) = |E(p_{1,n})| = |V(p_{1,n})| - 1$. A closed path is a *cycle* (i.e. *circuit*). The *set of all paths* in G is denoted by $P(G)$.

The *distance*, d_{ij}, between two vertices v_i and v_j is the length of a *shortest* path joining them, if exists: $d_{ij} = \min l(p_{ij})$; otherwise $d_{ij} = \infty$. A shortest path is often called a *geodesic*. The *eccentricity* of a vertex i, ecc_i, is the maximum distance between i and any vertex j of G: $ecc_i = \max d_{ij}$. The *radius* of a graph, $r(G)$, is the minimum eccentricity among all vertices i in G: $r(G) = \min ecc_i = \min \max d_{ij}$. Conversely, the *diameter*, $d(G)$, is the maximum eccentricity in G: $d(G) = \max ecc_i = \max \max d_{ij}$. The *set of all geodesics* (i.e. distances) in G is denoted by $D(G)$.

Cluj Indices

The *Cluj Sets,* $CJ_{i,j,p}$ collect vertices obeying the relation

$$CJ_{i,j,p} = \{v \mid v \in V(G);\ d(G)_{v,i} < d(G)_{v,j};\ \text{and } \exists\, w \in W_{v,i},\ V(w) \cap V(p) = \{i\}\} \qquad (1)$$

where $d(G)$ denotes the topological distance in G and $p \in D(G)$.

$CJ_{i,j,p}$ represent subgraphs (connected or not) in G, related to the endpoint i and referred to j and path p. In *Cluj criterion*, the *path* p (joining the vertices i and j) plays the central role in selecting the sets/fragments. In cycle-containing graphs, more than one path could join the pair (i,j) thus resulting more than one fragment related to i . We define the nondiagonal entries $[\mathbf{UCJ}]_{ij}$ in the Cluj matrices as

$$[\mathbf{UCJ}]_{ij} = \max_{p} \left| CJ_{i,j,p} \right| \quad (2)$$

where $\left| CJ_{i,j,p} \right|$ is the cardinality of the set $\boldsymbol{CJ_{i,j,p}}$. The diagonal entries are zero.

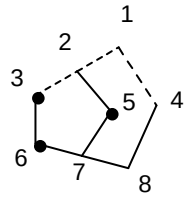
The above definition holds for any connected graph.

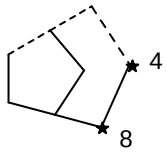
The Cluj matrices are square arrays, of dimension **N**x**N,** usually *unsymmetric* (excepting some symmetric regular graphs). They can be symmetrized, e.g., by the Hadamard product with their transposes

$$\mathbf{SCJ_p = UCJ \bullet (UCJ)^T} \quad (3)$$

$$\mathbf{SCJ_e = SCJ_p \bullet A} \quad (4)$$

The symbol • indicates the Hadamard (pairwise) matrix product[12] ($[\mathbf{M_a \bullet M_b}]_{ij} = [\mathbf{M_a}]_{ij} [\mathbf{M_b}]_{ij}$). For the symmetric matrices, the letter **S** is usually missing. In eq 4, the Hadamard product between the path-defined matrix $\mathbf{SCJ_p}$ and the adjacency matrix **A (**i.e. the matrix having the non-diagonal entries unity for two adjacent vertices and zero otherwise**)** provides the corresponding edge-defined matrix, $\mathbf{SCJ_e}$, which is a weighted adjacency matrix. An example of $\boldsymbol{CJ_{i,j,p}}$ set is illustrated in Figure 1.





$\boldsymbol{CJ}_{3,4,(3,2,1,4)} = \{3, 6, 5\}$
$\boldsymbol{CJ}_{4,3,(4,1,2,3)} = \{4, 8\}$

Cluj Matrix, **UCJ**

0	3	3	4	2	2	2	3	19
5	0	4	4	4	4	3	3	27
3	2	0	**3**	2	3	3	2	18
3	2	**2**	0	2	2	2	3	16
3	3	3	4	0	3	3	3	22
2	3	3	3	2	0	2	3	18
3	3	4	4	4	4	0	5	27
3	2	2	4	2	3	3	0	19
22	18	21	26	18	21	18	22	

$\boldsymbol{CJ_p} = 243$
$\boldsymbol{CJ_e} = 103$

Figure 1. Construction of Cluj matrix, **UCJ**

Cluj indices are calculated from the above matrices, by:

$CJ_e = \frac{1}{2}\Sigma\Sigma\,[\mathbf{M}]_{i,j}\,[\mathbf{A}]_{i,j}$; $CJ_e = \Sigma\Sigma\,[\mathbf{UM}]_{i,j}[\mathbf{UM}]_{j,i}[\mathbf{A}]_{i,j}$ (5)

Szeged Indices

Szeged Fragments, $\boldsymbol{SZ_{i,j}}$ representing the entries in the unsymmetric Szeged matrices, **USZ**, are defined by: [5,7,8]

$$\boldsymbol{SZ_{i,j}} = \{v \mid v \in V(G);\ d(G)_{v,i} < d(G)_{v,j}\} \quad (6)$$

and the corresponding indices are calculated cf. eq 5. In any graph, $CJ_e = SZ_e$ and, in general, $CJ_p \neq SZ_p$.

In the view to account for heteroatoms and multiple bonds in molecular graph, we introduced the ***Szeged property matrices***:[10]

$$[\mathbf{USZP}]_{ij} = \boldsymbol{P_{i,p}} \quad (7)$$

$$\boldsymbol{P_{i,p}} = f(\boldsymbol{P_v}) \mid v \in V(G);\ d(G)_{v,i} < d(G)_{v,j} \mid \quad (8)$$

$$f(\boldsymbol{P_v}) = m \sum_v \boldsymbol{P_v} \quad (9)$$

$$f(\boldsymbol{P_v}) = (\Pi_v \boldsymbol{P_v})^{1/N} \quad (10)$$

The summation and product in eqs 9 and 10 run over all vertices in graph.

Entries in a Szeged property matrix (see eq 7), in fact properties $\boldsymbol{P_{i,p}}$ of vertex ***i*** (with respect to the path ***p***), are defined by a function $f(\boldsymbol{P_v})$, evaluated on vertices ***v*** which obey the Szeged index condition (see eq 6). In other words, the set of such vertices can be viewed as a fragment (i.e., a subgraph). Consequently, $\boldsymbol{P_{i,p}}$ can be viewed as a fragmental property. $\boldsymbol{P_{i,p}}$ is mainly a topological (local) property (e.g., a topological index) but other physico-chemical properties are also considered (e.g., atomic mass or group electronegativities – see below). Two types of $f(\boldsymbol{P_v})$ are here proposed: an additive and a multiplicative one.

Several cases of the ***additive function*** (eq 9) are considered:

(a) $\boldsymbol{P_v}$ = 1 (i.e., the cardinality) and the weighting factor ***m*** = 1 (classical **USZ** matrix).

(b) $\boldsymbol{P_v}$ = some vertex property; ***m*** = 1; (property matrix, **USZP**).

(c) $\boldsymbol{P_v}$ = some vertex property; $\boldsymbol{m} = 1/\boldsymbol{P}(\boldsymbol{G})$; $\boldsymbol{P}(\boldsymbol{G})$ = a global property of the graph;

(d) $\boldsymbol{P_v} = \sum_v \boldsymbol{A_v}$; ***m*** = 1/12; A_v is the atomic mass and the matrix, **USZA** . The factor m indicates that $f(\boldsymbol{P_v})$ is a fragmental mass, relative to the carbon atomic mass.

The ***multiplicative function*** (eq 10) was used for group electronegativities: $\boldsymbol{P_v} = \boldsymbol{X_v}$; (**USZX** matrix). $\boldsymbol{X_v}$ is a local electronegativity, calculated from the Sanderson group electronegativities, for heteroatoms and fragments.[13]

The corresponding ***property indices*** are calculated cf. eq 5.

Correlating Studies

The mathematical models of a certain property are performed by MLR (Multiple Linear Regression) and/or CNN (Computational Neural Networks). In our case, the model is built by using MLR. Next, it is validated by the *leave-one-out* cross-validation procedure. In the following, the MLR procedure is presented.

MLR, for *n* observations and *m* independent variables is represented by

$$Y_i = b_0 + \sum_j^m b_{ij} X_{ij} \quad (11)$$

or, in matrix form as

$$\mathbf{Y} = \mathbf{bX} \quad (12)$$

where **Y** is the ***n*** x 1 vector of responses, **X** is an ***n*** x (***m*** + 1) matrix of independent variables and **b** is the (***m*** + 1) x 1 vector of regression coefficients. The regression coefficients can be determined by the least-squares solution of (12)

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \quad (13)$$

With **b** calculated, eq 12 can be used for estimating the chosen property for other chemical structures.

To avoid the chance correlations, it is recommended that the number of descriptors submitted to regression be less than 60 % of the number of observations in the training set.[14]

Set 1. Poly-Chlorinated Bipheniles.

A set of 15 Poly-Chlorinated Bipheniles (PCB - Table 1) was correlated against the vapor pressure (as log), at 25°C and 100°C, respectively. This property is important in connection with the toxicity of this class of compounds.

Statistics of the correlation are given in Tables 2 and 3.

It can be seen that our descriptors are able to model this property at least as well as the Wiener W, and hyper-Wiener indices.[15] The Szeged index weighted with the electronegativity brings some improvement in correlation. However, it is not a strong evidence of the electronegativity contribution since the set is a congeneric one.

The results of the bivariate regression can be used in prediction studies.

Table 2. Correlation of PCB: $\log VP(25^oC) = a+bx_1+cx_2...$

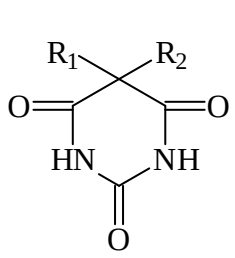
Indices	a	b	c	r	s	F
CJ_e	1.9971	-0.0059	-	0.9649	0.5259	175.82
CJ_p	1.6213	-0.0015	-	0.9704	0.484	209.98
SZ_e	2.0179	-0.0060	-	0.9646	0.528	174.32
W	1.6787	-0.0093	-	0.9630	0.5401	166.03
WW	1.2497	-0.0030	-	0.9703	0.4843	209.67
$CJ_e + SZ_pX$	-8.5702	-0.0258	0.0287	0.9817	0.397	159.67
$CJ_p + SZ_pX$	-5.2096	-0.0030	0.0174	0.9810	0.4044	153.65
$SZ_e + SZ_pX$	-8.4933	-0.0159	0.0287	0.9812	0.4023	155.33

Table 3. Correlation PCB: logVP(100^{o}C) = a+bx_1+cx_2 ...

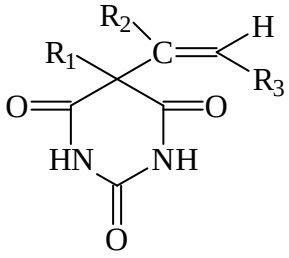
Indices	a	b	c	r	s	F
CJ_e	4.0565	-0.0045	-	0.9753	0.3398	156.22
CJ_p	3.7665	-0.0011	-	0.9833	0.2799	234.07
SZ_e	4.067	-0.0045	-	0.9742	0.3471	149.44
SZ_p	2.4867	-0.0003	-	0.9758	0.3362	159.75
W	3.8154	-0.0071	-	0.9744	0.3457	150.67
WW	3.4661	-0.0023	-	0.9819	0.2915	215.18
CJ_e + CJ_p	2.8123	0.0137	-0.0045	0.9934	0.1876	265.92
CJ_p + SZ_e	2.8591	-0.0042	0.0125	0.9936	0.1853	272.61
CJ_p + SZ_pX	-0.8422	-0.0021	0.0117	0.9930	0.1944	247.45
CJ_p + W	3.5364	-0.0044	0.0207	0.9941	0.1791	291.91

Set 2. Barbiturates.

A set of 25 barbiturates[16] (Table 4) was correlated against the logP. This property is important in connection with the drug membrane transport fenomena.



G_1



G_2

Statistics of data included in Table 4 are given in Table 5. For comparison, the correlation shown by the super-index $EATI_1$ in monovariate regression was: r = 0.9293; s = 0.232.[16] It is clear that the Szeged property index SZ_eA (Table 5, entry 4) is far more appropriate in modeling logP. The plot SZ_eA vs logP is shown in Figure 2.

Table 5. Statistics for Data in Table 4: *log P* = a + $\sum b_i X_i$

No.	**X_i**	**b**	**a**	**r**	**s**	**cv (%)**	**F**
1	SZ_p	0.0003	0.0604	0.9085	0.2627	15.924	108.82
2	SZ_e	0.0045	-1.1072	0.9562	0.1840	11.154	245.66
3	SZ_pA	0.0002	-0.1949	0.9541	0.1883	11.415	233.53
4	**SZ_eA**	**0.0041**	**0.9934**	**0.9935**	**0.0718**	**4.354**	**1740.58**
	Cross Validation		**-**	**0.9913**	**0.0824**	**4.999**	**-**

Table 5 (continued)

5	SZ_pX	0.0136	-0.8170	0.8310	0.3498	21.202	51.34
6	SZ_eX	0.1397	-1.6967	0.7075	0.4445	26.941	23.03
7	SZ_p SZ_e	-0.0005 0.0121	-2.9098	0.9872	0.1023	6.202	423.46
8	SZ_p SZ_eA	0.0000 0.0044	-1.7284	0.9938	0.0717	4.351	871.96
11	SZ_e SZ_pX	0.0080 -0.0131	-0.9247	0.9881	0.0991	6.006	452.38
12	SZ_e SZ_eX	0.0066 -0.1021	0.0401	0.9895	0.0931	5.643	513.80
13	**SZ_pA** **SZ_eA**	**-0.00001** **0.00486**	**-1.8626**	**0.9943**	**0.0685**	**4.157**	**956.23**
	Cross Validation		**-**	**0.9921**	**0.0721**	**4.796**	**-**

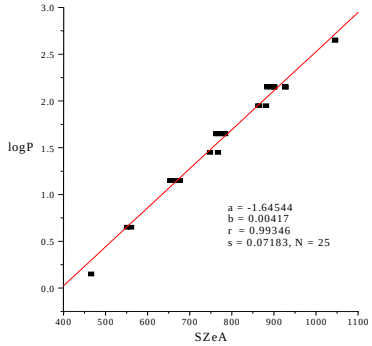


Figure 2. The plot SZ_eA vs logP

On the other hand, the Szeged property indices weighted with electronegativities show lower ability in describing this property (entries 5,6, 11 and 12). No significant improvement was recorded in bivariate regression (entries 7 - 13). The cross-validation test (no significant drop in r-value) indicates the good predicting ability of the equations given in entries 4 and 13.

Concluding, the original descriptors, Cluj and Szeged-property, demonstrated a good ability in modeling some important physico-chemical properties. In the particular case of logP of barbiturates, the recorded results surpass that reported in literature and can be used in predicting studies.

Acknowledgement. This work in under financial support of GRANT CNSIS, T 34, 2000.

REFERENCES

1. Free, S. M.; Wilson, J. W. A mathematical contribution to structure-activity studies, *J. Med. Chem*. **1964**, *7*, 395.
2. Gao, C.; Govind, R.; Tabak, H. H. Application of the group contribution method for predicting the toxicity of organic chemicals. *Environmental Toxicol. Chem.* **1992**, *11*, 631-636.
3. Diudea, M., Ed., QSPR/QSAR studies by molecular descriptors, NOVA SCIENCE Publishers, Inc., Huntington, N. Y., 2000
4. Diudea, M.V. Cluj matrix $\mathbf{CJ_u}$: source of various graph descriptors, *Commun. Math. Comput. Chem. (MATCH),* **1997**, *35*, 169-183.
5. Diudea, M.V.; Minailiuc, O.; Katona, G.; Gutman, I. Szeged matrices and related numbers, *Commun. Math. Comput. Chem. (MATCH),* **1997**, *35*, 129-143.
6. Diudea, M.V. Cluj matrix invariants, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 300-305.
7. Diudea, M.V.; Pârv, B.; Topan, M.I. Derived Szeged and Cluj indices, *J. Serb. Chem. Soc*. **1997**, 62, 267-276.
8. Kiss, A.A.; Katona, G.; Diudea, M.V. Szeged and Cluj matrices within the matrix operator $W_{(M1,M2,M3)}$ *Coll. Sci. Papers Fac. Sci. Kragujevac* **1997,** *19,* 95-107.
9. Gutman, I.; Diudea, M.V. Defining Cluj matrices and Cluj matrix invariants, *J. Serb. Chem. Soc.* **1998**, *63*, 497-504.
10. Minailiuc, O.; Katona, G.; Diudea, M.V.; Strunje, M., Graovac, A.; Gutman, I. Szeged fragmental indices, *Croat. Chem. Acta* **1998**, *71*, 473-488.
11. Jäntschi, L.; Katona, G.; Diudea, M. V. Modeling molecular properties by Cluj indices, *Commun. Math. Comput. Chem. (MATCH),* **2000**, *41*, 151-188.
12. Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge Univ. Press, Cambridge, **1985**.
13. Diudea M.V.; Kacso I.E.; Topan M.I. Molecular topology. 18. A Qspr/Qsar study by using new valence group carbon-related electronegativities. *Rev. Roum. Chim.* **1996**, *41*, 141-157.
14. Topliss, J. G.; Edwards, R. P. Chance factors in in studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238.
15. Diudea, M.V.; Gutman, I. Wiener-type tpological indices, *Croat. Chem. Acta* **1998**, *71*, 21-51.
16. 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. **1997**, *35*, 185-197.