Bulgarian Chemical Communications, Volume 42, Number 4 (pp.338 –342) 2010

Charge-related molecular index (CMI), a novel descriptor for quantitative structure/property relationship (QSPR) models. I. General considerations.

I. Bangov*, M. Moskovkina, A. Patleeva

'Episkop Konstantin Preslavski' University of Shumen, Department of Nature Sciences, 115 University St, Shumen 9712, Bulgaria

Received April 7, 20101; Revised June 7, 2010

The charge-related molecular index (*CMI*), developed by one of the authors (Bangov), and its use in chemoinformatics is discussed. A comparison is carried out between the *C*harge-related *T*opological *I*ndex (*CTI*) values and the *C*harge-related *G*eometrical *I*ndex (*CGI*) values, generated by using different quantum chemistry methods. It is shown that both indices can be successfully employed for isomorphic structure perception and for correlations with structure branching.

Key words: QSPR modelling; charge-related molecular index (CMI); charge-related topological index (*CTI*); charge-related geometrical index (*CGI*).

INTRODUCTION

A charge-related molecular index (*CMI*) was developed by one of us [1,2] in 1989. It has the following form:

$$CTI = \sum_i \sum_j \frac{L_i L_j}{D_{ij}} \qquad (1)$$

Here D_{ij} is the inter-atomic distances and L_i is the local indices, characterizing the individual heavy (non-hydrogen) atoms *i,* expressed as follows:

$$L_i = L_o - N_H + q_i \qquad (2)$$

L_o is the constant values for each atom for each hybridization state (they can be in some cases atom valences); N_H is the number of the hydrogen atoms, attached to a given heavy atom, and q_i is the corresponding charge densities. These are computed by either the topological empirical method of Gasteiger-Marsili [3] or by any of the sophisticated quantum chemistry methods on semi-empirical or non-empirical level. We used the Gasteiger-Marsili method for calculation of the atomic charges and the topological distance matrix inter-atomic distances when we take into consideration the 2D topology of the molecule. In this case we employ a *C*harge-related *T*opological *I*ndex (*CTI*). Vice versa, in case of using 3D molecular models, the distances and charge densities are calculated out from the atom coordinates, and in this case the index is no more topological. Further, we shall call it *C*harge-related *G*eometrical *I*ndex (*CGI*).

The L_o values with the corresponding hybridization states and valences for some elements are presented in **Table 1**.

Table 1. Valences, atom codes, and *Lo* initial values for some elements and their hybridization states.

Element and hybridization state	Valence	Code	L_o
C			
sp^3	4	C	4
sp^2 (olefinic)	3	=C	11
sp^2 (aromatic)	3	:C	13
sp	2	#C	7
N			
sp^3	3	N	15
sp^2	2	=N	18
sp	1	#N	20
O			
sp^3	2	O	23
sp^2	1	=O	25
S	2	S	28
F	1	F	32
Cl	1	Cl	33
Br	1	Br	34
I	1	I	35

Initially, the *CTI* was developed for perception of isomorphic (equivalent) complete molecular structures and substructures (fragments) in the process of 2D structure generation.[1] Although, it cannot be strictly mathematically proved, our practice shows that this index manifests an extremely good discriminating power and appears

* To whom all correspondence should be sent:
E-mail: ivanbangov@shu-bg.net

to be practically an index of no degeneracy. Thus, equivalent (isomorphic) structures produce the same *CTI* values (within the computer word precision), and different (non equivalent) structures, i.e. different values. Hence, the reason that we use the L_o values, is to achieve better discrimination (clustering) between the *L* values of the different atom types and their hybridization states within the chemical structure whereas no overlapping of their *L* values is produced and better perception of the isomorphism and automorphism is achieved.

We found that this index can be further developed to feature 3D chemical structures. This can be achieved by using the electron charge densities in the L_i values and the real inter-atomic distances R_{ij} rather than the topological distance matrix D_{ij} values in **Eqn. 1,** both generated by using quantum chemistry methods either on semi- or non-empirical level. Further on we shall call the latter index *C*harge-related *G*eometrical *I*ndex (*CGI*). Accordingly, the purpose of this and the following papers is to study the use of the *CGI* in the various chemoinformatics areas.

As seen from the relations (1) and (3), both *CTI* and *CGI* consists of two parts, numerator and denominator. Whereas the denominator accounts for the branching of the chemical structure in a way similar to that of the *Wiener* index [4], the numerator features the atomic type differences and their polarity. Furthermore, the charges (especially these produced from the Gasteiger - Marsili method using iteratively different atom environments) experience the influence of the electron density of the whole molecule on each separate atom. In this respect they algorithmically resemble the usual hashing procedures, used in chemistry. Accordingly, this combination of the two parts makes the index both very discriminative and a good descriptor for different types of quantitative structure/property relationships (QSPR), especially for the cases of interactions of polar moieties and media. Since it has the form of an electrostatic potential we can call it *molecular potential*.

EXPERIMENTAL.

Some of the *CTI* values are taken from a previous paper, written by Mr. Bangov [2], while other values are calculated together with the *CGI* values by using the *QSPR.exe* program which was written in Java by Bangov. The charges and the inter-atomic distances for the *CGI* calculations are generated by using quantum chemistry methods on a semi-empirical and non-empirical level from the GAMESS package (USA). Thus, the *CGI* values are compared with the corresponding *CTI* values.

RESULTS AND DISCUSSIONS.

3.1. CTI *Versu*s CGI *Results.*

The discriminating potential of our indices in comparison with the Wiener [4] and Randic [5] indices is illustrated on Fig. 1 for two couples of isomers. One can see that both the Wiener index in case *a* and the Randic index in case *b* produce degenerate (the same) values. In contrast, our *CTI* and *CGI* indices successfully discriminate the isomers.

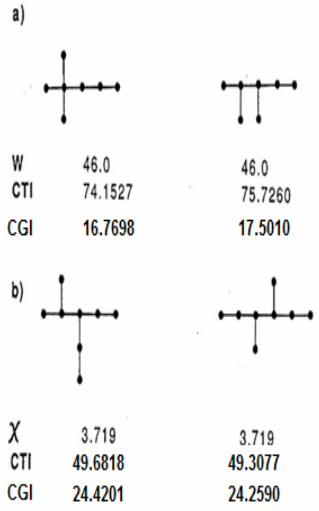

Fig. 1. a) Two isomers of different connectivity, providing the same Wiener index values W but different *CTI* and *CGI* values; **b)** two isomers of different connectivity, providing the same Randic index values χ but different *CTI* and *CGI* values.

It must be noticed here that the use of the *CGI* for perception of isomorphic (duplicated) structures depends on the precision of the SCF-procedure consistency and the conditions of the quantum chemistry calculations (parameterization, the basis set, geometry optimization, *etc*). The *CTI,* carried out with a fixed number of iterations of the Gasteiger-Marsili procedure (6 iterations in our case), is more suitable to this end on the one hand. On the other hand, it is much faster, hence it can be applied to large 2D structure datasets.

The L_i values were also used for perception of the constitutional molecular structure symmetry (automorphism). Thus, symmetric atoms have the same L_i values. [6]

First, it should be mentioned that both the *CTI* and the *CGI,* as well as the Wiener index, depend on the molecule size. This is illustrated in **Table 2** with the *CGI* results of the first 8 alkanes, generated from quantum chemical methods on different levels (the semi-empirical *AM1*, and *ab initio* calculations with **3**–**21G** and **6**–**31G** basis sets).

Table 2. *CGI* values for the first 8 alkanes, ethene, propene, butane, and butadiene.

Compound	Wiener	CTI	CGI (AM1)	CGI (3-21G)	CGI (6-31G)
Ethane	1.0	1.0940	0.4141	0.1032	0.18006
Propane	6.0	5.5174	3.4466	2.8160	1.2720
Butane	16.0	12.4114	6.8630	5.1819	3.8126
Pentane	20.0	21.0965	9.8116	6.0120	8.5504
Hexane	35.0	30.9838	19.4892	16.7350	11.3400
Heptane	56.0	41.9989	20.4546	13.5525	16.1190
Octane	84.0	53.7756	26.5340	17.9132	20.8957
Ethene	1.0	23.0940	17.2461	16.0280	16.3992
Propene	6.0	27.5174	25.6882	22.1318	23.3331
1-Butene	16.0	34.4114	35.6854	30.5694	32.4841
Butadiene	16.0	56.4114	94.9199	89.7188	91.6864

Clearly, the *CGI* takes into account the chemical structure diversity, i.e., both different types of the atoms and the bonds, while the *Wiener* index does not distinguish them, hence structures such as butane, 1-butene, butadiene produce the same values. The *CGI* values of four compounds which have double bonds: ethene, propene, 1-butene and butadiene are also presented in **Table 2**. Naturally, their values are larger because of the L_o values of the double bonds, as given in **Table 1** (compare the *CGI* values for butane, 1-butene and butadiene there).

CGI values of different alcohols are provided in **Table 3** at two *ab initio* basis sets (**3–21G** and **6-31G**), and are compared with the *CTI* values from our previous paper [2]

Table 3. Charge-related topological index (*CTI),* calculated for a series of alcohols, and compared with the charge-related geometrical index (*CGI*), calculated by *ab initio* of different basis sets.

No	Compound	CTI	CGI (3-21G basis set)	CGI (6-31G basis set)
1.	Ethanol	56.3998	32.2479	35.8291
2.	Propane-3-ol	78.9698	47.4938	51.8761
3.	Propane-2-ol	92.9637	53.9825	59.1265
4.	Butane-4-ol	97.7223	59.0230	64.1728
5.	2-methylpropane-3-ol	103.9319	64.25757	69.2459
6.	Butane-3-ol	117.8204	71.2747	77.3128
7.	2-methylpropane-2-ol	132.0219	76.6028	83.0625
8.	Pentane-5-ol	114.7124	69.0861	74.9591
9.	2-methylbutane-4-ol	118.8611	71.5210	77.3650
10.	2-methylbutane-1-ol	124.9558	76.9455	82.8386

Generally speaking, the results from the computing methods follow the same trend in the case of alcohols. Again, one can see that the *CTI*s give always much larger values than the *ab initio* results on both the **3**–**21G** and the **6**–**31G** basis set levels. The numerical results from the sophisticated **6-31G** basis set calculations are somewhat larger than these from the **3**–**21G** basis set. Several factors influence the *CGI* values: the first, as mentioned above, is the size of the molecule; the second is the position of the heteroatom; and the third is the branching of the structure. Actually, many molecular properties depend on the chemical structure branching.

In order to study the influence of the structure branching on the *CMIs,* we compare the *CTI* values with both the *Wiener* index and the *CGI* values, calculated via quantum chemistry methods on *semi-empirical* (**AM1** and **PM3**) and *non-empirical* levels (**6–31G** basis set). These were calculated for a series of isomers of the octane hydrocarbon (C_8H_{18}), having different branching together with their ranking reported by Bertz [7]. By inspecting the different isomers of the same size from **Table 4,** we can figure out that in general the most branched isomers produce higher *CMI* values both for the 2D (*CTI*) and 3D (*CGI*) cases. As expected the *Wiener* index produces values which decrease with the increase of the chemical structure branching.

Although the notion of branching has no strict definition, one can see that in some cases the *CMIs* provide much more reasonable results than the ranking of Bertz. One can see that the Wiener index, the *CTIs* of the 2D structures, and the *CGIs* of the 3D structures describe pretty well on *semi-empirical* level the branching of different types of chemical structures, while the *non-empirical* level produce some discrepancies. Thus, unlike the *Wiener* index, and the *CTIs* and *CGI*s on *semi-empirical* level, the most branched structure, the 2,2,3,3-tetramethylbutane, does not produce the highest *CGI* value on the *non-empirical* level, as expected. On the other hand the *Wiener index* has a serious disadvantage, it frequently provides degenerate (the same) values for different structures, as in the case of structures 3,3-dimethylhexane and 2-methyl-3-ethylpentane. They obviously have not only a different constitution but a different branching, too.

The results indicate that both 2D *CTI*s and 3D *CGIs* on semi-empirical level are more discriminative concerning branching than the *CGIs* on *non-empirical* level. The difference between the topological and the geometrical index is certainly in the charge densities which are influenced by the molecular geometry in the case of quantum chemistry calculations but the main contribution to this difference comes from the inter-atomic distances which are used explicitly. Whereas in the case of the *CTI* these distances account for the number of bonds between the non-bonded atoms, the distances between the non-bonded atoms are through space in the case of the *CGI*. Hence, the *CTI* always prefers the most folded structure, having the shortest topological distances (the number of bonds between different atoms). The latter appears to be the most branched. In the same time this is not the case with the *CGI* on *non-empirical* level where the distances between the atoms are estimated according to the molecular geometry. Obviously, the semi-empirical calculations produce some borderline results. This gives us a good opportunity to favor the much faster *empirical* and *semi-empirical* approaches.

Table 4. Ranking of octane isomers, compared with the *Wiener*, *CTI*, *CGI* indices (calculated by different quantum chemistry methods on semi- and non-empirical levels). The ranking of the different isomers, according to the branching index of Bertzc [7] is given in the brackets.

No	Compound	Wiener index	CTI	CGI (AM1)	CGA (PM3)	CGI (6-31G)
1	octane (1)	84	41.9796	26.5323	28.6211	20.8957
2	2-methylheptane (2)	79	43.6962	27.5524	28.7678	21.3851
3	3-methylheptane (4)	76	45.1788	28.5444	30.7575	22.2655
4	2,5-dimethylhexane (3)	74	45.4882	28.6227	30.9619	21.9044
5	4-methylheptha ne (5)	75	45.5535	28.8463	31.0634	22.5394
6	2,2dimethylhexane (12)	71	46.8757	29.2557	31.6989	22.1572
7	3-ethylhexane (7)	72	47.0352	30.1201	32.3656	23.5653
8	2,4-dimethylhexane (6)	71	47.1357	29.8083	32.1435	22.8984
9	2,3-dimethylhexane (8)	70	48.0352	30.2540	32.6000	23.3551
10	2,2,4-thrimethyl pentane (13)	66	48.9985	30.6053	33.1529	21.9044
11	3,4-dimethylhexane (9)	68	49.3077	31.0533	33.3759	24.0659
12	3,3-dimethylhexane (14)	67	49.4228	30.8739	33.2964	23.1588
13	2-methyl-3-ethylpentane (10)	67	49.6818	31.5014	33.8264	23.5878
14	2,3,4trimethyl pentane (11)	65	50.6825	31.8000	34.2331	24.4201
15	3-ethyl,3-methyl pentane (15)	64	51.5943	32.4283	34.8512	24.2955
16	2,2,3-trimethyl pentane (16)	63	51.6971	32.1138	34.6474	24.2326
17	2,3,3-trimethyl pentane (17)	62	52.5966	32.6185	35.1300	24.6297
18	2,2,3,3tetramethylbutane (18)	58	54.6152	33.1903	35.8574	24.4034

On the other hand the *CGI* can be used to distinguish between different conformations as shown on **Fig. 2.**

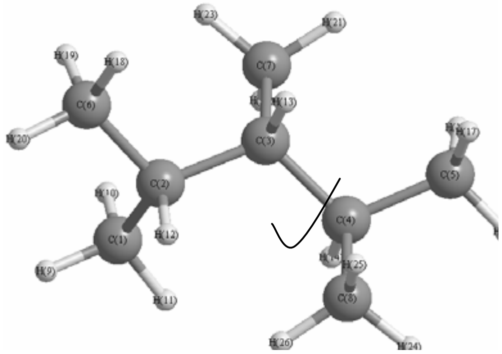

Fig. 2. CGI, total energy values of the lowest 2,3,4-trimethylpentane conformers (rotations at dihedral angle ω(bonds 2–3–4–5)) ***a***. ω = 163.116, E_{total}= -196677.684410672 kcal, *CGI*= 24.29638; ***b***. ω = 63.000 E_{total} = -196676.354433968 kcal, *CGI* = 24.24072; ***c***. ω = -72.700, E_{total} = -196676.472099028 kcal, *CGI*= 24.316437; **d.** ω = 96.400, E_{total} = -196676.354482224 kcal, *CGI*=24.24065

.

In the previous paper [2] we found a very good correlation between the *CTI* values and the enthalpies of formation (R = 0.968) of C_2–C_3 alkanes and with the octane numbers of some fuels (R=0.986). On the other hand,it was shown that the *CTI* can be used for description of the different environments of the structure atoms, forming

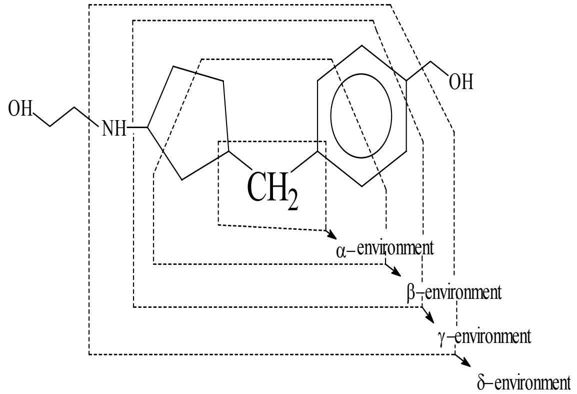

Fig. 3. The picture of the *α*-, β-, γ-, and *δ*-environments of a CH_2 carbon atom.

different fragments, hence a procedure similar to that of Bremser [8] for calculation of ^{13}C chemical shifts was reported by one of the authors (Bangov [9]). The method of Bremser was modified in such a way that instead of using Bremeser's HOSE code we used the *CTI*. As shown on **Fig. 3,** each carbon atom within a given structure is associated with α, *β*, *γ,* and *δ* environments.

These environments are fragments practically, and we assign a *CTI* value to each of them: *α* – (CTI_{α}), *β*(CTI_{β}), γ– (CTI_{γ}), and *δ*– ($CTI_{\delta\delta}$), and relate these environments (fragments) to their ^{13}C chemical shifts and coupling constants. We can use them either in a relational base or we can form fingerprints similar to these of Daylight [10], and assign to each key in the fingerprint both fragment and spectral characteristics.

CONCLUSIONS

It becomes obvious from our study of the *CTI*s and *CGI*s that they describe not only the connectivity within chemical structures (as the other most popular indices do) but also their multiple bond, heteroatom and polarity density owing to use of charges in their formation. As far as they have the form of an electrostatic potential (1), we can consider them as *molecular potential*.

Since our indices describe well the constitution, the diversity and the structural branching, they have the further potential for deriving efficient *QSAR* and *QSPR* models with *CTIs* and *CGIs),* derived from both 2D and 3D molecular structures. This might be especially useful for the cases of polar phases. Studies on their usage as basic descriptors for chromatography retention modelling have been carried out [11] and some novel results will be reported in the near future.

REFERENCES

1. I. Bangov, *J. Chem. Inf. Comput. Sci.*, **30,** 277 (1990).
2. P. A. Demirev, A. S. Dyulgerov, I. P. Bangov, *J. Math. Chem.,* **8,** 367 (1991).
3. J. Gasteiger, M. Marsili, *Tetrahedron*, **36,** 3219, (1980).
4. H. Wiener, *J. Am. Chem. Soc*., **69**(1) 17, (1947).
5. M. Randic, *J. Amer. Chem. Soc.* **97**, 6609, (1975).
6. I. Bangov, *J. Chem. Inf. Comput. Sci.,* **34,** 318, (1994).
7. St. Bertz, *Discr. Appl. Math.* **19,** 65 (1988).
8. W. Bremser, *Anal. Chim. Acta*, **103,** 355 (1978).
9. I. Bangov, *Annuaire Uni. Sofia,* **2001**, *91*, 103-113.
10. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html
11. M. Moskovkina, *Thesis*, Shumen Univ.Library, 118 (2004).

СВЪРЗАНИЯТ СЪС ЗАРЯДИТЕ МОЛЕКУЛЕН ИНДЕКС(CMI) – ЕДИН НОВ ДЕСКРИПТОР ЗА МОДЕЛИ НА СТРУКТУРА/СВОЙСТВА КОЛИЧЕСТВЕНИТЕ СЪОТНОШЕНИЯ (QSPR). I. ОБЩИ РАЗГЛЕЖДАНИЯ.

И.Бангов*, М. Московкина, А.Патлеева

Шуменски университет „Епископ Константин Преславски”, Факултет по природни науки“, ул. Университетска. 115, Шумен , 9712,България,

Постъпила на 7 април, 2010 г.; преработена на 7 юни, 2010 г.

Резюме

Дискутирани са свързаният със зарядите молекулен индекс (*CMI*), разработен от един от авторите (Бангов) и неговото използване в хемоинформатиката. Проведено е едно сравнение между стойностите на свързания със зарядите топологичен индекс (*CTI*) и на свързаният със зарядите геометричен индекс (*CGI*) получени при използването на различни квантово-химични методи. Показано е, че и двата индекса могат успешно да бъдат използвани, както за откриването на изоморфни структури, така за корелация със разклоняването на структурите.