J Mol Model (2001) 7:1–5
DOI 10.1007/s008940100005

ORIGINAL PAPER

Andrés Mercader · Eduardo A. Castro
Andrey A. Toropov

Calculation of total molecular electronic energies from Correlation Weighting of Local Graph Invariants

Received: 13 October 1999 / Accepted: 24 January 2001 / Published online: 8 March 2001

Abstract Using Correlation Weighting of Local Graph Invariants is considered for molecular descriptors for computing total molecular electronic energies. Instead of using prescribed weights for paths and vertices, we have optimized such weights so that the standard error in the regression analysis is as small as possible. Results compare favorably with respect to the employment of other common topological descriptors.

Keywords Electronic energy · Graph invariants · Theory QSAR/QSPR · Topological descriptors

Introduction

The employment of graph-theoretical structure-descriptors represents a meaningful step forward in the search for suitable predictive models in chemistry, biology and pharmacology and remains within the bounds of the philosophy of the increasing use of mathematical and computational methods in contemporary science.

The basis for these models in chemistry is the study of the **Q**uantitative **S**tructure-**P**roperty and **S**tructure-**A**ctivity **R**elationships (QSPR and QSAR, respectively), in which the structural information is encoded into numbers obtained from graph-theoretical invariants [1, 2, 3] (*i.e.* they do not depend on the vertex numbering).

Different graph characteristics or invariants have been used in the definition of molecular topological indices. Because the pool of molecular descriptors has increased dramatically during the last decade, the problem of selecting them is a current topic of interest to many researchers [4, 5]. A remarkable improvement is found in the regression results when they are based on the use of the path numbers as descriptors instead of the connectivity indices. The standard error is usually also improved (i.e. it decreases) when several suitable path numbers are applied [4].

The total molecular electronic energy is the most important molecular parameter from which one can derive the great majority of the microscopic properties necessary to characterize molecules. The usual way to compute this quantity for a given molecular arrangement is to resort to the standard methods of electronic structure theory [6]. But, strange to say, there has not be any study aimed at mimicking this relevant feature via QSPR theory, although Bonchev and Kier calculated electronic charges in alkanes through topological indices [7]. Needless to say, it would be really valuable to have at one's disposal a straightforward and economical method of computing total electronic molecular energies with good accuracy, avoiding the sometimes troublesome first-principles methods.

Thus, we have deemed it appropriate to present a simple procedure to calculate accurate total molecular electronic energies within the realm of QSPR theory resorting to **C**orrelation **W**eighting of **L**ocal **G**raph **I**nvariants (CWLGI).

The paper is organized as follows: the next section deals with the definition of the chosen descriptors, giving their foundations and pointing out their usefulness as well as discussing the antecedents for the particular form of the index. Then we show some illustrative numerical results derived from several common regression equations and compare them with exact data. Finally, we analyze the possibility of extending the employment of this sort of weighted path descriptor to study other physical-chemistry properties and biological activities and to apply it together with others complementing graph theoretical indices in order to reach an optimum description in QSPR/QSAR.

A. Mercader · E.A. Castro (✉)
CEQUINOR, Departamento de Química,
Facultad de Ciencias Exactas, UNLP, C.C. 962,
La Plata 1900, Argentina
e-mail: castro@dalton.quimica.unlp.edu.ar

A.A. Toropov
Vostok Innovation Company, S. Azimstreet 4,
700047 Taskent, Uzbekistan

Correlation weighting of local graph invariants

Although graph theory is now an integral branch of combinatorial analysis, it began as a part of topology and even today, most electrical engineers and many chemists working in network theory still consider "topology" to be entirely synonymous with "graph theory".

Graph theory was independently discovered on several occasions and three names deserve special mention: Euler, Kirchhoff, and Cayley [8]. Two kinds of correspondence between graphs and chemical categories have found numerous applications in chemistry:

a) A graph corresponds to a molecule, i.e. points symbolize atoms and lines symbolize chemical covalent bonds. These may be called **structural or constitutional graphs**.
b) A graph corresponds to a reaction mixture, i.e. points symbolize chemical species and lines symbolize conversions between these species. These may be called **reaction graphs**.

The former type of graph gave Cayley the incentive to derive a procedure for counting the constitutional isomers of alkanes and later it led Polya towards the discovery of his powerful counting theorem [9]. Thus, chemistry is an acknowledged origin for the beginning and development of graph theory, and the mathematicians Cayley and Polya published papers in chemical journals.

With chemistry as one of its breeding grounds, graph theory is well adapted for solving chemical problems, both from the high degree of abstraction evidenced by the generality of such concepts as points, lines, and neighbors, as well as by the combinatorial derivation of many graph-theoretical concepts, which correspond to the essence of chemistry viewed as the study of combinations between atoms.

Topological indices are numerical quantities derived from molecular graphs representing molecules. Sometimes weighted graphs, multigraphs, or weighted pseudographs are used to represent the relevant aspects of the chemical species [10].

Here we describe the CWLGI, which has the general form

$$DCW=\Sigma_{\text{all vertices}} [CW(a(i))+CW(VD(i))] \quad (1)$$

where Correlation Weights (CW) for a given element (a(i)) and for the corresponding Vertex Degree (VD) are determined by means of an optimization procedure to reproduce a physical-chemistry property. That is to say, CW(a(i)) and CW(VD(i)) are computed in such a way that they give the largest possible correlation coefficient between the numerical value of the property under consideration and the descriptor's value (DCW).

Obviously, the procedure demands the employment of a training set without any information at all about the structures of the test set. It must be pointed out that the usual additive contributions are found on the basis of the minimization of standard errors. Additive schema are based on computing the property P via a relationship such as

$$P=\Sigma_{\text{fragments}} AC(i) \quad (2)$$

where AC(i) is the additive contribution of the i-th fragment to the P value of the property. The AC(i) are usually found by means of an optimization procedure aiming to give as small as possible values of

$$s=\{\Sigma [P_{\text{experimental}}-P_{\text{calculated}}]^2\}^{1/2} \quad (3)$$

However, CW's are calculated in a quite different manner. In fact, they are computed in such a way as to maximize the correlation coefficient in Eq. (1). After determining the optimum CW's, one can calculate the desired property P via a general formula

$$P=f(DCW) \quad (4)$$

The most usual way to choose the function f(DCW) is to resort to a polynomial form, *i.e.*

$$P=a+b\ DCW+c\ (DCW)^{2+}d\ (DCW)^{3}+..... \quad (5)$$

where a, b, c, d,....are real numbers and are calculated through a standard fitting procedure.

Results

The methodology described in the previous section for calculating physical-chemistry properties from DCW has been applied before [11, 12, 13, 14] and results shown to be rather satisfactory. Here, we have chosen a set of 49 organic molecules previously analyzed to compute their enthalpies of formation from *ab initio* total electronic energies at the 6-31G* basis set level [15]. This rather modest set includes molecules composed of H, C, N, O, F, and Cl atoms and we have divided the whole group into two subsets of 25 and 24 molecules each (a training set, and a test set, respectively). We have not applied any special criteria for this particular choice, save that of obtaining two equilibrated subsets regarding the number and kind of atoms in each.

In Tables 1 and 2, we display the complete set of 49 molecules, denoting those included in the training set and those pertaining to the test set, respectively, together with the theoretical total electronic energies (TEE). The CW's for total electronic energies corresponding to the different atoms and vertex degrees are presented in Tables 3 and 4.

The linear correlation equation between TEE and DCW for the training set is

$$TEE=28.433\ DCW(TEE)-0.0629$$

$$n=25,\ s=0.02226\ a.u.,\ r=0.999999976,\ F=471629359 \quad (6)$$

while the statistical results of applying Eq. (4) to the test set are

$$n=24,\ s=0.03225\ a.u.,\ r=0.999999985,\ F=758011111$$

These results are not dependent on the particular selection of the molecules included in the two sets, since different choices yield practically the same statistical data.

Table 1 Theoretical *ab initio* total electronic energies (atomic units) calculated at the 6-31G* basis set level for the training set

Number	Molecule	Energy (a.u.)[a]
1	Methane	40.19517
2	Acetylene	76.81783
3	Allene	115.86110
4	Propene	117.07147
5	Dibutyne	152.49793
6	1-Butene	156.10499
7	Isobutene	156.11067
8	Isobutane	157.19896
9	Hydrogen cyanide	92.87520
10	Methylhydrazine	150.20108
11	Acetonitrile	131.92753
12	Cyanogen	184.59122
13	Formaldehyde	113.86633
14	Methanol	115.03542
15	Formic acid	188.76231
16	Dimethyl ether	154.06574
17	Urea	223.98219
18	Methylnitrite	243.66864
19	Difluoromethane	237.89635
20	Trifluoromethane	336.77164
21	Fluoroethane	178.07722
22	Vinylchloride	536.93369
23	Cyclopentane	195.16295
24	Bicyclo[1,1,0]butane	195.16295
25	Cubane	307.39383

[a] Taken from ref. [15]

Table 2 Theoretical *ab initio* total electronic energies (atomic units) calculated at the 6-31G* basis set level for the test set

Number	Molecule	Energy (atomic units)[a]
1	Ethylene	78.03172
2	Ethane	79.22785
3	Propyne	115.86432
4	Propane	118.26365
5	1,3-Butadiene	154.91960
6	2-Butyne	154.90925
7	1,4-Pentadiene	193.94093
8	Methylamine	95.20983
9	Ethylamine	134.24761
10	Dimethylamine	134.23885
11	Ketene	151.72467
12	Acetaldehyde	152.91569
13	Ethanol	154.07574
14	Glyoxal	226.59218
15	Acetone	191.96225
16	Fluoromethane	139.03461
17	Fluoroethene	176.88195
18	1,1-Difluoroethene	275.74000
19	Tetrafluoroethene	473.41567
20	1,1-Dicholoethane	997.03094
21	Cyclobutane	156.09720
22	Cyclopendiene	192.79192
23	Cyclopentene	193.97717
24	Cyclohexane	234.20796

[a] Taken from ref. [15]

We have also tried higher order fitting polynomials and numerical results are also nearly invariant, so that we have considered it unnecessary to report them here. Those readers interested in obtaining complete regression equations can request there from the corresponding author.

Table 3 Correlation Weights for elements on the total electronic energies

Atom	CW(a(i))
H	–0.0532
N	1.8420
F	3.4237
C	1.2592
O	2.5601
Cl	16.0860

Table 4 Correlation Weights for vertes degrees on the total electronic energy

Vi (i=1,2,3,4)	CW(VD(i))
V1	0.0745
V2	0.0715
V3	0.0716
V4	0.0712

Table 5 Calculated TEE(a.u.) through Eq. (6) for the training molecular set in Table 1

Molecule Number	DCW	–TEE	ΔTEE (TEE_{calc}–$TEE_{ab\ initio}$)
1	1.41560	40.18685	0.00832
2	2.70400	76.81993	–0.00210
3	4.07750	115.87266	–0.01156
4	4.11980	117.07537	–0.00390
5	5.36540	152.49152	0.00641
6	5.49280	156.11388	–0.00889
7	5.49280	156.11388	–0.00321
8	5.53460	157.30238	–0.00342
9	3.26850	92.87036	0.00484
10	5.28540	150.21688	–0.01580
11	4.64150	131.90887	0.01866
12	6.49440	184.59238	–0.00116
13	4.00800	113.89656	–0.03023
14	4.04720	115.01114	0.02428
15	6.63960	188.72085	0.04146
16	5.42020	154.04965	0.01609
17	7.87780	223.92659	0.05560
18	8.57400	243.72164	–0.05300
19	8.36940	237.90425	–0.00790
20	11.84630	336.76295	0.00869
21	6.26550	178.08406	–0.00684
22	18.88600	536.92274	0.01095
23	6.86500	195.12964	0.03331
24	5.44940	154.87989	–0.00820
25	10.81360	307.40019	–0.00636

Average absolute error=0.0156 a.u.

In Tables 5 and 6, we display the complete results for the prediction of TEE together with the DCW‘s. The average deviation for the training set is 0.0156 a.u. while this parameter for the test set is 0.0183 a.u. The comparison between calculated and predicted TEE for both sets is quite good and there is no “pathological behavior” among these molecules, which seems to show the quite satisfactory predictive capability of the present approach.

Table 6 Calculated TEE (a.u.) through Eq. (6) for the test molecular set in Table 2

Molecule Number	DCW	–TEE	ΔTEE (TEE_{calc}–$TEE_{ab\ initio}$)
1	2.74680	78.03686	–0.00514
2	2.78860	79.22536	0.00249
3	4.07700	115.85844	0.00588
4	4.16160	118.26387	–0.00022
5	5.45100	154.92538	–0.00578
6	5.45000	154.89695	0.01230
7	6.82400	193.96389	–0.02296
8	3.35050	95.20187	0.00796
9	4.72350	134.24038	0.00723
10	4.72350	134.24038	–0.00153
11	5.33870	151.73236	–0.00769
12	5.38100	152.93507	–0.01938
13	5.42020	154.04965	0.02609
14	7.97340	226.64478	–0.05260
15	6.75400	191.97358	–0.01133
16	4.89250	139.04555	–0.01094
17	6.22370	176.89556	–0.01361
18	9.70060	275.75426	–0.01426
19	16.65440	473.47166	–0.05599
20	35.06700	996.99711	0.03383
21	5.49200	156.09114	0.00606
22	6.78140	192.75265	0.03927
23	6.82320	193.94115	0.03602
24	8.23800	234.16815	0.03981

Average absolute error=0.01827 a.u.

Discussion

The numerical data given in the preceding section show the high-quality results based on CWLGI, which on one hand yield very accurate total molecular electronic energies and, on the other hand, give a correlation equation with significantly low standard error. These findings open the possibility of extending this sort of study for other physical-chemistry properties as well as biological activities using this new kind of topological descriptor.

It would be also interesting to employ multiple regression analysis based on suitable graph descriptors combined with the orthogonalization procedure in order to reach optimum structure-property-activity relationships which will surely lead to a meaningful interpretation of the results. This feature is currently missing from QSAR/QSPR studies. Work along these lines is presently being carried out in our laboratories and results will be published in the near future.

A final comment on the analytical formula (1) to compute DCW deserves to be made here. In fact, we have employed an additive relationship between CW(a(i)) and CW(VD(i)), but it should be equally valid to resort to another sort of connection between the CWLGI, for example:

$$DCW=\Sigma_{\text{all edges}}\,[CW(a(i))\ *\ CW(VD(i))\ *\ CW(a(j))\ *\ CW(VD(j))] \quad (7)$$

where (i,j) is an edge

$$DCW=\Sigma_{\text{all vertices}}\,[CW(a(i))\ *\ CW(VD(i))] \quad (8)$$

$$DCW=\Sigma_{\text{all vertices}}\,[CW(a(i))+CW(VD(i))] \quad (9)$$

$$DCW=\Pi_{\text{all vertices}}\,[CW(a(I))\ *\ CW(VD(i))] \quad (10)$$

Results obtained using a descriptor such as Eq. (7) for modeling QSPR on enthalpies of coordination compounds was reported in ref. [13], while descriptor Eq. (8) was applied in ref. [14]. The use of Eq. (9) constitutes an attempt to realise an additive scheme based on calculating the additive contributions of local graph invariants (LIs). However, computation via such models is based upon the formula

Property=A*DCW+B

The "classical" additive scheme based on the LIs may be organized through the following steps:

1) all CW(x) must be multiplied by the coefficient A, i.e.

Additive Contribution(x)=CW(x) * A

2) the B term must be employed as an extra term in the formula

Property=Additive Contribution(x)+B

where x is A(i) or VD(i)

The multiplicative scheme Eq. (10) may be an effective tool for QSAR/QSPR analysis in cases of non linear correlations between a given property and descriptors like Eq. (7), Eq. (8), or Eq. (9).

Thus, we have a wide range of possibilities at our disposal for deriving DCW based upon CWLGI. Furthermore, when one resorts to the use of other well-known topological indices, such as the Hosoya index, [16] Wiener index [17], Harary number [8], Randic connectivity indices [18], etc. to compute TEE, it is necessary to take recourse to equations with several variables to get an acceptable degree of accuracy as obtained from a linear one-parameter regression based on CWLGI.

In closing, we deem there are quite conclusive and well-grounded evidence that CWLGI are suitable tools for applications in QSAR/QSPR theory and it is worth analyzing the possibility of studying other properties and activities. Work on this issue is under current research and results will be given elsewhere in the future.

Acknowledgments The authors thank the useful comments of the referees which have helped to improve the final version of this paper.

References

1. *Graph Theoretical Approaches to Chemical Reactivity*; Bonchev, D.; Mekenyan, O., Eds.; Kluwer, Dordrecht, 1994.
2. Mihalic, Z.; Trinajstic, N. *J. Chem. Educ.* **1992**, 69, 701.
3. Randic, M. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672.
4. Randic, M.; Jansen, P. J.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1998**, *28*, 60.
5. Katristzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, *24*, 279.

6. Szabó, A.; Ostlund, N. S. *Modern Quantum Chemistry. Introduction to Advanced Electronic Structure Theory*; MacMillan, New York, 1982.
7. Bonchev, D.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **1992**, *9*, 75.
8. Harary, F. *Graph Theory*; Addison-Wesley, Reading, Mass, 1969.
9. *Chemical Applications of Graph Theory*; Balaban, A. T., Ed., Academic Press, London, 1976.
10. Trinajstic, N. *Chemical Graph Theory*; CRC Press, Boca Raton, Florida, 1983.
11. Baskin, I. I.; Stankevich, M. I.; Devdarian, R. O.; Zefirov, N. S. *Russ. J.Struct. Chem.* **1989**, *30*, 145.
12. Randic, M.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 261.
13. Toropov, A. A.; Toropova, A. P. *Russ. J. Coord. Chem.* **1998,** *24*, 89.
14. Toropov, A. A.; Toropova, A. P.; Voropaeva, N. L.; Ruban, I. N.; Rashidova, S. Sh. *Russ. J. Coord. Chem.* **1998**, *24*, 503.
15. Vericat, C.; Castro, E. A. *Commun. Math. Comput. Chem. MATCH* **1996**, *34*, 167
16. Hosoya, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
17. Wiener, H. *J. Am. Chem. Soc.* **1947,** *69*, 17.
18. Randic, M. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164.