DESIGN OF A QSRR - E INDEX WITH HIGH MOLECULAR INFORMATION CONTENT TO DIFFERENCIATE CIS AND TRANS ALQUENES.

E. CORNWELL[*] AND G. CORDANO[1]

[*] Facultad de Ciencias Químicas y Farmacéuticas. Universidad de Chile.
[1]Facultad de Ciencias Químicas y Farmacéuticas. Universidad de Chile.Santiago, Chile.
[*]E-mail: ecornwel@abello.dic.uchile.cl

SUMMARY

Using QSRR modelling, an E index was designed based on the H. P. Schultz MTI index. The E index is a molecular descriptor consisting of a vector and adjacent and distant matrices.

Linear regressions were obtained by means of a QSRR - GLC study of a set of 11 cis and 11 trans alquenes characterized by the E_{cis}, E_{trans} indices. The correlation indices (r) were 0.97495 and 0.95767 respectively. It was demonstrated that in the linear regressions above, a smaller r was obtained, when the E index was replaced by the refractive molar index taken as reference (IRM).

The discrimination E values for cis and trans alquenes were obtained by multiplying by (-1) the distance matrix diagonal elements, corresponding to the carbon atoms that support the π bonds in cis alquenes.

INTRODUCTION

QSRR modelling deals with the quantitative relationships between molecular structure and different chromatographic retention coeffecients[1]. QSRR is a subgroup of QSPR modelling which deals with the quantitative relationships between structure and chemical properties. A review of these models has been given in previous publications[2,3].

The primary aim of QSRR processes is to study the relationship between a dependent variable and one or more independent variables, including the analysis and interpretation of the correlation functions obtained. Retention times and volumes, in all their forms[4] and the Kovats index[5] are QSRR elements considered as dependent variables.

The dependent variable used for a set of 22 cis and trans alquenes, corresponds to the logarithm of relative retention time to the n-hexano (t_{rr}), obtained by means of gaseous chromatography[6]

It should be pointed out that in the independent variables set used in QSPR and QSRR modelling, some are described as compact whilst others we will call as permeable.

The former contain the molecular structure information obtained from their physical/ chemical properties; the latter allow change or re-definition of the elements which constitute the algorithm defining that particular variable (index) or, within a matrix dimension, allow the substitution of the matrix elements (a_{ij}; a_{ii}) corresponding to the parameters which characterize a certain molecular structure. Thus, an independent compact type variable, the boiling point of a molecule for example, does not permit modification of its value without consequently modifying its molecular structure or the conditions by which its value was obtained. However, a permeable independent variable, such as the H. B. Kier[7] valence connectivity chemical index originally created by Randic[8], shows properties subject to redefinition of its algorithm elements. In fact, its initial definition differs from the present one[7,8]

This study demonstrates that the H. P. Schultz (MTI) index can allow the substitution and/or inclusion of different vector and matrix elements (molecular properties) with the aim of producing an index with greater molecular structure information.

The fundamental basis of the work consisted in modifying the MTI index by introducing molecular properties and mathematical operations to cis and trans alquenes in the form of ($a_{ij,}$ a_{ii}) elements of the distance matrix $[D]_{nxn}$ and carbon charges (c_1, c_2,...c_n) in the multiplying vector V. The differentiating procedure was obtained by a −1 product applied to distance matrix $[D^{**}]_{nxn}$ diagonal elements, corresponding to the carbon atoms that support the π bonds in cis alquenes. Both changes enabled a significant correlation to be obtained with the dependent variable (log t_{rr}) and in addition, the ability to differentiate the cis and trans alquenes. These changes to the MTI index were made without modifying the matrix or the vectorial dimensions.

The definition of the E index was based on the MTI index, which is defined by means of equations 1, 2 and 3,

$$Q = V^*([A]_{nxn} + [D]_{nxn}) \qquad (1)$$

$$Q = [\, e_1, e_2, \ldots e_n] \qquad (2)$$

$$MTI = \sum_{i=1}^{n} e_i \qquad (3)$$

where $[A]_{nxn}$ represents the adyacent topological matrix, $[D]_{nxn}$ the distant topological matrix and V the previous multiplying vector. The definitions and characteristics of these elements, are defined in the literature[10, 11, 12]

The next step was to make modifications to the MTI index to structure the E index, and to evaluate these modifications according to the statistical parameters appropriate to the mathematical regressions relating these indices to the dependent variable (log t_{rr}).

First modification to the MTI index. Obtaining the MTI(c) index.

The topological matrix $[D^{**}]_{nxn}$ is used instead of topological matrix $[D]_{nxn.}$ In the former, each element a_{ij} corresponds to the distance in Armstrong (A°) between the graph carbon atoms corresponding to the alquene that is described, and each a_{ii} element is equal to 0. These distance values were obtained by means of specific software[13]. Clearly, a $[D^{**}]_{nxn}$ matrix corresponding to a cis alquene is little different from that corresponding to a trans alquene, due to distances differences. The previous multiplying vector is identical to that used in the MTI index and its elements are equal to the diagonal elements a_{ii} of the adyacent $[A]^2_{nxn}$ matrix. In this modification the structure of adyacent matrix appropriate to MTI $[A]_{nxn}$ is retained.

Second modification to the MTI index. Obtaining the MTI(cr) index.

This index was obtained using the MTI(c) index modification by means of the following procedure: The multiplying vector V with elements $e_1, e_2, \ldots e_{n,}$ was replaced by the supported charge values on each graph carbon atom corresponding to the alquene to be characterized. The charge values for each carbon alquene were calculated by means of Hyperchem[14] software that included and the AM1semi-empirical Steep-Descent algorithm method with RMS of (1Kcal/A° mol) gradient until convergence. This change defines the new multiplying vector V_c .

Third modification to the MTI index. Obtaining the E index.

With the MTI(cr) index, instead of using a_{ii} elements of the $[D^{**}]_{nxn}$ matrix (whose values are equal to 0), d_i, (dj) elements were used corresponding to the $(^1\chi_v)$[7] index calculated by means of equation 4.

$$d_i , d_j = \{(Z^v - h) / (Z - Z^v - 1)\} \qquad 4$$

where Z^v represents the valence electrons, Z the total electrons and h the protons bonded to the carbon atom whose (d_i, d_j) values are to be evaluated, corresponding also to the a_{ii} elements of the $[A]^2_{nn}$ matrix. In order to differentiate cis alquenes from trans alquenes when the value a_{ii} represents a cis alquene, the a_{ii} value of the adjacent carbon atom to a π bond was multiplied by (-1)

Taking account of the above, the E index is defined through equations 5, 6 and 7.

$$Q_E = V_c*([A]_{nxn} + [D^{**}]_{nxn}) \qquad (5)$$

$$Q_E = [\, c_1, c_2, \ldots c_n] \qquad (6)$$

$$MTI = \sum_{I=1}^{n} c_I \qquad (7)$$

The alquene molar refractions index was used as a compact independent variable to differentiate the cis and trans alquenes and thus validate the E index .

When an index is modelled, independent topological variables are used to distinguish cis and trans alquenes and it is necessary to define a priori differentiating criteria[15]. When compact independent variables are used, the difference between isomers is defined per se, for example, when the molar refractive index is used, because it is a molecular structure function.

The inclusion of molecular information in the matrix space corresponding to the molecular characteristics of the substances studied, avoids the use of molecular parameters as independent variables in multi-variate regression systems. The latter type of correlation present two kinds of problems:

a.- Statistical parameter regressions, whose indices of determination (R^2), Fisher index (F), and standard deviation (SD) are acceptable within standard statistical criteria, are nevertheless less significant than their p value indicates. This happens when there is a mis-match between the number of cases treated versus the number of independent variables used[16,17].

b.- For correct interpretation, the multi-regression equation multiplying factors (vectors) of the independent variables must be changed into their orthoganalized form because in their original structure, they exhibit superposition properties[18]

The use of the E index in this study assumes the simple lineal regression equation line (log t_{rr} = m E + n) and that the E index is a function of the molecular distance between carbon atoms, the charge supported by each carbon atom, molecular structure and the differentiating power between cis and trans alquenes.

METHODOLOGY

Table 1 shows a model of 11 cis alquenes and 11 trans alquenes, the relative retention times (t_{rr}) with respect to n-hexane, its logarithm values, and the values of the MTI(c), MTI(cr) and E indices.

For the substances used in this study, the values of the refractive indices, molecular densities and molecular weights are provided, enabling the calculation of their molar refractive indices (IRM) by means of the Lorenz-Lorenz[19] equation. This index (IRM) is used to define the E index by means of its properties, as indicated later. For the calculation of the E index, programs available in the HP 48GX calculator were used, allowing the input of values for equations 5, 6 and 7. The dimensions of the matrix used are related to the number of carbon atoms shown in the molecular graph. Each element a_{ij} of the adjacent matrix $[A]_{nxn}$, is equal to 1 if I and j carbon atoms are adjacent, otherwise they are 0 . Adjacent and distant matrices $[A]_{nxn}$, $[D^{**}]_{nxn}$ are not singular and are symmetrical with respect to the main diagonal. The elements a_{ii} of the $[D^{**}]_{nxn}$ matrix are the elements of the $[A]^2_{nxn}$ diagonal matrix, equivalent to d_i, d_j in the $(^1\chi_v)$[7] index.

When applying equations 5 and 6, the information contained in matrices $[A]_{nxn}$ and $[D^{**}]_{nxn}$ and vector v_c are translated to a vector Q_E, and through equation 7 are transformed into positive scalar values corresponding to the the E index value, where the values of Q_E are considered absolute.

By means of this procedure, given molecular structure characteristics are associated with information on the E index scalar number originating in the MTI and MTI index transformations. The new contributions are: charge on each carbon atom (elements of vector Vc), molecular volume and ramification degree (elements d_i, d_j instead of the a_{ii} $[D^{**}]_{nxn}$ matrix elements) and the cis alquene differentiating power when the a_{ii} value of the adjacent carbon atom to a π bond, was multiplied by (-1).

Tabla 1 Value of differents indeces and parameters used in this study

Num Subst.	Substances	t_{rr} GLC	log t_{rr}	Indice MTI (c)	Indice MTI (cr)	Indice E	Indice refrac.	densi. g/cm^3	PM	Indice IRM
1	Cis-2-butene	0,18	-0,74473	60,878	5,55974	4,91212	1,3931	0,616	56,11	21,7444
2	Cis-2-pentene	0,41	-0,38722	107,738	9,72185	9,35685	1,3830	0,6556	70,14	24,9560
3	Cis-4-methyl-2-pentene	0,71	-0,14874	176,776	14,16704	13,89628	1,3800	0,6690	84,16	29,1398
4	Cis-3-hexene	1,03	0,01284	169,350	14,93102	14,85702	1,3947	0,6796	84,16	29,6690
5	Cis-2-hexene	1,16	0,06446	170,755	15,04087	14,99187	1,3979	0,6869	84,16	29,5643
6	Cis-3-methyl-2-pentene	1,23	0,08991	173,550	14,97476	15,03013	1,4016	0,6886	84,16	29,7335
7	Cis-4,4-dimethyl-2-pentene	1,53	0,18469	251,744	20,24670	20,15949	1,4026	0,6951	98,19	34,4415
8	Cis-4-methyl-2-hexene	1,97	0,29447	247,696	21,49898	21,64232	1,4026	0,6952	98,19	34,4365
9	Cis-3-heptene	2,84	0,45332	275,063	24,39419	21,93609	1,4059	0,7030	98,19	34,3006
10	Cis-3-methyl-3-hexene	2,90	0,46240	241,283	20,93203	21,08603	*	*	98,19	*
11	Cis-3-methyl-2-hexene	3,07	0,48714	246,397	21,49388	21,64951	1,4126	0,712	98,19	34,3587
12	Trans-2-butene	0,14	-0,85387	62,046	5,81054	7,20055	1,3848	0,599	56,11	21,9418
13	Trans-2-pentene	0,38	-0,42022	112,548	10,41859	12,09940	1,3793	0,6431	70,14	25,2220
14	Trans-4-methyl-2-pentene	0,74	-0,13077	183,160	15,56185	17,17705	1,3889	0,6686	84,16	29,7638
15	Trans-3-hexene	1,03	0,01284	180,480	16,39444	18,35992	1,3943	0,6772	84,16	29,7474
16	Trans-2-hexene	1,05	0,02119	177,787	15,96053	17,99213	1,3936	0,6732	84,16	29,8771
17	Trans-3-methyl-2-pentene	1,11	0,04532	158,594	14,52353	16,65429	1,4045	0,693	84,16	29,7330
18	Trans-4,4-dimethyl-2-pentene	1,31	0,11727	260,244	20,81285	22,85365	1,3982	0,6889	98,19	34,4156
19	Trans-4-methyl-2-hexene	1,97	0,29447	248,420	21,85606	23,83006	1,4025	0,6925	98,19	34,5632
20	Trans-3-heptene	2,76	0,44091	271,181	24,15347	26,43532	1,4043	0,6981	98,19	34,4212
21	Trans-3-methyl-3-hexene	2,65	0,42325	229,655	19,71520	21,81004	*	*	98,19	*
22	Trans-3-methyl-2-hexene	2,84	0,45332	250,029	21,99903	24,11103	*	*	98,19	*

* Value not found in the information source used in this work23)

Table 2 shows the cis and trans alquene values for correlation (r), Fisher (F) and standard deviation (SD) indices. For the lineal regressions (y=mx+n) between log t_{rr} and E, MTI(c), MTI(cr) and IRM respectively, it is possible to note the difference between E and IRM index correlations and the changes that occur when molecular information is supplemented in each index (MTI(c) and MTI(cr)) to obtain the E index.

Table 2. Statistical parameters of different correlations studied in this issue.

Indices	Correlation indices (r)	Fischer ratio (F)	Standard desviation S.D.
Cis alquenes			
MTI (c)	0.95021	74.3900	0.12529
MTI (cr)	0.96457	120.3200	0.10636
E	0.97494	172.8783	0.08967
IRM	0.96145	97.8054	0.11056
Trans alquenes			
MTI (c)	0.93173	46.0800	0.150079
MTI (cr)	0.952842	88.7300	0.12756
E	0.95757	99.3569	0.12114
IRM	0.95502	72.6069	0.12254

Table 3 shows the correlation index (r) between the E index versus MTI(c) , MTI(cr) and IRM indices respectively. The values indicate that when two regressions are more similar in their statistical indices (r), the greater is the degree of information superposition by these regressions. This indicates the existence of topologically redundant indices[20)] and is a logical fact, because the E index is the product of a staggered transformation process from the MTI index where there is also a degree of redundancy with the IRM index.

Tabla 3 The regressions correlation indices (r) between the proposed E index and other indices considered in this study.

Proposed	MTI	MTI	IRM
index E	(c) index	(cr) index	index
Cis alquenes	0.98968	0.99120	0.99517
Trans Alquenes	0.99057	0.99970	0.98335

The relationships between log t_{rr} versus $E_{(cis)}$ and $E_{(trans)}$ are expressed by means of equations 8 and 9 respectively.

Cis alquenes.

$$\log t_{rr} = -1.01291(\pm 0.08667) + 0.06634(\pm 0.00504)^{*} E_{(cis)} \qquad 8$$

$r = 0.97499$

$r^2 = 95.0526\%$

$SD = 0.08967$

$F = 172.91$

The p ANOVA calculation is less than 0.01, indicating a significant relationship between the variables at a level of confidence 99%. The correlation (value of r) in the model studied explain the 95.05% of the log t_{rr} dependent variable showing a strong relationship between the variables. The sample SD standard deviation of residual value is 0.08967.

Trans alquenes.

$$\log t_{rr} = -1.23388(\pm 0.13259) + 0.06702(\pm 0.00672)^{*} E_{(trans)} \qquad 9$$

$r = 0.95757$

$r^2 = 91.6941\%$

$SD = 0.12114$

$F = 99.36$

The interpretation of the statistical parameters of equation 9 is in accordance with the previous analysis but based on the particular case values. All the statistical correlations and parameters were calculated and interpreted by means of Statgraphics[21)] software.

The MTI Index does not differentiate cis and trans alquenes. An E index based on the MTI index was designed in this study with differentiating characteristics incorporating two contributions: one, intrinsic isomeric differentiation utilising the different distances between carbon atoms, the other, an a priori configuration of the design of the E index (aii values of the $[D^{**}]_{nxn}$ matrix multiplied by -1 on double bonded adjacent carbon atoms)

Table 2 indicates that the relationship between log t_{rr} versus IRM for both types of isomers presents a significant correlation. The molar refraction index is the sum of the atomic refractions as well as the molecular refraction bonds [22)]. The IRM index is extremely responsive to the charges that are involved in enzyme-substrate interactions. For that reason, the IRM index is used in QSAR studies[22)] (the relationship between biological activity and chemical structure) In a similar way, there is a significant correlation between the E and IRM indices, seeTable 3, and a significant correlation between log t_{rr} versus E_{cis}, thus allowing the conclusion that in a chromatographic system the interaction between the charges in the stationary phase (silanol groups) and the dipole moments due to the presence of cis alquene double bounds are interpreted better by E_{cis} (greater (r) coefficient correlation) than in trans alquenes interpreted by the E_{trans} index. This is because in the cis alquene structure the tiny polarty of the double bonds is more exposed to opposite charges in the stationary phase. For this reason, in general, cis alquenes present a greater chromatographic retention time than their trans isomers.

The reason why the equation regressions obtained in this study present correlation indices which are not significantly high, is due to the fact that in the group of substances chosen for the model, some differ from others in the position of the double bond as well as in the location of the methyl groups with respect to each other. They therefore belong to different substance families[23)] and are classified in different subgroups.

Table 4 presents a linear regression study in the form of y = m*x + n, between the experimental dependent variable log t_{rr} versus the log t_{rr} calculated by means of the E_{cis} and E_{trans} indices (equations 8 and 9 respectively) and for the IRM index . The data collected confirm that the E_{cis} index is more specific for cis alquenes than the E_{trans} index for trans alquenes. The same applies to the IRM. index.

Table 4 Statistical parameters of lineal regressions between experimental log $t_{rr\,exper.}$ and calculated log $t_{rr\,calc.}$ using E and IRM indices.

Statisticals parameters	Calculated base of E_{cis} on the on	Calculated base of E_{trans}	Calculated the base of IRM (cis alquenes)	Calculated the base of IRM (trans alquenes)
r	0.97494	0.95757	0.96145	0.95502
F	172.88049	99.35683	97.80890	72.60701
S.D.	0.08968	0.12133	0,11297	0.12364
n	11	11	10	9

CONCLUSIONS

In this study, different molecular properties were integrated in a single independent variable, E index, by means of a series of simple matrix processes to obtain a scalar containing the molecular structure information introduced originally in the matrix and vector elements.

For cis and trans alquenes, the proposed index E, presents a greater correlation with the dependent variable log t_{rr} than the molar refractive index IRM.

ACKNOWLEDGEMENTS

The authors greatly appreciate the idiomatic support to Victoria Hare Cornwell

REFERENCES

1. R. Kaliszan. Anal Chem 64, 619A (1992).
2. E, Cornwell. Bol. Soc. Chil. Quim. 45, 649 (2000).
3. E, Cornwell. Bol. Soc. Chil. Quim. 47, 53 (2002).
4. C. F. Poole and S. K. Poole "Chromatography today" Edit Elsevier (1991).
5. Report for Analytical Chemistry. Anal. Chem. 36, 31A (1964).
6. G. Zweig, J. Sherma "Hanbook of Chromatography" Vol I Edit Chemical Rubber Co. pag. 143 (1972).
7. L. B. Kier, L. H. Hall. J. Pharm Sci. 72, 1170 (1983).
8. M. Randic. J. A C. S. 97, 6609 (1975).

9. H. P. Schultz.J. Chem. Inf. Comput. Sci 29, 227 (1989).
10. H. P. Schultz, E. B. Schultz, T. P. Schultz. 30, 27 (1990).
11. H. P. Schultz.J. Chem. Inf. Comput. Sci.33, 863 (1993).
12. H. P. Schultz.J. Chem. Inf. Comput. Sci. 36, 996 (1996).
13. C. S Chem 3D Pro. Cambridge Soft Corporation 875 Massachusetts Avenue. Cambridge , M. A. 0213 e U. S. A. Version 4.0
14. Hyperchem. Hypercube and autodesk ,Inc. Developed by Hypercube Inc. (1993).
15. A. Sabljic, O. Horvatic. J. Chem. Inf. Comput. Sci. 33, 292 (1993).
16. E. Estrada, a. Ramirez, J. Chem. Inf. Comput. Sci. 36, 877 (1995).
17. J. G. Toplis, R.J. Castello J. Med. Chem 15, 1066 (1972).
18. J. G. Toplis, R. P. EdwardsJ. Med. Chem. 22, 1238 (1979).
19. A. L.eo, A. Weininger "2CMR Reference Manual" (1995): www.daylight.com/dayhtml/doc/cmr/index.html.
20. M. Randic. J. Chem. Inf. Comput. Sci. 37, 672 (19997).
21. Statgraphic Plus Windows 4.0 Profesional Version Copyright 1994-1999 by Statistical Graphic Corp.
22. Y. Du. Y. Liang, D. Yun. J. Chem. Inf. Comput. Sci. 42, 1283 (2002).
23. D. R. Lide. "CRC Handbook of Chemistry and Physics 75th ed. Edit. CRC Press, Boca Raton FL.. (1994).