

**TOPOLOGICAL QUANTUM SIMILARITY INDICES AND THEIR USE
IN QSAR: APPLICATION TO SEVERAL FAMILIES
OF ANTIMALARIAL COMPOUNDS**

Emili Besalú*, Ana Gallegos and Ramon Carbó-Dorca

Institute of Computational Chemistry and Department of Chemistry

University of Girona

17071 Girona

Spain

e-mail: emili@iqc.udg.es

Abstract. New 3D molecular topological indices are described. From the underlying theoretical foundation, it is revealed the connection between molecular topology and the general theory of Quantum Similarity. Finally, results concerning the establishment of QSAR models, related to five antimalarial molecular families, are presented.

INTRODUCTION

An important field of research in contemporaneous Chemistry is based on the prediction of molecular properties, either physicochemical or biological. Within this paradigm, the QSPR/QSAR (Quantitative Structure-Property and Structure-Activity Relationships) fields are one of the most known, see for example the reviews [1-7] and the references cited therein. One of the oldest and most successful QSPR approaches relies on the topological paradigm.

Within this field, numerical correlation between quantities derived from topological graphs and physicochemical or biological properties usually produces acceptable results [8]. Many times, the mathematically related methods are based on the manipulation of molecular indices as descriptors using linear techniques as Multilinear Regression [9] and the computation of the called $q^{(2)}$ coefficient [9,10].

Nowadays, a relatively large amount of methods, which uses old and new molecular indices, are combined to extract information on molecular properties. Simultaneously, the chemical and pharmacological community needs extended tools capable to grasp complex information, most of it coming from the three-dimensional (3D) molecular structure. Unfortunately, when only classical Topological Indices (TI) derived from Topological Matrices (TMs) are used, part of the spatial information of the molecule may be lost. Some studies are focussed to overcome this problem [11,12].

Parallel to this panorama, our laboratory has been developing the so-called Quantum Similarity Theory [13-20]. In this context, a link between fundamental similarity theory and classical topological approach has been found [16]. As a result, a new kind of tridimensional TI has been described. These indices can be called Topological Quantum Similarity Indices (TQSI).

ATOMIC QUANTUM SIMILARITY MEASURES AND TOPOLOGICAL QUANTUM SIMILARITY MATRICES

The fundamental idea beyond Quantum Similarity Theory consists in measuring, in some way, the similarities between pairwise quantum object density functions. In this context, for every pair of quantum objects, A and B , with respective first order density functions ρ_A and ρ_B , an estimation of the similarities between the particle distributions can be obtained from the following Quantum Similarity Measure:

$$Z_{AB}(\Omega) = \max \left(\int \rho_A \Omega \rho_B dV \right) \quad (1)$$

Ω being a weighting operator. In this expression, when A and B are molecules, the integral argument is optimised choosing the molecular alignment that maximises its value.

Starting from this idea, a simplified application can be envisaged. A similarity measure can be established among every pair of atoms, i and j , of a given molecule defining in this way an Atomic Quantum Similarity Measure (AQSM):

$$A_{ij}(\Omega) = \int \rho_i \Omega \rho_j dV. \quad (2)$$

Here, it must be understood that a density function, ρ , is attached to every atom. If this density function is approximated in such a manner as to possess spherical symmetry, then the AQSM will only depend on the interatomic distance R_{ij} .

In this way, for a molecule with n atoms, a symmetric $n \times n$ Topological Quantum Similarity Matrix (TQSM), $A = A(\Omega)$, is obtained when collecting the measures of type (2) coming from all the atomic pairs, that is, $A = \{A_{ij}\}$. The interesting point is that the TQSM can be easily seen as an extension of the classical TM concept [8,16]. Once the matrix A is known, their elements can be combined in several ways in order to obtain the TQSI, which are well suited to design predictive models in the field of the QSPR/QSAR framework. Consequently, the general QSM theory leads to the generation of new *ab initio* molecular descriptors. Several interesting results have been obtained concerning the numerical correlation between indices derived from the TQSM and physicochemical or biological properties [12,14]. Here, only the main ideas will be outlined. More details can be found in references [12,14,21].

AQSMs are obtained from the 3D molecular geometry coming from crystallographic data or by means of any *ab initio* or semiempirical methodology level. In our laboratory, and for practical purposes, a simple set of 1s GTO basis functions is employed in order to describe atomic densities. The constructed computer program, once given the molecular geometry, associates a unique function of the type

$$g_i = g(\mathbf{r} - \mathbf{R}_i, \zeta_i) = N_i e^{-\zeta_i (\mathbf{r} - \mathbf{R}_i)^2} \quad (3)$$

to each atom. The term N_i is a normalisation factor, $N_i = \left(\frac{2\zeta_i}{\pi} \right)^{\frac{3}{4}} \left(v_i^T \right)^{\frac{1}{2}}$, v_i^T being the topological atomic valence of the atom i . The function exponent ζ_i is one of the parameters which can be numerically modulated for every atom. Table 1 summarises the exponents which are needed in this work [14].

When codifying the molecular structure, an important issue associated to the decision about the presence or absence of the hydrogen atoms is present. As it is customary in the classical topological approaches, in this paper the criteria of ignoring them has been followed.

TABLE 1. Atomic exponents (in a.u.) used in this work.

Atom	Exponent	Atom	Exponent
H	3.436350	F	0.548240
C	0.467380	S	0.249715
N	0.497510	Cl	0.268040
O	0.530530	Br	0.196090

Every TQSM element is computed by means of an integral of the type (2) involving two 1s GTO functions, as defined in equation (3), centered at atoms i and j belonging to the same molecule. Several AQSMs can be defined depending on the nature of the Ω operator appearing in (2) [14,21]. The most relevant are:

- Classical TM, $A = T$: the operator Ω is defined in such a way that, for every element T_{ij} , a value of 1 or 0 is obtained depending on the fact that the atoms i and j are bonded or not. Due to the artificial and ambiguous definition of a chemical bond, such operator must be defined according to a previously established criterion. Classically, the Lewis concept of chemical bond has been extensively used. In our laboratory, as it is customary in other fields, a euclidian distance choice is used to design automatised molecular treatment procedures. In more sophisticated environments, a criterion based on the Bader analysis [22] could be also considered, but this would extremely slow down the computation process.
- Overlap matrix: this is a well-known matrix which is reproduced when the Ω operator becomes unit: $S = A(1)$.
- Cioslowski-like matrix C : the Ω operator is defined in order to give, for every matrix element, the square of the corresponding overlap matrix: $C_{ij} = S_{ij}^2$.

Following a similar criteria which is found in the classical TM construction, the diagonal elements of these matrices are defined as null.

TOPOLOGICAL QUANTUM SIMILARITY INDICES

As in the classical topological approach, appropriate manipulations of the elements embedded in the TQSMs permit to generate TQSI. Such is the origin of the new molecular descriptors proposed here. The details of this new approach are given in reference [14]. From an $n \times n$ TQSM, A , a general expression for a TQSI can be written

$$I = I(C, A) = \sum_n C(n) A(n). \quad (4)$$

Here, a Nested Summation Symbol notation has been used [23-26]. In this context, the vector index $n = \{n_1, n_2, \dots\}$ runs over all the combinations of atomic index numbers. Every index contained in this vector ranges from 1 to n , the number of atoms in the molecule. At the same time, the terms $C(n)$ are numerical coefficients. Their presence and specific structure allow the generation of several kinds of index combinations. Finally, the factor $A(n)$ is a product of elements taken from the TQSM, for example:

$$A(n) = A(n_1, n_2, \dots) = \prod_{i,j} A_{n_i n_j}. \quad (5)$$

This general notation permits to reproduce, among others, the classical TI formulation as can be found in references [8,12,27] such as: Wiener (W), Randic (χ), Schultz (MTI), Balaban (B), Hosoya (Z), and Harary (H) indices, the generalised connectivity indices (${}^m\chi$) of Kier and Hall, and so on. In order to obtain the respective expressions, it is only necessary to invoke the same classical index generating rules but replacing the numbers coming from the TM the ones arising from the TQSM. For example, the valence vectors are generated from TQSM as:

$$\mathbf{v}^A = \{\mathbf{v}_i^A\} = \left\{ \sum_{j=1}^n A_{ij} \right\} \quad (6)$$

Topological distance matrices, within TQSM, will use three dimensional Euclidean distances [12,14].

In this way, several sets of TI can be obtained by variate choices of the operator Ω providing different forms of the TQSM. As commented before, the classical TM and the classical related indices are particular cases arising from the present general formulation.

Due to the generalised molecular topology procedure described here, particular considerations must be taken into account when defining some of the new indices. Special cases have been discussed in references [12,14] and will not be repeated. An example is given below.

The Randić index formula from TM is well-known and reads:

$$\chi^T = \sum_{bonds} \frac{1}{\left(v_i^T v_j^T\right)^{\frac{1}{2}}}. \quad (7)$$

Noted that, in the numerator, a TM element is present. The following generalised definition is here proposed:

$$\chi^A = \sum_{bonds} \frac{A_{ij}}{\left(v_i^A v_j^A\right)^{\frac{1}{2}}}. \quad (8)$$

Of course, other combinations can be considered.

The methodology can be extended to other indices (see Table 2). There, p_3 is the number of atoms separated by three bonds in the molecule, the symbol $[B]_i$ stands for the i -th row of matrix B , μ is the number of cycles, n_e is the number of edges of the related graph, $(D)_i$ stands for the sum of distances from vertex i and n_t is the number of connected subgraphs of type t . Within the classical approach, $p^T(i)$ is the number of ways to draw i non-adjacent bonds in the molecular graph. As a particular case, it is defined $p^T(0) = 1$.

In general, TQSI are constructed based on the same mathematical formulas, as those appearing in Table 2, but the matrix or vector elements are replaced by the corresponding TQSM ones. The summations run over the same integer indices but it is expected that the new indices include, in some way, information about the 3D molecular structure.

When TQSI are considered, the Hosoya index is treated as a slightly special case. Despite the same classical and discrete contributions must be generated automatically, in the present calculations, the $p^A(i)$ term (with $A \neq T$ and $i > 0$) is obtained in a special manner. The product of the TQSM elements attached to the selected bonds is computed and, then, the

negative of the natural logarithm of this product is considered and all the resulting contributions are added.

TABLE 2. Definition of several TIs (see text)

Index	Definition
Wiener	$W = \sum_{i=1}^{n-1} \sum_{j=i+1}^n D_{ij}$
Randic	$\chi^A = \sum_{bonds} A_{ij} \left(v_i^A v_j^A \right)^{-\frac{1}{2}}$
Schultz	$MTI = \sum_{i=1}^n \left[v^A (A' + D') \right]_i$
Harary	$H = \sum_{i=1}^{n-1} \sum_{j=i+1}^n D_{ij}^{-2}$
Balaban	$B = \frac{n_e}{\mu + 1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \left[(D)_i (D)_j \right]^{-\frac{1}{2}}$
Hosoya	$Z^A = \sum_{i=0}^{n/2} p^A(i)$
Generalised connectivity indices of order m and type t	${}^m\chi_t^A = \sum_{s=1}^{n_t} \prod_{i=1}^{m+1} \left(v_i^A \right)^{-\frac{1}{2}}$

Some 3D variants of well-known TI can be also defined. Such indices are the 3D Wiener path number (${}^{3D}W$), 3D Shultz index (${}^{3D}MTI$) and the 3D Harary number (${}^{3D}H$). Their definition is the same as the related in Table 2, but the distance entering into the above indices is the the Euclidean distance between pairs.

Caution must be addressed to the Schultz index definition appearing in Table 2. In order to prevent the dependence from the distance measure units, one possible choice consists in defining the adimensional topological and 3D distance matrices, T^* and D^* :

$$T' = \frac{T}{\max_{i,j}\{T_{ij}\}} \quad \text{and} \quad D' = \frac{D}{\min_{i,j}\{D_{ij} | D_{ij} \neq 0\}}.$$

In this way, the adimensional maximal entry on the matrix T' and the minimal non-zero element on matrix D' are 1.

A MOLECULAR EXAMPLE

Before presenting the application section, here it will be shown how to compute some TQSI for a simple molecule as ethanol. In its hydrogen-suppressed form, only $n=3$ atoms are relevant. In Table 3 the employed cartesian coordinates and the full topological valences attached to every atom have been indicated, while the atomic exponents can be read in Table 1.

TABLE 3. Atomic data related to the hydrogen-suppressed form of ethanol molecule. The molecular skeleton is C_2-C_1-O . The topological valence is referred to the original graph, having hydrogen atoms.

Atom	Topological valence	Cartesian coordinates / a.u.		
		x	y	z
C_2	4	-0.85096	2.7269	0.0000
C_1	4	0.0000	0.0000	0.0000
O	2	2.6827	0.0000	0.0000

In Table 4, the TQSM and the corresponding valence vectors are shown together with the euclidian distance matrix.

TABLE 4. Lower triangles of the TQSMs and Euclidean distance matrix attached to the ethanol molecule. The atom numbering corresponding to the molecular skeleton C_2-C_1-O is ordered: 2-1-3.

Topological	Overlap
$T = \begin{pmatrix} 0 & & \\ 1 & 0 & \\ 1 & 0 & 0 \end{pmatrix}$	$S = \begin{pmatrix} 0 & & \\ 0.54503 & 0 & \\ 0.57662 & 0.08389 & 0 \end{pmatrix}$
$v^T = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$	$v^S = \begin{pmatrix} 1.1217 \\ 0.62892 \\ 0.66051 \end{pmatrix}$

TABLE 4. (continued).

Cioslowski	Distance (a. u.)
$C = \begin{pmatrix} 0 & & \\ 0.29706 & 0 & \\ 0.33249 & 0.00704 & 0 \end{pmatrix}$ $v^C = \begin{pmatrix} 0.62955 \\ 0.3041 \\ 0.33953 \end{pmatrix}$	$D = \begin{pmatrix} 0 & & \\ 2.8566 & 0 & \\ 2.6827 & 4.4635 & 0 \end{pmatrix}$

From the tabulated numerical values in tables 3 and 4, and following equation (8), the Randic index can be computed adding a term for each molecular bond. The computations coming from each of the matrices T , S and C are as follows:

$$\chi^T = \frac{T_{12}}{\left(v_1^T v_2^T\right)^{\frac{1}{2}}} + \frac{T_{13}}{\left(v_1^T v_3^T\right)^{\frac{1}{2}}} = \frac{1}{(2 \times 1)^{\frac{1}{2}}} + \frac{1}{(2 \times 1)^{\frac{1}{2}}} = \sqrt{2}$$

$$\chi^S = \frac{S_{12}}{\left(v_1^S v_2^S\right)^{\frac{1}{2}}} + \frac{S_{13}}{\left(v_1^S v_3^S\right)^{\frac{1}{2}}} = \frac{0.54503}{(1.1217 \times 0.62892)^{\frac{1}{2}}} + \frac{0.57662}{(1.1217 \times 0.66051)^{\frac{1}{2}}} = 1.3188$$

$$\chi^C = \frac{C_{12}}{\left(v_1^C v_2^C\right)^{\frac{1}{2}}} + \frac{C_{13}}{\left(v_1^C v_3^C\right)^{\frac{1}{2}}} = \frac{0.29706}{(0.62955 \times 0.3041)^{\frac{1}{2}}} + \frac{0.33249}{(0.62955 \times 0.33953)^{\frac{1}{2}}} = 1.3981$$

It is well known that, within the classical topological formulation, the Randic index is equivalent to the connectivity path one [28]: $\chi^T = {}^1\chi_p^T$. In the present approach, and looking at the corresponding definitions in Table 2, it can be realised that these indices may have different values. This is due to the contribution of the A_{ij} terms into the first one, while in the later only topological valences are present. For the ethanol molecule, the connectivity path ${}^1\chi_p^A$ indexes are

$${}^1\chi_p^T = \chi^T = \sqrt{2}$$

$${}^1\chi_p^S = \frac{1}{\left(v_1^S v_2^S\right)^{\frac{1}{2}}} + \frac{1}{\left(v_1^S v_3^S\right)^{\frac{1}{2}}} = \frac{1}{(1.1217 \times 0.62892)^{\frac{1}{2}}} + \frac{1}{(1.1217 \times 0.66051)^{\frac{1}{2}}} = 2.3524$$

$${}^1\chi_P^C = \frac{1}{\left(v_1^C v_2^C\right)^{\frac{1}{2}}} + \frac{1}{\left(v_1^C v_3^C\right)^{\frac{1}{2}}} = \frac{1}{\left(0.62955 \times 0.3041\right)^{\frac{1}{2}}} + \frac{1}{\left(0.62955 \times 0.33953\right)^{\frac{1}{2}}} = 4.4485$$

and from here it can be seen how different TQSM can really lead to different index values.

Concerning the Hosoya index, for this particular molecular example only two kinds of contributions are relevant: $p^T(0)$ and $p^A(1)$. Their values, apart from being considered molecular descriptors themselves, once added, give the Z^A values:

$$Z^T = 1 + 2 = 3$$

$$Z^S = -(\ln S_{12} + \ln S_{13}) = -\ln(S_{12}S_{13}) = 1.1575$$

$$Z^C = -(\ln C_{12} + \ln C_{13}) = -(\ln S_{12}^2 + \ln S_{12}^2) = -2(\ln S_{12} + \ln C_{13}) = 2Z^S$$

This computation procedure demands a routine able to generate explicitly all the Hosoya contributions. Reference [29] provides an appropriate algorithm to calculate it.

APPLICATION EXAMPLES

As one of the current interests in our laboratory is related to the QSAR study of antimalarial compounds, five molecular families of antimalarial agents have been chosen as an application example. The linear models, which will be presented here, have always been obtained following the protocol described below. In this way, the final results and conclusions will possess the additional property that they were obtained using exactly the same systematic procedure. This characteristic enriches the practical utility of the methodology.

All the molecules were initially optimised using the semiempirical AM1 calculations from the AMPAC 6.55 [30] program. Then, the structures were sent to the program package developed in our laboratory where the TM and TQSI were computed. During the phase of molecular indices generation, some additional restrictions have been considered in order to reduce the amount of data to be analysed: the generalised connectivity indices of Kier and Hall have been computed only up to order 9 and only contributions up to also order 9 were considered to obtain the Hosoya index.

Once the TQSI matrices were obtained, each array was sent to a multiple linear regression program. All the combinations of 2, 3 and 4 descriptors were generated and the ones attached to the highest values of the $q^{(2)}$ coefficient [9,10] are reported in the tables presented below. There, the linear correlation coefficients arising from a true cross-validation procedure (Leave-one-out method), r^2_{cv} , and the data fitting, r^2 , are given. Exceptuating for the particular cases which are specifically indicated, along the presented tables the statistical significance parameter coming from the Snedecor F-test, p , is lesser than 0.0001 for all the correlation coefficients attached to the cross-validation processes.

The studied five molecular families, the original experimental references and more numerical details are given next. Linear models are specified in terms of the TQSI.

System 1

This molecular family is composed by a set of 15 3-alkyl substituted analogs of artemisinin [31] (see Figure 1 and Table 5). In vitro activities against W-2 and D-6 strains of *Plasmodium falciparum* are reported in the original article in terms of relative IC₅₀ value.

The optimal linear models and the related r^2_{cv} cross-validation parameters are reported in Tables 6 and 7, where k stands for the number of indices entering in the model.

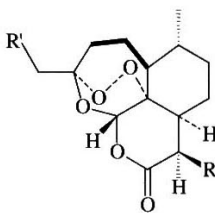


FIGURE 1. General molecular structure of the 3-alkyl substituted analogs of artemisinin molecules of system 1.

TABLE 5. Activities and structures of the molecules of system 1
(see Figure 1)

Molecule	R'	R	Relative activity ^a	
			D-6	W-2
1	H	CH ₃	100	100
2	CH ₃	H	88	112
3	CH ₃ CH ₂	H	2102	673
4	CH ₃ (CH ₂) ₂	H	20	18
5	(CH ₃) ₂ CH	H	53	45
6	EtO ₂ CCH ₂	H	232	232
7	C ₆ H ₅ CH ₂	H	3	1
8	<i>p</i> -ClC ₆ H ₄ (CH ₂) ₂	H	114	127
9	C ₆ H ₅ (CH ₂) ₃	H	220	281
10	CH ₃	CH ₃ (CH ₂) ₃	184	257
11	CH ₃ (CH ₂) ₂	CH ₃ (CH ₂) ₃	28	33
12	C ₆ H ₅ CH ₂	CH ₃ (CH ₂) ₃	1	1
13	<i>p</i> -ClC ₆ H ₄ (CH ₂) ₂	CH ₃ (CH ₂) ₃	43	53
14	C ₆ H ₅ (CH ₂) ₃	CH ₃ (CH ₂) ₃	39	48
15	EtO ₂ CCH ₂	CH ₃ (CH ₂) ₃	1382	2285

^aIt was computed as the relative quantity $100 \frac{(IC_{50})_{artemisinin}(W)_{analog}}{(IC_{50})_{analog}(W)_{artemisinin}}$,

where *W* stands for the molecular weight [31].

In both series of results a qualitative improvement of the model is obtained when 3 descriptors are considered. The linear equations involving 4 descriptors are also indicated but the possibility to deal with an overparametrised model shall be taken into account. This idea can be also applied to other families presented in this study.

TABLE 6. Linear models having a maximal r^2_{cv} value for every set of *k* descriptors. The variable *y* stands for the logarithm of the D-6 activity measure reported in Table 5.

<i>k</i>	r^2 and r^2_{cv}	Linear model equation (log D-6 activity)
2	0.590 0.422	$y = 0.0107418 WPN^T - 0.00570827 P^T(3) + 1.09179$
3	0.839 0.72	$y = -14.2616 {}^4\chi_P^S + 6.81012 {}^4\chi_P^C + 20.4000 {}^8\chi_{CH}^C - 13.0016$
4	0.933 0.851	$y = 13.9258 {}^4\chi_{PC}^T - 5.88954 {}^5\chi_{PC}^T - 19.1486 {}^6\chi_P^S + 8.31544 {}^8\chi_P^C - 0.961919$

TABLE 7. Linear models having a maximal r^2_{cv} value for every set of k descriptors. The variable y stands for the logarithm of the W-2 activity measure reported in Table 5.

k	r^2 and r^2_{cv}	Linear model equation (log W-2 activity)
2	0.631 0.476	$y = 0.0121877 W^T - 6.46274 \cdot 10^{-3} p^T(3) + 0.262345$
3	0.854 0.744	$y = -3.98021 {}^8\chi_p^T + 1.88598 {}^4\chi_p^C - 13.6583 {}^6\chi_{CH}^C + 2.47056$
4	0.961 0.897	$y = 2.54094 \chi^T - 6.07153 {}^5\chi_p^S - 52.5154 {}^8\chi_p^S + 11.4215 {}^8\chi_p^C + 4.32928$

System 2

This family is composed by a set of 17 analogs of 10-deoxoartemisinin substituted at positions C-3 and C-9 [32] (see Figure 2). In vitro molecular activities are of the same nature as the ones reported for system 1.

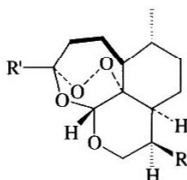


FIGURE 2. General molecular structure of the analogs of 10-deoxoartemisinin antimalarial molecules of system 2.

TABLE 8. Activities and structures of the molecules of system 2

Molecule	R'	R	Relative activity ^a	
			D-6	W-2
1	CH ₃	CH ₃	659	567
2	CH ₃	H	237	190
3	CH ₃	CH ₃ CH ₂	914	466
4	CH ₃	CH ₃ (CH ₂) ₂	473	550
5	CH ₃	CH ₃ (CH ₂) ₃	5826	2090
6	CH ₃	CH ₃ (CH ₂) ₄	170	145
7	CH ₃	C ₆ H ₅ (CH ₂) ₃	5073	2506
8	CH ₃	<i>p</i> -ClC ₆ H ₄ (CH ₂) ₃	6991	3317

TABLE 8. (continued)

9	CH ₃ CH ₂	H	10	10
10	CH ₃ (CH ₂) ₂	H	722	685
11	CH ₃ (CH ₂) ₃	H	653	556
12	(CH ₃) ₂ CHCH ₂	H	183	250
13	C ₆ H ₅ (CH ₂) ₄	H	336	380
14	C ₆ H ₅ (CH ₂) ₂	H	6	2
15	<i>p</i> -ClC ₆ H ₄ (CH ₂) ₃	H	13	28
16	(CH ₂) ₂ CO ₂ Et	H	422	506
17	(CH ₂) ₂ CO ₂ H	H	0.09	0.09

^aObtained in reference [32] in the same way as in Table 5.

In the experimental paper [32] the authors reported some results related to qualitative SAR but no satisfactory QSAR equations were obtained even using, among others, topological and connectivity/shape indices from the Tsar [33] program. The present results in tables 9-12 show how the combination of several kinds of TQSI activates a synergic effect which leads to acceptable linear models.

TABLE 9. Linear models having a maximal r_{cv}^2 value for every set of k descriptors. The variable y stands for the logarithm of the D-6 activity measure reported in Table 8.

k	r^2 and r_{cv}^2	Linear model equation (log D-6 activity)
3	0.711 0.486	$y = 44.6304^4 \chi_C^T - 5.62376^3 \chi_C^S + 2.12991^5 \chi_{PC}^C - 18.6223$
4	0.808 0.639	$y = 33.6116^3 \chi_C^T - 7.93018^4 \chi_{PC}^T - 37.1546^3 \chi_C^S + 2.65114^5 \chi_{PC}^C - 6.38758$

TABLE 10. Linear models having a maximal r_{cv}^2 value for every set of k descriptors. The variable y stands for the logarithm of the W-2 activity measure reported in Table 8.

k	r^2 and r_{cv}^2	Linear model equation (log W-2 activity)
3	0.699 0.478	$y = -6.40997^7 \chi_P^T + 2.43942^4 \chi_P^C + 1.27781^7 \chi_{PC}^C - 14.8043$
4	0.824 0.657	$y = 31.5817^3 \chi_C^T - 7.79599^4 \chi_{PC}^T - 34.9313^3 \chi_C^S + 2.77878^5 \chi_{PC}^C - 6.50120$

TABLE 11. Linear models having a maximal r_{cv}^2 value for every set of k descriptors. The variable y stands for the logarithm of the D-6 activity measure reported in Table 8. The models were obtained without considering the molecule number 17.

k	r^2 and r_{cv}^2	Linear model equation (log D-6 activity)
2	0.611	$y = 2.32173 \cdot 10^{-5} p^T(9) - 22.0545^7 \chi_{CH}^C + 12.0052$
	0.460	
3	0.682	$y = 6.33065 \cdot 10^{-6} Z^T - 1.03067^7 \chi_P^C - 29.0119^7 \chi_{CH}^C + 19.2279$
	0.541	
4	0.837	$y = -6.99820^3 \chi_P^T + 4.42378 \cdot 10^{-3} MTT^S - 3.17224^3 \chi_C^C + 1.57109^7 \chi_{PC}^C + 24.2764$
	0.744	

TABLE 12. Linear models having a maximal r_{cv}^2 value for every set of k descriptors. The variable y stands for the logarithm of the W-2 activity measure reported in Table 8. The models were obtained without considering the molecule number 17.

k	r^2 and r_{cv}^2	Linear model equation for (log W-2 activity)
2	0.613	$y = 2.18192 \cdot 10^{-5} p^T(9) - 20.7077^7 \chi_{CH}^C + 11.3316$
	0.443	
3	0.762	$y = 33.2729 B^T - 7.59291^4 \chi_{PC}^S + 5.73675 \cdot 10^{-3} MTT^C - 48.8439$
	0.627	
4	0.881	$y = -7.09264^3 \chi_P^T + 4.58703 \cdot 10^{-3} MTT^S - 7.39280^3 \chi_C^S + 1.57086^7 \chi_{PC}^C + 26.5073$
	0.784	

Molecule 17 exhibits a very low value of both experimental values shown in Table 8. The point lies quite far away from the remaining molecular data. Usually, this effect allows to apparently improve the value of the cross-validation parameter. However, this time the model is enhanced just removing the analog as it can be seen from the results of tables 11 and 12. Such an effect can be related to the fact that activity of molecule number 17 is substantially different from the ones corresponding to other molecules of the same family and having similar chemical structures.

System 3

The antimalarial system 3 is constituted by 21 β -methoxyacrylates having different linkers against chloroquine-sensitive (NF54) and chloroquine-resistant (K1) *P. falciparum* in vitro [34]. For the last molecule only a unique semiquantitative value is available. Apparently, good results are obtained when considering molecule number 21 but this can be due to the

artificial and ambiguous extrapolated experimental value attached to it. On the other hand, the low activity of molecule number 20 also seems to distort the molecular data cloud. If the first 19 molecules are taken into account, the molecular activity distribution becomes more uniform and satisfactory models are obtained, as presented below.

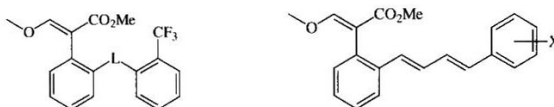


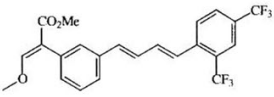
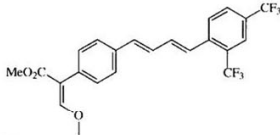
FIGURE 3. Molecular structures of the molecules of system 3 (see also Table 13).

Tables 14 and 15 show the linear models obtained for the two reported activities. As it was cited above, molecules number 20 and 21 were discarded.

TABLE 13. Molecular structures and activities of the antimalarial molecules of system 3 (see also Figure 3).

Molecule	Structures (see Figure 3)	Activity ^a	
		NF54	K1
1	$L = \text{CH}_2\text{CH}_2\text{SCH}_2$	14.4	75.8
2	$L = \text{CH}_2\text{CH}=\text{CHCH}_2$	8.31	21.5
3	$L = \text{CH}_2\text{CH}=\text{CHCH}_2\text{SCH}_2$	4.29	6.15
4	$L = \text{CH}_2\text{CH}_2\text{ON}=\text{C}(\text{CH}_3)\text{CH}_2$	0.42	1.6
5	$L = \text{CH}_2\text{CH}=\text{CHCH}=\text{CHCH}_2$	0.15	0.39
6	$L = (\text{CH}_2)_6$	1.38	3.9
7		0.91	4.2
8	$X = \text{H}$	2.5	11.5
9	$X = 2\text{-Cl}$	0.25	1.01
10	$X = 2\text{-CN}$	1.51	4.63
11	$X = 3\text{-F}$	4.43	24.8

TABLE 13. (continued)

12	X = 3-CF ₃	20.1	43.
13	X = 3-Br	15.3	78.6
14	X = 4-Cl	1.47	5.64
15	X = 2,4-di-CF ₃	0.13	0.28
16	X = 2,4-di-Cl	0.08	0.26
17	X = 2,4-di-Me	0.09	0.14
18	X = 2-Cl, 4-F	0.16	0.51
19	X = 3-MeO, 2-NO ₂	0.27	1.40
20		385.4	868.6
21		>11000	>11000

^aActivity is reported as IC₅₀ in nmol l⁻¹, a quantity derived from the original data of reference [34].

TABLE 14. Linear models having a maximal r^2_{cv} value for every set of k descriptors. The variable y stands for the logarithm of the activity measure reported in Table 13.

k	r^2 and r^2_{cv}	Linear model equation (log NF54 activity)
2	0.652 0.508 ^a	$y = -4.26130^3 \chi_p^S + 14.9392^7 \chi_p^S + 8.06932$
3	0.797 0.726	$y = -2.62207^3 \chi_p^T + 9.55966^7 \chi_p^T - 6.38591^8 \chi_p^C + 8.78029$
4	0.815 0.75	$y = -2.45384^3 \chi_p^T + 10.6961^7 \chi_p^T - 7.62112^8 \chi_p^C - 0.328296^3 \chi_p^S + 9.22398$

^aThe statistical significance parameter is $p = 0.00062$.

TABLE 15. Linear models having a maximal r^2_{cv} value for every set of k descriptors. The variable y stands for the logarithm of the activity measure reported in Table 13.

k	r^2 and r^2_{cv}	Linear model equation (log K1 activity)
3	0.759	$y = -1.03123 p^T(1) + 12.5640^7 \chi_p^T - 9.69714^8 \chi_p^C + 13.4701$
	0.676	
4	0.870	$y = 1.99351^6 \chi_p^T - 13.4758^3 \chi_p^S + 71.7140^9 \chi_p^S - 50.3532^8 \chi_C^C + 12.3526$
	0.766	

System 4

This molecular group consists on a series of 17 3-methyl-10-(substituted-phenyl)flavins (see structure in Figure 4 and in Table 16). The activity is given in Table 16 as the action versus the lethal parasite *Plasmodium vinckei* in mice [35]. The best linear models obtained for this family can be read in Table 17.

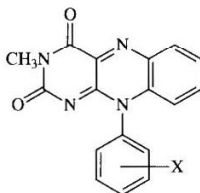


FIGURE 4. General structure of the 3-methyl-10-(substituted-phenyl)flavins of molecular system 4.

TABLE 16. Molecules and biological activity consisting the antimalarial system 4.

Molecule	Structures (X, Figure 4)	Activity ^a (ED ₄₀)
1	4-Br	38.4
2	4-Cl	38.8
3	3,5-di-Cl	40.2
4	3-CF ₃	79.3
5	3-Cl, 5-Me	85.7
6	4-F	103
7	3,5-di-Me	105

TABLE 16. (continued)

8	4-CF ₃	135
9	4-OMe	138
10	3-Br	148
11	4-Cl, 3-Me	182
12	3,4-di-Me	210
13	3,5-di-OMe	219
14	3-Cl	229
15	H	248
16	4-Et	281
17	3-Cl, 4-Me	456

^aActivity is given as the effective dose (in mmol kg⁻¹ 10⁻³) required to obtain a parasitemia of 40% in 48h (ED₄₀) [35].

TABLE 17. Linear models having a maximal r^2_{cv} value for every set of k descriptors. The variable y stands for the logarithm of the activity measure reported in Table 16.

k	r^2 and r^2_{cv}	Linear model equation (log ED ₄₀ activity)
3	0.679 0.51 ^a	$y = 0.0443328 p^T(2) - 16.3798 {}^9\chi_p^T + 3.17847 {}^3\chi_p^S + 8.07947$
4	0.777 0.673	$y = 9.44469 \cdot 10^{-3} MTT^T - 1.41134 {}^5\chi_p^T - 0.0111852 {}^{3D}MTT^S - 967.435 {}^7\chi_C^S + 77.8123$

^aStatistical significance $p = 0.0013$.

System 5

This family was constituted originally by 27 phenothiazine derivatives [36] with capacity to inhibit the *Plasmodium falciparum* cysteine protease falcipain activity. Several trials were carried out among this molecular set and no satisfactory results have been yet obtained except for one case. If the set of 11 molecules having a sulphur atom is considered (see Figure 5 and Table 18), a good linear model is obtained with 2 descriptors (see Table 19). Results can be apparently improved by means of the construction of models involving 3 or more descriptors but this can be a spurious result arising from the system overparametrisation.

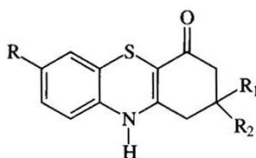


FIGURE 5. Molecular structures of the 16 molecules (taken from reference [36]) consisting the system 5.

TABLE 18. Molecules and biological activities consisting the system 5.

Molecule	R	R ₁	R ₂	Activity ^a (IC ₅₀)
1	Cl	H	H	40
2	Cl	CH ₃	CH ₃	30
3	Cl	H	C ₆ H ₅	10
4	Cl	H	3-CH ₃ OC ₆ H ₄	20
5	Cl	H	4-CH ₃ OC ₆ H ₄	10
6	Cl	H	2,3-(CH ₃ O) ₂ C ₆ H ₃	10
7	Cl	H	3,4-(CH ₃ O) ₂ C ₆ H ₃	30
8	Cl	H	4-ClC ₆ H ₄	4
9	Cl	H	2,4-Cl ₂ C ₆ H ₃	10
10	F	CH ₃	CH ₃	60
11	F	H	C ₆ H ₅	20
12	F	H	3-CH ₃ OC ₆ H ₄	20
13	F	H	2,3-(CH ₃ O) ₂ C ₆ H ₃	20
14	F	H	2,4-Cl ₂ C ₆ H ₃	10
15	F	H	4-ClC ₆ H ₄	5
16	F	H	3,4-(CH ₃ O) ₂ C ₆ H ₃	20

^aIC₅₀ for inhibition of falcipain activity measured as the hydrolysis of Z-Phe-Arg-AMC [36].

TABLE 19. Linear models having a maximal r^2_{cv} value for every set of k descriptors. The variable y stands for the logarithm of the activity measure reported in Table 18

k	r^2 and r^2_{cv}	Linear model equation (log IC ₅₀ activity)
1	0.575 0.437 ^a	$y = -7.18658^7 \chi_{CH}^C + 3.91614$
2	0.765 0.671	$y = 24.7079^8 \chi_{CH}^S - 0.680162^2 \chi_P^C + 1.67282$
3	0.884 0.811	$y = 1.35265^3 \chi_P^C - 16.0903^7 \chi_{CH}^C - 8.27100^9 \chi_P^C + 6.72932$

^aStatistical significance $p=0.0053$.

GENERAL DETAILS AND OTHER CONSIDERATIONS

For each family of n members a matrix of descriptors of dimension $n \times m$ (m =number of descriptors) was obtained. Originally, for all the systems m was $46 \times 3 = 138$ because the matrix of indices was obtained by juxtaposition of the tree kinds of available TQSMs. Then, after removing null or other kind of irrelevant columns (for instance, those originating linear dependencies), for the studied system 4 the parameter m becomes 121 and $m=126$ for the rest.

From the obtained linear models it can be seen that connectivity indices are used many times and, among them, these of higher order are commonly requested. Indices derived from the classical TM ($A=T$) are used but many indices coming from matrices S and C are also employed too. This facts indicate that this new approach really contributes to improve the old methodological capabilities. Nevertheless, at least for the families studied here, only in some cases the Hosoya index or its contributions are used and, when requested, only those coming from the classical definition [37,38] occurs in the models. In previous work, were a family of 31 steroids was studied [14], the Hosoya indices coming from any kind of TQSM took a relevant role. Then, one can deduce that the presence or absence of indices in the linear models is related to the nature of the studied molecular sets and activities.

CONCLUSIONS

A new method for the calculation of molecular TI has been described. It has been put in evidence the connection of molecular topology with the general theory of Quantum Similarity. Satisfactory results concerning the establishment of QSAR models, related to five antimalarial molecular families, have been obtained.

ACKNOWLEDGEMENTS

Financial resources from the European Community Commission contract #ENV4-CT97-0508, the CICYT Research Project: #SAF 96-0158, the Fundació Maria Francisca de Roviralta and a University of Girona grant for the investigation project (session date 3/00) are also acknowledged.

REFERENCES

- [1] Y. C. Martin, *Quantitative Drug Design*, Medicinal Research Series, vol 8, Marcel Dekker, Inc., New York, 1978.
- [2] E. J. Lien. *SAR, Side effects and Drug Design*. Medicinal Research Series, vol 11, Marcel Dekker, Inc., New York, 1987.
- [3] Y. C. Martin, E. Kutter and V. Austel (eds.), *Modern Drug Research*, Medicinal Research Series, vol 12, Marcel Dekker, Inc., New York, 1989.
- [4] C. G. Wermuth (ed.), *Trends in QSAR Molecular Modelling* 92, Escom, Leiden, 1993.
- [5] H. Kubinyi (ed.), *3D QSAR in Drug Design*, ESCOM, Leiden, 1993.
- [6] H. van de WaterBeemd (ed.), *Structure-Property Correlations in Drug Research*, Acad. Press, San Diego (CA), 1996.
- [7] M. Charton (ed.), *Advances in Quantitative Structure-Property Relationships*, vol 1, Jai Press. Greenwich (Conn.), 1996.
- [8] A. T. Balaban, *J. Chem. Inf. Comput. Sci.*, **1995**, 35, 339-350.
- [9] D. C. Montgomery and E. A. Peck, *Introduction to Linear Regression Analysis*, Wiley, New York, 1992.
- [10] D. M. Allen, *Technometrics*, **1974**, 16, 125-127.

- [11] M. Randić, B. Jerman-Blazić and N. Trinajstić, *Computers Chem.*, **1990**, *14*, 237-246.
- [12] M. Lobato, L.I. Amat, E. Besalú and R. Carbó-Dorca. *Quant. Struct.-Act. Relat.*, **1997**, *16*, 465-472
- [13] R. Carbó-Dorca, D. Robert, L.I. Amat, X. Gironés and E. Besalú, *Molecular Quantum Similarity in QSAR and Drug Design*, Lecture Notes in Chemistry vol. 73, Springer Verlag, Berlin, 2000.
- [14] R. Carbó-Dorca, L. Amat, E. Besalú, X. Gironés and D. Robert, Quantum Molecular Similarity: Theory and Applications to the Evaluation of Molecular Properties, Biological Activities and Toxicity. In: *The Fundamentals of Molecular Similarity*, Ed.: R. Carbó-Dorca. Kluwer Acad. Press, Dordrecht, 2000, (in press).
- [15] R. Carbó-Dorca, L. Amat, E. Besalú, X. Gironés, D. Robert. *J. Mol. Struct. (Theochem)*, **2000**, *504*, 181-228.
- [16] R. Carbó-Dorca, L.I. Amat, E. Besalú, M. Lobato, Quantum Molecular Similarity. In: *Advances in Molecular Similarity*, vol. 2, Eds.: R. Carbó-Dorca and P. G. Mezey, JAI Press, London, 1998, pp. 1-42.
- [17] R. Carbó, M. Arnau, L. Leyda, *Int. J. Quant. Chem.*, **1980**, *17*, 1185.
- [18] R. Carbó and E. Besalú, Theoretical Foundations of Quantum Similarity. In: *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*, Ed.: R. Carbó, Kluwer Academic Publishers, Amsterdam, 1995, pp. 3-30.
- [19] R. Carbó-Dorca, E. Besalú, L.I. Amat and X. Fradera, Quantum Molecular Similarity Measures: Concepts, Definitions and Applications to QSAR. In: *Advances in Molecular Similarity*, vol. 1, Eds.: R. Carbó-Dorca and P. G. Mezey. JAI Press, London, 1996, pp. 1-42.
- [20] R. Carbó, B. Calabuig, L. Vera and E. Besalú, *Adv. Quantum Chem.*, **1994**, *25*, 253.
- [21] E. Besalú and R. Carbó, *Scientia Gerundensis*, **1995**, *21*, 145-152.
- [22] R. F. W. Bader, *Atoms in Molecules, a Quantum Theory*, Clarendon Press, Oxford, 1990.
- [23] R. Carbó and E. Besalú, *J. Math. Chem.*, **1993**, *13*, 331-342.
- [24] R. Carbó and E. Besalú, *Computers and Chem.*, **1994**, *18*, 117-126.
- [25] R. Carbó and E. Besalú, *J. Math. Chem.*, **1995**, *18*, 37-72.
- [26] R. Carbó and E. Besalú, Applications of Nested Summation Symbols to Quantum Chemistry: Formalism and Programming Techniques. In: *Strategies and Applications in Quantum Mechanics*, Eds.: Y. Ellinger and M. Defranceschi. Kluwer Acad. Pub., Dordrecht, 1996, pp. 229-248.
- [27] Z. Mihalic and N. Trinajstić, *J. Chem. Educ.*, **1992**, *69*, 701-712.
- [28] L. B. Kier and L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976.
- [29] M. M. Balakrishnarajan and P. Venuvanalingam, *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 1122-1126.

- [30] AMPAC 6.55, **1999**, Semichem, 7128 Summit, Shawnee, KS 66216 D. A.
- [31] M. A. Avery, S. Mehrotra, J. D. Bonk, J. A. Vroman, D. K. Goins and R. Miller, *J. Med. Chem.*, **1996**, *39*, 2900-2906.
- [32] M. A. Avery, S. Mehrotra, T. L. Johnson, J. D. Bonk, J. A. Vroman and R. Miller, *J. Med. Chem.*, **1996**, *39*, 4149-4155.
- [33] Tsar, version 2.41, Oxford Molecular Group, Inc. CAChe Scientific, Beaverton, OR.
- [34] J. Alzeer, J. Chollet, I. Heinze-Krauss, C. Hubschwerlen, H. Matile and R. G. Ridley, *J. Med. Chem.*, **2000**, *43*, 560-568.
- [35] W. B. Cowden, P. K. Halladay, R. B. Cunningham, N. H. Hunt and I. A. Clark, *J. Med. Chem.*, **1991**, *34*, 1818-1822.
- [36] J. N. Domínguez, S. López, J. Charris, L. Iarruso, G. Lobo, A. Semenov, J. E. Olson and P. J. Rosenthal, *J. Med. Chem.*, **1997**, *40*, 2726-2732.
- [37] H. Hosoya, *Bull. Chem. Soc. Jap.*, **1971**, *44*, 2332-2339.
- [38] H. Hosoya, *Bull. Chem. Soc. Jap.*, **1971**, *45*, 3415-3421.