Internet Electronic Journal of Molecular Design **2004**, *3*, 426–442 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

July 2004, Volume 3, Number 7, Pages 426–442

Editor: Ovidiu Ivanciuc

Similarity Matrices Quantitative Structure–Activity Relationships for Anticonvulsant Phenylacetanilides

Ovidiu Ivanciuc

Sealy Center for Structural Biology, Department of Human Biological Chemistry & Genetics, University of Texas Medical Branch, Galveston, Texas 77555–0857

Received: February 15, 2003; Revised: November 28, 2003; Accepted: March 17, 2003; Published: July 31, 2004

Citation of the article:
O. Ivanciuc, Similarity Matrices Quantitative Structure–Activity Relationships for Anticonvulsant Phenylacetanilides, *Internet Electron. J. Mol. Des.* **2004**, *3*, 426–442, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Similarity Matrices Quantitative Structure–Activity Relationships for Anticonvulsant Phenylacetanilides

Ovidiu Ivanciuc*

Sealy Center for Structural Biology, Department of Human Biological Chemistry & Genetics, University of Texas Medical Branch, Galveston, Texas 77555–0857

Received: February 15, 2003; Revised: November 28, 2003; Accepted: March 17, 2003; Published: July 31, 2004

Internet Electron. J. Mol. Des. **2004**, *3* (7), 426–442

Abstract

Molecular graph descriptors are used in developing structure–property models, in drug design, virtual synthesis, similarity and diversity assessment. We present a new application of topological indices in computing similarity matrices that are subsequently used to develop quantitative structure–property relationship and quantitative structure–activity relationship models. The molecular structure is described by similarity matrices obtained from similarity indices calculations, when each molecule is compared to every other from the data set. Four similarity indices are introduced for the computation of the molecular similarity from a set of topological indices that numerically characterize the structure of chemical compounds. Using the multilinear regression model, the significant columns from the similarity matrices are selected as independent variables in a structure–activity study of anticonvulsant phenylacetanilides. The results obtained show that similarity matrices derived from molecular graph descriptors can provide the basis for the investigation of quantitative structure–activity relationships.

Keywords. QSAR; quantitative structure–activity relationships; molecular similarity; similarity matrices; molecular graph; topological indices; molecular graph operators.

1 INTRODUCTION

Structural descriptors that express in numerical form the chemical structure are used in quantitative structure–property relationships (QSPR) and quantitative structure–activity relationships (QSAR) studies, in drug design, virtual synthesis, similarity and diversity assessment of combinatorial libraries. They belong to several classes of descriptors, depending on the model used to represent the chemical compounds, *i.e.* constitutional, graph–theoretical, topological, geometrical, electrostatic, quantum, and grid descriptors. Modern 3D QSAR techniques generate thousands of highly correlated grid descriptors that describe various molecular fields. The resulting QSAR data matrix, containing thousands of columns corresponding to individual grid points, cannot be correlated with the multilinear regression (MLR) method. Usually, the partial least squares (PLS)

* Correspondence author; E–mail: ivanciuc@netscape.net.

method is used to extract from this data matrix the information regarding the ligand–receptor interactions and to generate a 3D QSAR. PLS extracts principal component–like vectors (latent variables) from the matrices of independent and dependent variables. This method takes a matrix containing a large number of potentially useful structural descriptors, which can be highly intercorrelated, and offers a correlation using the latent variables.

A procedure for reducing the too large number of independent variables from grid 3D QSAR was developed using techniques from the similarity analysis. At a qualitative level, molecular similarity has played a fundamental role in chemistry due to the principle that similar structures have similar properties. Quantitative applications of this principle include molecular superposition, searching of common structural fragments, similarity searching in chemical databases, diversity (dissimilarity) selection in virtual combinatorial libraries, QSPR, and QSAR [1–5]. Several similarity indices are used in modern theoretical chemistry for defining molecular similarity measures computed from grid or field descriptors, quantum chemistry indices, molecular graph descriptors, or counts of various molecular subgraphs [6–8].

Carbó computed a quantum similarity index R_{AB} by comparing the electron densities, ρ_A and ρ_B, of two molecules *A* and *B* [9–12]:

$$R_{AB} = \frac{\int \rho_A \rho_B \mathrm{d}v}{\left(\int \rho_A^2 \mathrm{d}v\right)^{1/2} \left(\int \rho_B^2 \mathrm{d}v\right)^{1/2}} \quad (1)$$

where the integrations are considered over all space. Willett revealed that the Carbó index formula computed for a property distributed on a grid surrounding a molecule is essentially equivalent to the Cosine coefficient [6,8]. Hodgkin and Richards [13] pointed that in the Carbó index the denominator is a normalizing constant and R_{AB} varies in the range 0 to 1. Such an index of similarity is required to have a value of 1 when the electron density distributions in the two molecules are identical. However, substitution of $\rho_A = a\rho_B$ into the above equation, where *a* is a constant, gives an index of unity. Thus the Carbó index represents the similarity of the shapes of the density distributions but not of the magnitudes as well.

Although originally proposed as a method of comparing molecules in terms of electron density, Hodgkin and Richards [13] proposed to use the formula of the Carbó index with other quantum properties, such as molecular electrostatic potential (MEP) or molecular electrostatic field (MEF). The use of electrostatic potentials and electrostatic fields is particularly attractive since they are better discriminators than charge and problems can be avoided if only values external to a van der Waals volume of the molecule are considered. The electrostatic potentials and electrostatic fields can be calculated over a grid of points surrounding a molecule.

In an attempt to increase the magnitude sensitivity of similarity calculations, Hodgkin and Richards proposed the Hodgkin index [13]:

$$H_{AB} = \frac{2\int \rho_A \rho_B dv}{\int \rho_A^2 dv + \int \rho_B^2 dv} \quad (2)$$

Analogously with the Carbó index, the Hodgkin index can be used with grid–based properties such as MEP or MEF, when its formula is identical with that of the Dice coefficient, as revealed by Willett [6,8]. Richards [14] proposed a linear index for the computation of grid–based molecular similarity descriptors, while Good [15] introduced a related formula to define an exponential similarity index.

Good and Richards [16,17] developed 3D QSAR models from similarity matrices computed with the Carbó, Hodgkin, linear, and exponential similarity indices. The similarity matrices, derived from the shape and electrostatic potential molecular fields, are correlated with the biological activities of the molecules, using either neural networks or PLS models. So and Karplus [18,19] used the Carbó and Hodgkin similarity indices to compute shape and electrostatic similarity matrices. The 3D QSAR, developed with multi–layer feedforward neural networks, used structural descriptors (columns from the similarity matrices) selected with a genetic algorithm. Kubinyi [20] employed SEAL (Steric and Electrostatic ALignment) similarity matrices and SEAL–based fields (hydrophobic, electrostatic and steric) to compute distance and covariance matrices. PLS models derived with all these matrices showed good calibration and prediction results.

The computation of the three–dimensional similarity of two molecules, irrespective of the similarity index or grid property, is a nonlinear optimization process that is computationally very expensive. Various time consuming procedures, such as the Simplex, Monte Carlo, or genetic algorithms were proposed for this task. On the other hand, molecular graph descriptors are particularly efficient in measuring the molecular similarity. Although several applications using graph similarity were published, this class of 2D similarity descriptors was not extensively used in QSPR and QSAR models studies [21–28].

Molecular path sequences and atomic identification numbers were used with success to compute molecular similarity indices based on the Euclidean distance and information theory [21–26]; this approach was found useful in selecting similar compounds on a rational basis and in ordering chemical compounds. Herndon computed similarity matrices based on the subgraphs count and various distance metrics, such as the Hamming, Manhattan, or Euclidean [27]. The similarity matrices of 47 steroids were used to obtain a MLR model for the binding affinity constants for human corticosteroid binding globulin [28]. In this approach, Rum and Herndon proposed a linear code for the molecular structure; the similarity of two molecules is computed by comparing the canonical notations of the two compounds.

Topological indices (TI), representing an important class of structural descriptors, are derived from the molecular graph and encode in numerical form information regarding molecular size,

shape, branching, presence of heteroatoms and multiple bonds. Numerous articles [29–41] present the theory and applications of topological indices in developing QSPR and QSAR models. Although topological indices are important structural descriptors in QSPR and QSAR studies, they were not used in the computation of molecular similarity indices. In this paper we present an application of topological indices for the computation of similarity matrices; the columns of these matrices (describing the pairwise molecular similarity) are used as structural descriptors in MLR equations that model the anticonvulsant activity of a set of phenylacetanilides.

2 MOLECULAR SIMILARITY INDICES

The chemical structure of each molecule A is encoded into a set of n structural descriptors **SD** collected into the vector $\mathbf{X} = \mathbf{X}(A)$, $\mathbf{X}(A) = \{\mathbf{SD}_1, \mathbf{SD}_2, \mathbf{SD}_3, \ldots, \mathbf{SD}_n\}$. For a set **M** of m molecules, $\mathbf{M} = \{A, B, C, \ldots\}$ all the structural descriptors are collected into an $m \times n$ matrix where each row corresponds to a molecule and each column corresponds to a particular structural descriptor. For the computation of the similarity indices the structural descriptors can be standardized by the Z_Score method [6] (autoscaling) that gives variables centered to have zero mean and scaled to unit variance. Each structural descriptor (column from the QSAR data table) is individually autoscaled. For a vector of N variables $\mathbf{Y} = \{y_1, y_2, y_3, \ldots, y_N\}$ the autoscaling is performed with the formula:

$$y_i \leftarrow \frac{y_i - \bar{y}}{s} \tag{3}$$

where $\bar{y}$ is the mean of the variables:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{4}$$

and $s = s(\mathbf{Y})$ is the standard deviation of the vector **Y**:

$$s = \sqrt{V(\mathbf{Y})} \tag{5}$$

i.e. the square root of the variance $V = V(\mathbf{Y})$ of the vector **Y**:

$$V(\mathbf{Y}) = \frac{1}{N-1}\sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{6}$$

From the similarity indices used in 3D QSAR studies, we have selected four for the computation of the similarity matrices from vectors of molecular graph descriptors, namely the Cosine, Dice, Richards, and Good similarity indices.

Cosine similarity index. The Cosine coefficient, C_S, for the similarity between two molecules A and B is given by:

$$C_S(A,B) = \frac{\sum_{i=1}^{n} \mathbf{X}(A)_i \mathbf{X}(B)_i}{\left[\sum_{i=1}^{n} \mathbf{X}(A)_i^2 \sum_{i=1}^{n} \mathbf{X}(B)_i^2\right]^{1/2}} \tag{7}$$

with the property $-1 \le C_S \le 1$. As pointed above, Carbó [9–12] used a form of the Cosine similarity index defined for an integral of electron densities over all space. The Cosine coefficient measures the deviation of two datasets from proportionality.

Dice similarity index. The Dice coefficient, D_S, for the similarity between two vectors of structural descriptors $\mathbf{X}(A)$ and $\mathbf{X}(B)$ is given by:

$$D_S(A,B) = \frac{2\sum_{i=1}^{n} \mathbf{X}(A)_i \mathbf{X}(B)_i}{\sum_{i=1}^{n} \mathbf{X}(A)_i^2 + \sum_{i=1}^{n} \mathbf{X}(B)_i^2} \tag{8}$$

with the property $-1 \le D_S \le 1$.

Richards similarity index. The Richards similarity index is defined by the equation [14]:

$$R_S(A,B) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{\left|\mathbf{X}(A)_i - \mathbf{X}(B)_i\right|}{\max\left(\left|\mathbf{X}(A)_i\right|, \left|\mathbf{X}(B)_i\right|\right)}\right) \tag{9}$$

where $\max(|\mathbf{X}(A)_i|,|\mathbf{X}(B)_i|)$ equals the larger absolute value at the position i of the two vectors $\mathbf{X}(A)$ and $\mathbf{X}(B)$ that collect the structural descriptors for molecules A and B.

Good similarity index. The Good similarity index is an exponential variant of the Richards index [15]:

$$G_S(A,B) = \frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{\left|\mathbf{X}(A)_i - \mathbf{X}(B)_i\right|}{\max\left(\left|\mathbf{X}(A)_i\right|, \left|\mathbf{X}(B)_i\right|\right)}\right) \tag{10}$$

Usually, the above four similarity indices were used with three–dimensional grid descriptors to compute the similarity of steric, electrostatic, or lipophilic fields. The first practical application of 3D similarity indices is for the alignment (superposition) of the molecules, which is an important step in any 3D QSAR study. A second application is the computation of similarity descriptors used to develop a 3D QSAR equation. In this paper we use these four similarity indices to compute molecular similarity matrices from a 2D numerical characterization of the chemical structure represented by a set of topological indices. An important advantage of the computation of the molecular similarity from graph descriptors is the straightforward application of Eqs. (7)–(10) and the elimination of the optimization (maximization) of the similarity indices. The molecular graph operators that are relevant for the present paper are introduced in the next section.

3 MOLECULAR GRAPH DESCRIPTORS

In this paper chemical structures are represented as molecular graphs. By removing all hydrogen atoms from the chemical formula of a compound containing covalent bonds one obtains the hydrogen–depleted (or hydrogen–suppressed) molecular graph of that compound, whose vertices correspond to non–hydrogen atoms and whose edges correspond to covalent bonds [1]. A graph $G = G(V,E)$ is an ordered pair consisting of two sets $V = V(G)$ and $E = E(G)$. Elements of the set $V(G)$ are called vertices and elements of the set $E(G)$, involving the binary relation between the vertices, are called edges. The number of vertices N represents the number of elements in $V(G)$, $N = |V(G)|$, and the number of edges M represents the number of elements in $E(G)$, $M = |E(G)|$. The graph vertices are labeled from 1 to N, $V(G) = \{v_1, v_2, ..., v_N\}$, and the edge connecting vertices v_i and v_j is denoted by e_{ij}.

Using graph theory, an organic compound containing heteroatoms and/or multiple bonds can be represented as a vertex– and edge–weighted molecular graph. A vertex– and edge–weighted (VEW) molecular graph $G = G(V,E,Sy,Bo,Vw,Ew,w)$ consists of a vertex set $V = V(G)$, an edge set $E = E(G)$, a set of chemical symbols of the vertices $Sy = Sy(G)$, a set of topological bond orders of the edges $Bo = Bo(G)$, a vertex weight set $Vw(w) = Vw(w,G)$, and an edge weight set $Ew(w) = Ew(G)$. The elements of the vertex and edge weight sets are computed with the weighting scheme w. Usually, hydrogen atoms are not considered in the molecular graph, and in a VEW graph the weight of a vertex corresponding to a carbon atom is 0, while the weight of an edge corresponding to a carbon–carbon single bond is 1. Also, the topological bond order Bo_{ij} of an edge e_{ij} takes the value 1 for single bonds, 2 for double bonds, 3 for triple bonds and 1.5 for aromatic bonds. Several procedures for computing vertex and edge weights were proposed in the literature [42–46]. From them, five weighting schemes for molecular graphs will be used in this study to compute the vertex Vw and edge Ew parameters [46]: P, the atomic polarizability weighting scheme; E, the atomic electronegativity weighting scheme; R, the atomic radius weighting scheme; A, the atomic mass weighting scheme; AH, the atomic mass weighting scheme that considers the hydrogen atoms.

In a weighting scheme w the vertex Vw and edge Ew parameters are computed from a property p_i associated with every vertex v_i from G, $v_i \in V(G)$, and the topological bond order Bo of all edges from the molecular graph. The vertex parameter $Vw(w)_i$ for the vertex v_i is:

$$Vw(w)_i = 1 - p_C/p_i \tag{11}$$

and the edge parameter $Ew(w)_{ij}$ for the edge between vertices v_i and v_j is:

$$Ew(w)_{ij} = p_C p_C/Bo_{ij} p_i p_j \tag{12}$$

where p_i is the atomic property of vertex v_i, p_j is the atomic property of vertex v_j, and p_C is the atomic property for carbon atom. Several weighting schemes can be obtained when p represents different atomic properties: Z, when p is the atomic number [42]; A, when p is the atomic mass; P, when p is the atomic polarizability; E, when p is the atomic electronegativity; R, when p is the

atomic radius [46]. The atomic properties for the *P*, *E*, and *R* weighting schemes are taken from a recent report [47]; selected values of atomic properties for the computation of vertex and edge parameters are presented in Table 1. Similar equations were used to define the *X* and *Y* weighting schemes, using different sets of values for the atomic radius and electronegativity [45].

Table 1. Selected set of atomic properties used with different weighting schemes: the atomic number *Z*, the atomic mass *A*, the polarizability α_v (Å^3), the atomic radius r_α (Å), and the electronegativity χ_α.

Element	*Z*	*A*	α_v	r_α	χ_α
B	5	10.811	3.03	1.45	2.02
C	6	12.011	1.76	1.21	2.55
N	7	14.007	1.10	1.03	3.12
O	8	15.999	0.802	0.93	3.62
F	9	18.998	0.557	0.82	4.23
Si	14	28.086	5.38	1.75	1.87
P	15	30.974	3.63	1.54	2.22
S	16	32.066	2.90	1.43	2.49
Cl	17	35.453	2.18	1.30	2.82
As	33	74.922	4.31	1.63	2.11
Se	34	78.960	3.77	1.56	2.31
Br	35	79.904	3.05	1.45	2.56
Te	52	127.60	5.5	1.77	2.08
I	53	126.90	4.7	1.68	2.27

The *AH* weighting scheme uses the following equation to define the vertex parameter $Vw(AH)_i$ for the non–hydrogen atom *i*:

$$Vw(AH)_i = 1 - A_C/(A_i + NoH_iA_H) = 1 - 12.011/(A_i + 1.0079NoH_i) \quad (13)$$

The edge parameter $Ew(AH)_{ij}$ for the bond between atoms *i* and *j* is defined with the equation:

$$Ew(AH)_{ij} = A_CA_C/Bo_{ij}(A_i + NoH_iA_H)(A_j + NoH_jA_H) = 12.011\cdot12.011/Bo_{ij}(A_i + 1.0079NoH_i)(A_j + 1.0079NoH_j) \quad (14)$$

where A_C = 12.011 is the atomic mass for carbon, A_H = 1.0079 is the atomic mass for hydrogen, NoH_i is the number of hydrogen atoms bonded to the heavy atom *i*, and NoH_j is the number of hydrogen atoms bonded to the heavy atom *j*.

A molecular graph can be represented as a molecular matrix, such as the adjacency or distance matrix. The structural descriptors used in QSPR and QSAR studies can be computed from a large variety of molecular matrices [1,40]; from the group of the most widely used molecular matrices we mention the adjacency **A**, distance **D**, reciprocal distance **RD** [48–50], distance–path $\mathbf{D_p}$, and reciprocal distance–path $\mathbf{RD_p}$ matrices [59].

Consider the vertex– and edge–weighted graph *G* with *N* vertices and its distance matrix $\mathbf{D}(w) = \mathbf{D}(w,G)$ computed with the weighting scheme *w*. The reciprocal distance matrix of a weighted graph *G* with *N* vertices, $\mathbf{RD}(w) = \mathbf{RD}(w,G)$, is a square $N\times N$ symmetric matrix, whose entries $[\mathbf{RD}(w)]_{ij}$ are equal to the reciprocal of the corresponding value of the $\mathbf{D}(w)$ matrix (*i.e.* $1/[\mathbf{D}(w)]_{ij}$) for non–diagonal elements, and equal to $[\mathbf{D}(w)]_{ii}$ for the diagonal elements:

$$[\mathbf{RD}(w)]_{ij} = \begin{cases} 1/[\mathbf{D}(w)]_{ii} & \text{if } i \neq j \\ [\mathbf{D}(w)]_{ij} & \text{if } i = j \end{cases} \quad (15)$$

The distance–path matrix of the weighted graph G, $\mathbf{D_p}(w) = \mathbf{D_p}(w,G)$, is the square $N \times N$ symmetric matrix whose element $[\mathbf{D_p}(w)]_{ij}$ is defined with the formula:

$$[\mathbf{D_p}(w,G)]_{ij} = [\mathbf{D}(w,G)]_{ij}([\mathbf{D}(w,G)]_{ij} + 1)/2 \quad (16)$$

The reciprocal distance–path matrix of a weighted graph G with N vertices, $\mathbf{RD_p}(w) = \mathbf{RD_p}(w,G)$, is:

$$[\mathbf{RD_p}(w)]_{ij} = \begin{cases} 1/[\mathbf{D_p}(w)]_{ii} & \text{if } i \neq j \\ [\mathbf{D_p}(w)]_{ij} & \text{if } i = j \end{cases} \quad (17)$$

We have to mention that the distance–path and reciprocal distance–path matrices for alkanes and cycloalkanes were introduced by Diudea [51–53]. However, for the computation of the structural descriptors derived from weighted molecular graphs we use the above two equations.

The vertex sum operator VS. Consider the vertex v_i from the VEW graph G with N vertices and the symmetric graph matrix $\mathbf{M}(w) = \mathbf{M}(w,G)$ computed with the weighting scheme w. The vertex sum of the vertex v_i, $\mathbf{VS}(\mathbf{M},w)_i = \mathbf{VS}(\mathbf{M},w,G)_i$, is defined as the sum of the elements in the column i, or row i of the molecular matrix **M** [46]:

$$\mathbf{VS}(\mathbf{M},w,G)_i = \sum_{j=1}^{N}[\mathbf{M}(w)]_{ij} = \sum_{j=1}^{N}[\mathbf{M}(w)]_{ji} \quad (18)$$

The Chi operator. The **Chi** operator [54] is derived from the Kier and Hall connectivity indices [29] by replacing the local invariant δ^v with any other vertex invariant. Consider a vertex structural descriptor $\mathbf{VSD}(\mathbf{M},w) = \mathbf{VSD}(\mathbf{M},w,G)$ that assigns a numerical invariant $\mathbf{VSD}(\mathbf{M},w)_i$ to each vertex v_i from the VEW molecular graph G. The **Chi** operator $\mathbf{Chi}(\mathbf{VSD},\mathbf{M},w) = \mathbf{Chi}(\mathbf{VSD},\mathbf{M},w,G)$ of the graph G is:

$${}^{m}\mathbf{Chi}(\mathbf{VSD},\mathbf{M},w)_t = \sum_{i=1}^{s}\prod_{j=1}^{n}(\mathbf{VSD}(\mathbf{M},w)_j)^{-1/2} \quad (19)$$

where s is the number of connected subgraphs of type t with m edges, n is the number of vertices of the subgraph, and w is the weighting scheme. In this study we use two **VSD** atomic descriptors, namely the valency **val** and the vertex sum **VS**. The valency of the vertex v_i, $\mathbf{val}(w)_i = \mathbf{val}(w,G)_i$, is defined as the sum of the weights $Ew(w)_{ij}$ of all edges e_{ij} incident with vertex v_i:

$$\mathbf{val}(w)_i = \sum_{e_{ij} \in E(G)} Ew(w)_{ij} \quad (20)$$

where w is the weighting scheme used to compute the Ew parameters. Alternatively, the valency of the vertex v_i may be computed as the sum of the non–diagonal elements in the row i, or column i, of the adjacency matrix $\mathbf{A}(w) = \mathbf{A}(w,G)$, of a molecular graph G with N vertices:

$$\mathbf{val}(w)_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} [\mathbf{A}(w)]_{ij} = \sum_{\substack{j=1 \\ j \neq i}}^{N} [\mathbf{A}(w)]_{ji} \tag{21}$$

The set of valency values for all vertices in a graph forms the vector **Val** = **Val**(G) whose ith element represents the valency of the vertex v_i.

The Wiener operator Wi. Consider the vertex– and edge–weighted molecular graph G with N vertices and its symmetric molecular matrix **M**(w) = **M**(w,G) computed with the weighting scheme w. The Wiener operator **Wi**(**M**,w) = **Wi**(**M**,w,G) is [46,54]:

$$\mathbf{Wi}(\mathbf{M}, w, G) = \sum_{i=1}^{N} \sum_{j=i}^{N} [\mathbf{M}(w, G)]_{ij} \tag{22}$$

The hyper–Wiener operator HyWi. Based on a symmetric molecular matrix **M**(w) = **M**(w,G) computed with the weighting scheme w, the hyper–Wiener operator **HyWi**(**M**,w) = **HyWi**(**M**,w,G) is defined with the equation [54]:

$$\mathbf{HyWi}(\mathbf{M}, w, G) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=i}^{N} \left([\mathbf{M}(w)]_{ij}^{\ 2} + [\mathbf{M}(w)]_{ij} \right) \tag{23}$$

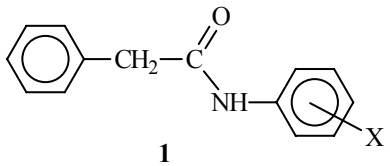
The spectrum operators MinSp and MaxSp. The matrix spectrum operator **Sp**(**M**,w,G) = {x_i, i = 1, 2, ..., N} represents the eigenvalues of the matrix **M**(w) or the roots of the polynomial **Ch**(**M**,w,G,x), **Ch**(**M**,w,G,x) = 0 [46]. The **MinSp**(**M**,w,G) and **MaxSp**(**M**,w,G) spectral operators are equal to the minimum and maximum values of **Sp**(**M**,w,G), respectively [54,58]:

$$\mathbf{MinSp}(\mathbf{M},w,G) = \mathbf{min}\{\mathbf{Sp}(\mathbf{M},w,G)\} \tag{24}$$

$$\mathbf{MaxSp}(\mathbf{M},w,G) = \mathbf{max}\{\mathbf{Sp}(\mathbf{M},w,G)\} \tag{25}$$

4 SIMILARITY QSAR FOR ANTICONVULSANT PHENYLACETANILIDES

Data. The similarity QSAR models are developed for a data set consisting of 30 phenylacetanilides with the general formula **1**, presented in Table 2 together with their anticonvulsant activity –log ED_{50} taken from the literature [55]. The anticonvulsant activity ED_{50} (mol kg^{-1}) is evaluated by the maximal electroshock seizure method in mice.



1

A Hansch–type QSAR analysis of these 30 compounds was performed with the aid of the following descriptors: log P, the octanol–water partition coefficient; σ, the Hammett electronic constant; I_p, an indicator variable which takes the value 1 for p–derivatives and 0 for other

compounds; E_S, the Taft steric constant for *o*–derivatives; *R*, the electronic parameter for *o*–derivatives. The Hansch equations with four, five, and six independent variables are:

$$-\log ED_{50} = 2.280 + 0.264(\log P)^2 + 1.222(\log P) - 0.161\sigma - 0.079I_p$$
$$n = 30 \quad r = 0.700 \quad s = 0.228 \quad F = 5.99$$

$$-\log ED_{50} = 2.311 + 0.290(\log P)^2 + 1.309(\log P) - 0.135\sigma - 0.157I_p + 0.404E_S$$
$$n = 30 \quad r = 0.800 \quad s = 0.195 \quad F = 10.05$$

$$-\log ED_{50} = 2.478 + 0.276(\log P)^2 + 1.229(\log P) - 0.353\sigma - 0.223I_p + 0.278E_S + 0.621R$$
$$n = 30 \quad r = 0.855 \quad s = 0.172 \quad F = 7.83$$

Table 2. Structure of substituted phenylacetanilides, molecular similarity descriptors, and anticonvulsant activity.

No.	X	**S4**	**S10**	**S20**	**S26**	$-\log ED_{50}$,exp	$-\log ED_{50}$,res
1	H	0.99828	0.99978	0.99981	0.99944	3.77	0.28
2	*m*–Me	0.99817	0.99986	0.99999	0.99943	3.75	0.16
3	*m*–Et	0.99756	0.99958	0.99988	0.99900	3.67	0.01
4	*m*–F	1	0.99903	0.99799	0.99962	3.34	–0.23
5	*m*–Cl	0.99836	0.99983	0.99982	0.99951	3.40	–0.24
6	*m*–Br	0.99735	0.99896	0.99896	0.99864	3.32	0.06
7	*m*–I	0.99331	0.99514	0.99515	0.99481	2.64	–0.03
8	*m*–CF_3	0.99820	0.99520	0.99331	0.99655	2.84	–0.29
9	*m*–OH	0.99962	0.99986	0.99935	0.99999	3.58	–0.00
10	*m*–NH_2	0.99903	1	0.99980	0.99985	3.81	0.12
11	*m*–NHMe	0.99960	0.99964	0.99905	0.99982	4.03	0.14
12	*m*–NHEt	0.99944	0.99901	0.99824	0.99935	3.91	0.03
13	*m*–OMe	0.99961	0.99768	0.99623	0.99862	3.55	–0.08
14	*m*–CN	0.99795	0.99973	0.99989	0.99928	3.44	–0.01
15	*m*–NO_2	0.99898	0.99639	0.99466	0.99759	3.62	0.30
16	*m*–COMe	0.99775	0.99950	0.99967	0.99903	3.95	0.27
17	*m*–OAc	0.99604	0.99161	0.98907	0.99349	3.48	0.15
18	*m*–OEt	0.99823	0.99506	0.99308	0.99645	3.42	–0.23
19	*m*–OSO_2Me	0.99562	0.99192	0.98958	0.99357	3.77	–0.03
20	*p*–Me	0.99799	0.99980	1	0.99932	3.26	–0.21
21	*p*–F	0.99999	0.99904	0.99804	0.99961	3.49	0.04
22	*p*–OH	0.99957	0.99988	0.99941	0.99997	3.72	0.03
23	*p*–OMe	0.99965	0.99787	0.99651	0.99873	3.78	0.04
24	*p*–COMe	0.99734	0.99933	0.99963	0.99875	3.51	–0.17
25	*o*–F	0.99999	0.99904	0.99800	0.99964	3.48	0.08
26	*o*–OH	0.99962	0.99985	0.99932	1	3.33	–0.17
27	*o*–NH_2	0.99907	0.99999	0.99976	0.99988	3.40	–0.06
28	*o*–OMe	0.99957	0.99753	0.99602	0.99853	3.43	–0.03
29	*o*–NO_2	0.99880	0.99600	0.99415	0.99731	3.29	0.17
30	*o*–COMe	0.99806	0.99960	0.99965	0.99924	3.41	–0.11

Structural descriptors. In this study we use structural descriptors computed with the following operators: Wiener **Wi**, hyper–Wiener **HyWi**, **Chi**, and the spectral operators **MinSp** and **MaxSp**. The molecular graph descriptors were computed with five weighting schemes [46], namely *P*, *E*, *R*, *A*, and *AH*, using the distance–path $\mathbf{D_p}(w)$ and reciprocal distance–path $\mathbf{RD_p}(w)$ matrices. The list of the 111 structural descriptors used in the QSAR study is: (1) the molecular weight, **MW**; (2) the **Chi** indices: ^{0}**Chi**(**VSD**,**M**,*w*), ^{1}**Chi**(**VSD**,**M**,*w*), ^{2}**Chi**(**VSD**,**M**,*w*), ^{3}**Chi**(**VSD**,**M**,*w*)$_p$,

^{3}**Chi**(**VSD**,**M**,*w*)$_c$, computed with the vertex descriptors **val**(*w*) and **VS**(**M**,*w*); (3) Wiener indices computed with the Wiener operator **Wi**(**M**,*w*); (4) hyper–Wiener indices computed with the hyper–Wiener operator **HyWi**(**RD$_p$**,*w*) using the reciprocal distance–path matrix; the distance–path matrix was not used because it gives too large values for the hyper–Wiener operator; (5) the spectral operators **MinSp**(**M**,*w*) and **MaxSp**(**M**,*w*).

Similarity matrices. Using the 111 structural descriptors, four similarity indices, namely the Cosine, Dice, Richards, and Good similarity indices, are employed for the computation of the similarity matrices. Before computing the similarity matrices, the individual descriptors can be standardized. For each similarity matrix, three experiments were performed: with original, not scaled descriptors; with descriptors scaled in the range [0, 1]; with autoscaled descriptors. All resulting similarity matrices contain 30 rows and 30 columns. The column *j* of a similarity matrix describes the pairwise molecular similarity between molecule *j* and all other molecules in the QSAR set; each column represents a structural descriptor that is used to develop MLR equations that model the anticonvulsant activity of the phenylacetanilides.

Table 3. Coefficients, structural descriptors **SD**$_i$ (i = 1–4), and statistical indices for the best two MLR equations with four independent variables that model the anticonvulsant activity of substituted phenylacetanilides using structural descriptors form the similarity matrices computed with the Cosine, Dice, Richards, and Good indices. The MLR equations have the general form: $\log 1/ED_{50} = a_0 + a_1\mathbf{SD}_1 + a_2\mathbf{SD}_2 + a_3\mathbf{SD}_3 + a_4\mathbf{SD}_4$.

Index	Scale[a]	a_0	a_1	**SD$_1$**	a_2	**SD$_2$**	a_3	**SD$_3$**	a_4	**SD$_4$**	r	s	F
Cosine	n	–88.391	1396.8	**S4**	23800	**S10**	–10025	**S20**	–15083	**S26**	0.8413	0.17	15.1
Cosine	n	–39.463	4447.3	**S13**	–2337.3	**S24**	–22103	**S26**	20041	**S27**	0.8400	0.17	15.0
Cosine	r	3.0999	–6.5609	**S7**	–6.1930	**S8**	5.4709	**S19**	7.5149	**S20**	0.7491	0.21	8.0
Cosine	r	3.4384	–11.120	**S6**	–7.7669	**S8**	5.8386	**S19**	13.131	**S20**	0.7453	0.21	7.8
Cosine	a	3.4850	–2.5011	**S7**	–1.6925	**S8**	–1.3511	**S15**	2.2037	**S19**	0.8024	0.19	11.3
Cosine	a	3.4849	–2.0890	**S7**	–1.5538	**S8**	2.2647	**S19**	1.2876	**S20**	0.7817	0.20	9.8
Dice	n	55.224	–4915.9	**S3**	9123.4	**S11**	–2922.5	**S28**	–1362.4	**S29**	0.8061	0.19	11.6
Dice	n	35.050	–5304.3	**S3**	13638	**S11**	–6819.3	**S13**	–1567.5	**S14**	0.8023	0.19	11.3
Dice	r	3.7097	–8.8579	**S6**	–6.1424	**S8**	4.6256	**S19**	10.215	**S20**	0.7521	0.21	8.1
Dice	r	3.3188	13.216	**S2**	–11.972	**S6**	–6.4669	**S8**	5.8498	**S19**	0.7508	0.21	8.1
Dice	a	3.4953	3.3340	**S1**	–1.6396	**S7**	1.4625	**S19**	–2.0775	**S25**	0.7833	0.20	9.9
Dice	a	3.4328	0.45397	**S3**	–1.4013	**S7**	–1.5621	**S8**	1.2142	**S19**	0.7818	0.20	9.8
Richards	n	–27.253	15.663	**S1**	4.7674	**S16**	2.7624	**S17**	15.085	**S19**	0.8122	0.19	12.1
Richards	n	–22.462	12.779	**S1**	1.9093	**S12**	4.2061	**S16**	13.379	**S19**	0.7951	0.19	10.7
Richards	r	4.0788	–1.3329	**S7**	–1.7179	**S8**	0.60656	**S12**	0.88140	**S16**	0.7883	0.20	10.3
Richards	r	4.2665	–1.3140	**S7**	–1.7262	**S8**	2.0974	**S16**	–0.93004	**S30**	0.7797	0.20	9.7
Richards	a	3.5704	–1.0934	**S7**	–1.6166	**S8**	0.65009	**S16**	0.53826	**S19**	0.8318	0.18	14.0
Richards	a	3.7747	–1.3118	**S7**	–1.5581	**S8**	–0.81551	**S26**	–0.43779	**S29**	0.8295	0.18	13.8
Good	n	11.274	4.1843	**S6**	–10.337	**S7**	–7.2693	**S8**	4.0588	**S16**	0.7913	0.19	10.5
Good	n	–26.432	13.766	**S1**	2.3235	**S12**	5.8791	**S16**	13.696	**S19**	0.7911	0.19	10.5
Good	r	4.9709	–2.0017	**S7**	–2.3389	**S8**	0.78786	**S12**	1.0041	**S16**	0.7987	0.19	11.0
Good	r	5.3629	–2.0431	**S7**	–2.2805	**S8**	2.3661	**S16**	–1.1605	**S30**	0.7926	0.19	10.6
Good	a	5.9327	–0.71027	**S4**	–2.0121	**S7**	–1.8099	**S8**	–0.73567	**S29**	0.7935	0.19	10.6
Good	a	4.6152	–1.6790	**S7**	–2.1219	**S8**	0.75784	**S16**	0.63128	**S19**	0.7930	0.19	10.6

[a] n: not scaled; r: range scaled between 0 and 1; a: autoscaled

QSAR model. The QSAR models were obtained by selecting the best combination of structural descriptors that correspond to certain conditions. This algorithm takes a similarity matrix and from the pool of 30 similarity descriptors generates the best QSAR equations by applying the following steps: (1) All one–parameter correlation equations are computed. All descriptors with a correlation coefficient greater than a threshold, $|r_{min}| > 0.15$, are selected for further use. (2) MLR regression equations are computed with all possible groups of k descriptors selected in step (1). The most significant MLR equations are reported. (3) Step (2) is performed for k from 2 and 5.

Table 4. Coefficients, structural descriptors $\mathbf{SD}_i$ (i = 1–5), and statistical indices for the best two MLR equations with five independent variables that model the anticonvulsant activity of substituted phenylacetanilides using structural descriptors form the similarity matrices computed with the Cosine, Dice, Richards, and Good indices. The MLR equations have the general form: $\log 1/ED_{50} = a_0 + a_1\mathbf{SD}_1 + a_2\mathbf{SD}_2 + a_3\mathbf{SD}_3 + a_4\mathbf{SD}_4 + a_5\mathbf{SD}_5$.

Index	S[a]	a_0	a_1	$\mathbf{SD_1}$	a_2	$\mathbf{SD_2}$	a_3	$\mathbf{SD_3}$	a_4	$\mathbf{SD_4}$	a_5	$\mathbf{SD_5}$	r	s	F
Cosine	n	–82.057	–5197.4	**S3**	27654	**S10**	–14624	**S22**	4947.3	**S23**	–12690	**S26**	0.8615	0.17	13.8
Cosine	n	–83.187	–4703.9	**S3**	25018	**S10**	4285.9	**S13**	–9328.8	**S22**	–15179	**S26**	0.8611	0.17	13.8
Cosine	r	2.8204	–5.0924	**S7**	–8.1697	**S8**	5.9184	**S19**	20.499	**S22**	–13.031	**S25**	0.8133	0.19	9.4
Cosine	r	2.6897	–8.7206	**S6**	–9.7796	**S8**	6.8259	**S19**	25.849	**S22**	–13.574	**S25**	0.8128	0.19	9.3
Cosine	a	3.4833	11.909	**S1**	1.7526	**S8**	4.3125	**S13**	6.5682	**S16**	6.1041	**S19**	0.8493	0.17	12.4
Cosine	a	3.4808	–31.700	**S3**	–1.9399	**S8**	–4.4082	**S9**	47.166	**S11**	–32.381	**S13**	0.8472	0.17	12.2
Dice	n	51.910	–3438.1	**S3**	–82.869	**S8**	10700	**S11**	–5218.5	**S13**	–2033.3	**S14**	0.8309	0.18	10.7
Dice	n	54.253	–4189.0	**S3**	–38.424	**S7**	7824.0	**S11**	–2445.8	**S28**	–1225.6	**S29**	0.8285	0.18	10.5
Dice	r	0.17410	45.257	**S1**	–9.8488	**S6**	8.4885	**S19**	9.8679	**S23**	–41.699	**S26**	0.8286	0.18	10.5
Dice	r	0.99701	34.634	**S1**	–5.0754	**S7**	6.6589	**S19**	7.4544	**S23**	–34.911	**S26**	0.8145	0.19	9.5
Dice	a	3.4894	2.2011	**S1**	–1.8243	**S7**	–1.0581	**S8**	1.5030	**S19**	–1.6456	**S25**	0.8254	0.18	10.3
Dice	a	3.4806	3.9353	**S1**	–1.6453	**S7**	1.7071	**S19**	–2.6696	**S25**	0.37711	**S28**	0.8135	0.19	9.4
Richards	n	–22.492	13.819	**S1**	8.4152	**S11**	3.7387	**S17**	12.722	**S19**	–6.3204	**S28**	0.8194	0.19	9.8
Richards	n	–28.498	15.275	**S1**	1.2637	**S2**	4.3818	**S16**	3.1213	**S17**	15.740	**S19**	0.8183	0.19	9.7
Richards	r	4.2673	–1.1731	**S4**	–1.6137	**S7**	–1.6980	**S8**	1.1104	**S10**	1.3446	**S16**	0.8179	0.19	9.7
Richards	r	4.2176	–1.3595	**S7**	–1.5093	**S8**	1.4011	**S11**	0.82144	**S16**	–1.2194	**S28**	0.8150	0.19	9.5
Richards	a	3.5613	0.58982	**S6**	–1.6190	**S7**	–1.5770	**S8**	0.68508	**S16**	0.57048	**S19**	0.8482	0.17	12.3
Richards	a	3.7240	–0.91898	**S4**	–1.3081	**S7**	–1.5589	**S8**	1.1572	**S10**	–0.88832	**S27**	0.8444	0.17	11.9
Good	n	11.372	5.7891	**S6**	–11.135	**S7**	–8.1684	**S8**	8.5086	**S16**	–4.4612	**S30**	0.8177	0.19	9.7
Good	n	–30.688	15.070	**S1**	2.9583	**S3**	4.5937	**S16**	2.9444	**S17**	15.157	**S19**	0.8144	0.19	9.5
Good	r	5.5134	–1.4742	**S4**	–2.3986	**S7**	–2.3354	**S8**	1.3116	**S10**	1.4569	**S16**	0.8275	0.18	10.4
Good	r	5.5437	1.0090	**S2**	–0.99854	**S4**	–2.5549	**S7**	–2.3565	**S8**	1.4112	**S16**	0.8227	0.18	10.1
Good	a	4.0933	0.61235	**S2**	–1.8269	**S7**	–1.9259	**S8**	0.92862	**S16**	0.88068	**S19**	0.8170	0.19	9.6
Good	a	5.7841	–1.3736	**S4**	–2.1027	**S7**	–1.7042	**S8**	0.84170	**S10**	–0.63587	**S29**	0.8142	0.19	9.4

[a] Scale: n: not scaled; r: range scaled between 0 and 1; a: autoscaled

Results. In Tables 3 and 4 we present the statistical indices (r, correlation coefficient, s, standard deviation, and F, Fisher test) for the best two QSAR models with 4 and 5 similarity descriptors, together with the coefficients of the MLR equations. An inspection of these results indicates a significant difference of the results obtained with the four similarity indices and scaling methods. The best QSAR model having four similarity descriptors, with r = 0.8413, s = 0.17, F = 15.1, is obtained using the Cosine and unscaled topological indices and contains the descriptors **S4**, **S10**, **S20**, and **S26**. The similarity descriptor **S4**, representing the 4–th column from the similarity matrix, represents the similarity of the 30 compounds with molecule **4**, with X = *m*–F. The remaining three descriptors represent the similarity of the investigated molecules with compound **10** (X = *m*–NH_2), **20** (X = *p*–Me), and **26** (X = *o*–OH). We have to point that the most active compound, **11**, was not selected as a similarity standard. Also, in the four molecules that represent the similarity standard

one can recognize the three substitution modes from the set of compounds, *i.e*. *ortho*, *meta*, and *para*. The statistical quality of this equation is comparable with that of the Hansch model with six parameters. However, an important advantage of this approach that uses theoretical parameters is the possibility to compute the structural descriptors for all organic compounds; a Hansch analysis uses empirical substituent constants and frequently, when the proper value is missing, one uses approximations or values from similar groups. The values of the similarity descriptors **S4**, **S10**, **S20**, and **S26** are given in Table 2, together with the residual values ($\log 1/ED_{50,res} = \log 1/ED_{50,exp} - \log 1/ED_{50,calc}$). An inspection of the residual values shows that all biological activities are well estimated, with no statistical outliers.

Table 5. Structure of substituted phenylacetanilides, molecular similarity descriptors, and anticonvulsant activity.

No	X	**S3**	**S10**	**S22**	**S23**	**S26**	$-\log ED_{50,exp}$	$-\log ED_{50,res}$
1	H	0.99941	0.99978	0.99946	0.99661	0.99944	3.77	0.20
2	*m*–Me	0.99985	0.99986	0.99949	0.99670	0.99943	3.75	0.12
3	*m*–Et	1	0.99958	0.99912	0.99622	0.99900	3.67	0.07
4	*m*–F	0.99756	0.99903	0.99957	0.99965	0.99962	3.34	–0.26
5	*m*–Cl	0.99957	0.99983	0.99952	0.99685	0.99951	3.40	–0.15
6	*m*–Br	0.9985	0.99896	0.99854	0.99539	0.99864	3.32	0.12
7	*m*–I	0.99438	0.99514	0.99455	0.99069	0.99481	2.64	–0.03
8	*m*–CF_3	0.99319	0.99520	0.99646	0.99929	0.99655	2.84	–0.21
9	*m*–OH	0.99905	0.99986	0.99999	0.99877	0.99999	3.58	–0.03
10	*m*–NH_2	0.99958	1	0.99988	0.99787	0.99985	3.81	0.15
11	*m*–NHMe	0.99895	0.99964	0.99986	0.99913	0.99982	4.03	0.14
12	*m*–NHEt	0.99829	0.99901	0.99941	0.99943	0.99935	3.91	–0.02
13	*m*–OMe	0.99592	0.99768	0.99858	0.99998	0.99862	3.55	–0.04
14	*m*–CN	0.99990	0.99973	0.99933	0.99660	0.99928	3.44	–0.09
15	*m*–NO_2	0.99441	0.99639	0.99750	0.99972	0.99759	3.62	0.28
16	*m*–COMe	0.99988	0.99950	0.99913	0.99666	0.99903	3.95	0.29
17	*m*–OAc	0.98882	0.99161	0.99337	0.99785	0.99349	3.48	0.10
18	*m*–OEt	0.99284	0.99506	0.99640	0.99938	0.99645	3.42	–0.17
19	*m*–OSO_2Me	0.98927	0.99192	0.99337	0.99689	0.99357	3.77	–0.08
20	*p*–Me	0.99988	0.99980	0.99941	0.99651	0.99932	3.26	–0.18
21	*p*–F	0.99761	0.99904	0.99959	0.99967	0.99961	3.49	–0.06
22	*p*–OH	0.99912	0.99988	1	0.99874	0.99997	3.72	–0.04
23	*p*–OMe	0.99622	0.99787	0.99874	1	0.99873	3.78	0.14
24	*p*–COMe	0.99989	0.99933	0.99890	0.99622	0.99875	3.51	–0.14
25	*o*–F	0.99756	0.99904	0.99957	0.99959	0.99964	3.48	0.16
26	*o*–OH	0.99900	0.99985	0.99997	0.99873	1	3.33	–0.23
27	*o*–NH_2	0.99951	0.99999	0.99988	0.99789	0.99988	3.40	–0.07
28	*o*–OMe	0.99568	0.99753	0.99845	0.99992	0.99853	3.43	–0.00
29	*o*–NO_2	0.99385	0.99600	0.99716	0.99955	0.99731	3.29	0.14
30	*o*–COMe	0.99981	0.99960	0.99928	0.99698	0.99924	3.41	–0.11

As already mentioned, the results reported in Table 3 show an important dependence of the statistical quality of the QSAR model on the similarity index used to compute the similarity matrix and on the scaling of the topological indices. The best results are obtained with the Cosine index computed with unscaled data ($s = 0.17$), followed by the Richards index computed with autoscaled

data (s = 0.18). Lower quality results are obtained with the Dice and Good indices.

From the results presented in Table 4 one can see that the best QSAR model containing five similarity descriptors, with r = 0.8615, s = 0.17, F = 13.8, is obtained with the Cosine index and unscaled topological indices; we have to mention that this similarity matrix gives also the best results in QSAR models with four similarity descriptors. The five similarity descriptors (**S3**, **S10**, **S22**, **S23**, and **S26**) represent the similarity of the investigated molecules with compound **3** (X = *m*–Et), **10** (X = *m*–NH_2), **22** (X = *p*–OH), **23** (X = *p*–OMe), and **26** (X = *o*–OH). The values of the similarity descriptors **S3**, **S10**, **S22**, **S23**, and **S26** are given in Table 5, together with the residual values for the 30 molecules in the QSAR set. Overall, the statistical indices of the QSAR equations from Table 4 do not present a significant improvement when compared with the results from Table 3, obtained with four descriptors. A significant influence of the similarity matrix on the statistical quality of the QSAR model is observed also from Table 4, with best results are obtained with the Cosine index computed with unscaled data, followed by the Cosine index computed with autoscaled data and the Richards index computed with autoscaled data. More QSAR models have to be investigated in order to identify the best way of computing the similarity matrix. Until then, the single advice is to experiment with similarity matrices computed with several similarity indices and to use the three scaling procedures for the molecular graph descriptors.

5 CONCLUSIONS

Molecular similarity plays a fundamental role in structure–activity relationship due to the principle that similar structures have similar biological activities. A quantitative model that implements this principle is represented by the use of similarity matrices. The large majority of QSAR equations developed from similarity matrices are based on three–dimensional field descriptors computed on a grid. The computation of the three–dimensional similarity of two molecules, irrespective of the similarity index or grid property, is a nonlinear optimization process that is computationally very expensive. Various time consuming procedures, such as the Simplex, Monte Carlo, or genetic algorithms were proposed for this task. Therefore, the 3D similarity matrices are obtained in a computationally intensive process, and offer only a limited and partial measure of the molecular similarity, because only one field is used in their calculation.

On the other hand, molecular graph descriptors are particularly efficient in measuring the molecular similarity. Although several applications using graph similarity were published, this class of 2D similarity descriptors was not extensively used in QSPR and QSAR studies. In this paper we have presented a new application of topological indices for the computation of similarity matrices in which the numerical descriptors are obtained by computing a similarity index between a molecule and every other from the data set. In order to analyze a data set of N molecules, in the $N \times N$ similarity matrix the column j describes the pairwise molecular similarity between molecule j and

all other molecules in the QSAR set; each column represents a structural descriptor that is used to develop QSAR models.

In this paper we have described the molecular structure with 111 graph descriptors computed from six operators, namely the Wiener **Wi**, hyper–Wiener **HyWi**, **Chi**, and the spectral operators **MinSp** and **MaxSp**. Two molecular matrices (the distance–path $\mathbf{D_p}$ and reciprocal distance–path $\mathbf{RD_p}$) and five weighting schemes (*P*, *E*, *R*, *A*, and *AH*) [46,56,57] were used to compute the topological indices. The similarity matrices were computed using four similarity indices, namely the Cosine, Dice, Richards, and Good similarity indices. For each similarity matrix, three experiments were performed: with original, not scaled descriptors; with descriptors scaled in the range [0, 1]; with autoscaled descriptors.

The similarity matrices were used to develop multilinear regression QSAR models of the anticonvulsant activity of 30 phenylacetanilides; the significant columns from the similarity matrices were selected as independent variables in QSAR equations. The best models with 4 and 5 descriptors were obtained with similarity indices computed with the Cosine index from unscaled topological indices. The results obtained show that similarity matrices derived from molecular graph descriptors represent an efficient model for the investigation of quantitative structure–activity relationships.

6 REFERENCES

[1] O. Ivanciuc, Graph Theory in Chemistry. In: *Handbook of Chemoinformatics*, Ed.: J. Gasteiger. Wiley–VCH, 2003, pp. 103–138.

[2] M. A. Johnson and G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, John Wiley & Sons, New York, 1990.

[3] K. Sen, *Molecular Similarity*, Vols. 1 and 2, Springer–Verlag, Berlin, 1995.

[4] R. Carbó–Dorca and P. G. Mezey, *Advances in Molecular Similarity*, Vol. 1, JAI Press, Greenwich, CT, 1996.

[5] M. A. Johnson, *J. Math. Chem.* **1989**, *3*, 117–145.

[6] D. B. Turner, P. Willett, A. M. Ferguson, and T. W. Heritage, *SAR QSAR Environ. Res.* **1995**, *3*, 101–130.

[7] M. I. Skvortsova, I. I. Baskin, I. V. Stankevich, V. A. Palyulin, and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 785–790.

[8] P. Willett, J. M. Barnard, and G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

[9] R. Carbó, L. Leyda, and M. Arnau, *Int. J. Quantum Chem.* **1980**, *17*, 1185–1189.

[10] R. Carbó and L. Domingo, *Int. J. Quantum Chem.* **1987**, *32*, 517–545.

[11] R. Carbó and B. Calabuig, Molecular Similarity and Quantum Chemistry. In: *Concepts and Applications of Molecular Similarity*, Eds.: M. A. Johnson and G. M. Maggiora, John Wiley & Sons: New York, 1990, pp. 147–171.

[12] R. Carbó–Dorca, E. Besalú, L. Amat, and X. Fradera, Quantum Molecular Similarity Measures: Concepts, Definitions, and Applications to Quantitative Structure–Property Relationships. In: *Advances in Molecular Similarity*, Eds.: R. Carbó–Dorca and P. G. Mezey, JAI Press, Greenwich, CT, 1996, Vol 1, pp. 1–42.

[13] E. E. Hodgkin and W. G. Richards, *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1987**, *14*, 105–110.

[14] C. A. Reynolds, C. Burt, and W. G. Richards, *Quant. Struct.–Act. Relat.* **1992**, 11, 34–35.

[15] A. C. Good, *J. Mol. Graphics* **1992**, *10*, 144–151.

[16] A. C. Good, S.–S. So, and W. G. Richards, *J. Med. Chem.* **1993**, *36*, 433–438.

[17] A. C. Good, S.J. Peterson, and W. G. Richards, *J. Med. Chem.* **1993**, *36*, 2929–2937.

[18] S.–S. So and M. Karplus, *J. Med. Chem.* **1997**, *40*, 4347–4359.

[19] S.–S. So and M. Karplus, *J. Med. Chem.* **1997**, *40*, 4360–4371.

[20] H. Kubinyi, F. A. Hamprecht, and T. Mietzner, *J. Med. Chem.* **1998**, *41*, 2553–2564.

[21] M. Randić and C. L. Wilkins, *J. Chem.Inf. Comput. Sci.* **1979**, *19*, 31–37.

[22] M. Barysz, N. Trinajstić, and J. V. Knop, *Int. J. Quantum Chem.: Quantum. Chem. Symp.* **1983**, *17*, 441–451.

[23] B. Jerman–Blažič, I. Fabič, and M. Randić, *J. Comput. Chem.* **1986**, *7*, 176–188.

[24] M. Randić, B. Jerman–Blažič, D. H. Rouvray, P. G. Seybold, and S. C. Grossman, *Int. J. Quantum Chem.: Quantum. Biol. Symp.* **1987**, *14*, 245–260.

[25] B. Jerman–Blažič, I. Fabič–Petrač, and M. Randić, *Chemom. Intell. Lab. Syst.* **1989**, *6*, 49–63.

[26] M. Randić, *J. Chem.Inf. Comput. Sci.* **1996**, *36*, 1092–1097.

[27] S. H. Bertz and W. C. Herndon. The Similarity of Graphs and Molecules. In: *Artificial Intelligence Applications in Chemistry*, ACS Symposium Series 306, Eds.: T. H. Pierce and B. A. Hohne. American Chemical Society, Washington, 1986, pp. 169–175.

[28] G. Rum and W. C. Herndon, *J. Am. Chem. Soc.* **1991**, *113*, 9055–9060.

[29] L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure–Activity Analysis*, Research Studies Press, Letchworth, 1986.

[30] O. Ivanciuc, T. Ivanciuc, D. Cabrol–Bass, and A. T. Balaban, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 631–643.

[31] O. Ivanciuc, T. Ivanciuc, D. Cabrol–Bass, and A. T. Balaban, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 732–743.

[32] O. Ivanciuc and A. T. Balaban, Graph Theory in Chemistry. In: *The Encyclopedia of Computational Chemistry*, Eds.: P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, and P. R. Schreiner. John Wiley & Sons, Chichester, 1998, pp. 1169–1190.

[33] A. T. Balaban and O. Ivanciuc, Historical Development of Topological Indices. In: *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds.: J. Devillers and A. T. Balaban. Gordon and Breach Science Publishers, The Netherlands, 1999, pp. 21–57.

[34] O. Ivanciuc and A. T. Balaban, The Graph Description of Chemical Structures. In: *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds.: J. Devillers and A. T. Balaban. Gordon and Breach Science Publishers, The Netherlands, 1999, pp. 59–167.

[35] O. Ivanciuc, T. Ivanciuc, and A. T. Balaban, Vertex– and Edge–Weighted Molecular Graphs and Derived Structural Descriptors. In: *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds.: J. Devillers and A. T. Balaban. Gordon and Breach Science Publishers, The Netherlands, 1999, pp. 169–220.

[36] O. Ivanciuc and T. Ivanciuc, Matrices and Structural Descriptors Computed from Molecular Graph Distances. In: *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds.: J. Devillers and A. T. Balaban. Gordon and Breach Science Publishers, The Netherlands, 1999, pp. 221–277.

[37] O. Ivanciuc and A. T. Balaban, *Rev. Roum. Chim.* **1999**, *44*, 479–489.

[38] O. Ivanciuc and A. T. Balaban, *Rev. Roum. Chim.* **1999**, *44*, 539–547.

[39] O. Ivanciuc, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1412–1422.

[40] O. Ivanciuc, *Rev. Roum. Chim.* **1999**, *44*, 519–528.

[41] O. Ivanciuc, T. Ivanciuc, and A. T. Balaban, *ACH Models Chem.* **2000**, *137*, 57–82; A. T. Balaban, D. Mills, O. Ivanciuc, and S. C. Basak, *Croat. Chem. Acta* **2000**, *73*, 923–941.

[42] M. Barysz, G. Jashari, R. S. Lall, V. K. Srivastava, and N. Trinajstić, On the Distance Matrix of Molecules Containing Heteroatoms. In: *Chemical Applications of Topology and Graph Theory*, Ed.: R. B. King. Elsevier, Amsterdam, 1983, pp. 222–227.

[43] A. T. Balaban, *MATCH* (*Commun. Math. Chem.*) **1986**, *21*, 115–122.

[44] A. T. Balaban and O. Ivanciuc. FORTRAN 77 Computer Program for Calculating the Topological Index J for Molecules Containing Heteroatoms. *MATH/CHEM/COMP 1988*, Proceedings of an International Course and Conference on the Interfaces Between Mathematics, Chemistry and Computer Sciences, Dubrovnik, Yugoslavia,

20–25 June 1988, Ed.: A. Graovac. *Studies in Physical and Theoretical Chemistry*, Vol. 63, pp. 193–211, Elsevier, Amsterdam, 1989.

[45] O. Ivanciuc, T. Ivanciuc, and A. T. Balaban, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 395–401.

[46] O. Ivanciuc, *Rev. Roum. Chim.* **2000**, *45*, 289–301.

[47] J. K. Nagle, *J. Am. Chem. Soc.* **1990**, *112*, 4741–4747.

[48] O. Ivanciuc, *Rev. Roum. Chim.* **1989**, *34*, 1361–1368.

[49] O. Ivanciuc, T.–S. Balaban, and A. T. Balaban, *J. Math. Chem.* **1993**, *12*, 309–318.

[50] M. V. Diudea, O. Ivanciuc, S. Nikolić, and N. Trinajstić, *MATCH* (*Commun. Math. Comput. Chem.*) **1997**, *35*, 41–64.

[51] M. V. Diudea, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 535–540.

[52] M. V. Diudea, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 833–836.

[53] O. Ivanciuc, M. V. Diudea, and P. V. Khadikar, *Ind. J. Chem.* **1998**, *37A*, 574–585.

[54] O. Ivanciuc, *Rev. Roum. Chim.* **2001**, *46*, 243–253.

[55] C. Yamagami, N. Takao, M. Tanaka, K. Horisaka, S. Asada, T. Fujita, *Chem. Pharm. Bull.* **1984**, *32*, 5003–5009.

[56] O. Ivanciuc, S. L. Taraviras, and D. Cabrol–Bass, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 126–134.

[57] S. Taraviras, O. Ivanciuc, and D. Cabrol–Bass, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1128–1146.

[58] O. Ivanciuc, Topological Indices. In: *Handbook of Chemoinformatics*, Ed.: J. Gasteiger. Wiley–VCH, 2003, pp. 981–1003.

[59] O. Ivanciuc, *Rev. Roum. Chim.* **2001**, *46*, 543–552.