

# TOPOLOGICAL SIMILARITY / DISSIMILARITY INDICATORS: APPLICATION TO CYTOCHROME P450 INHIBITION BY ALCOHOLS

Dan Dragos,<sup>1</sup> Alina Heghes,<sup>1</sup> Mihai Medeleanu,<sup>2</sup> Vicentiu Vlaia<sup>1</sup>,  
Corina Seiman,<sup>3</sup> Adriana Kaycsa<sup>1</sup>, Dan Ciubotariu<sup>1</sup>

## REZUMAT

Au fost examinați 10 indicatori topologici ai similarității moleculare derivată din analiza de grupare și bazându-se pe măsurători de distanță binară. Coeficienții de corelare între aceste măsurători de similaritate și activitate moleculară a seriilor de 22 alcooli alkilați ce inhibă competitiv p-hidroxilarea microzomală a anilinei au avut valori între 0,032 și 0,907. Ne propunem de asemenea măsurarea similarității/disimilarității topologice (MSDT) pe baza metodei originale a diferenței topologice minime, denumită MSD. Pentru aceasta moleculele bioactive sunt tratate ca reprezentări grafice cu depleție de hidrogen, iar congruențele geometrice ale moleculei cu cea mai mare activitate față de moleculele comparate sunt efectuate căutând suprapunerea maximă. MSDT a fost obținută prin normalizarea valorilor MSD, pe un palier de valori [0,1]. O valoare a MSDT apropiată de 1 implică o similaritate topologică înaltă, în timp ce disimilaritatea este maximă dacă MSDT=0. Rezultatele obținute prin analiza QSAR a aceleiași serii de alcooli cu valorile MSDT au fost următoarele: coeficientul de corelare=0,944, coeficientul de validare încrucișată  $r^2_{cv}=0,736$ . Aceste date susțin ipoteza că legarea alcoolilor de situl enzimatic al citocromului P-450 are loc printr-un mecanism în două etape, așa numitul mecanism al fermoarului.

**Cuvinte cheie:** măsurarea similarității/disimilarității topologice (MSDT), QSAR, citocrom P-450, MSD

## ABSTRACT

10 topological indicators of molecular similarity derived from cluster analysis, and based on binary distance measures were examined. The correlation coefficients ( $r$ ) between these similarity measures and the biological activity ( $A$ ) of a series of 22 alkyl alcohols that competitively inhibit the microsomal p-hydroxylation of aniline ( $A$  is  $pI_{50}$ , that is, the logarithm of reciprocal concentration that causes 50% inhibition) are in the range of 0.032 – 0.907. We also propose a topological similarity/dissimilarity measure (TSDM) developed on the basis of original Minimal Topological Difference (MTD) method, named MSD. Thus, the bioactive molecules are treated as hydrogen depleted graphs and the geometrical congruencies of the molecule with the highest activity vs. compared molecules are performed seeking the maximal superposition. The TSDM was obtained by normalization of MSD values. The range of variation is [0,1]. A value of TSDM close to 1 implies a high topological similarity. The degree of dissimilarity is maximum if TSDM=0. The results obtained by QSAR analysis of the same series of alcohols with TSDM values were: the correlation coefficient  $r=0.944$  and the cross-validation coefficient was  $r^2_{cv}=0.736$ . This supports the hypothesis that the binding of alcohols to the enzymatic site of cytochrome P-450 takes place in a two-stage mechanism, the so-called “zipper” mechanism.

**Key Words:** topological similarity/dissimilarity measure (TSDM), QSAR, cytochrom P-450, MSD

## INTRODUCTION

In computer aided molecular design, very often no information about the receptor site is available. In such circumstances, the rational development of new drugs can only take place by considering the properties of known ligands that are selective for the site and mimicking their properties in an attempt to produce new leads. The determination and use of molecular similarity is currently of intense interest in drug design.<sup>1</sup>

The main difficulties of design, based on similarity, are associated with recognizing the important common properties of several molecules from a structurally dissimilar set. The structures must be superposed in their optimal orientation, taking into account their individual flexibility. Nevertheless, numerous alternative superpositions may be possible with comparable similarity values.

The notion of similarity that we have is strongly dependent on the current use to which similarity is put. Basically, molecules can be described in three fundamentally distinct ways: by their molecular graphs, by their atom positions and by their molecular fields. The quantitative similarity measures can be developed for each of the above molecular characteristics.

It is an important goal to find a definition of molecular similarity, which is reliable and simple enough to allow rapid comparisons in large databases. Easy selection of a subset of molecules having a property falling within a certain prescribed range is

<sup>1</sup> Faculty of Pharmacy, Victor Babes University of Medicine and Pharmacy, Timisoara, <sup>2</sup> Faculty of Chemical Engineering, Technical University, Timisoara, <sup>3</sup> Faculty of Chemistry-Biology-Geography, West University, Timisoara,

Correspondence to:

Dan Ciubotariu, Faculty of Pharmacy, Victor Babes University of Medicine and Pharmacy, Piata Eftimie Murgu 2, Timisoara, Tel: +40.256.201883; +40.722.579.220; E-mail: dciubotariu@mail.dnttm.ro

Received for publication: Sep. 17, 2003. Revised: Jan. 17, 2004.

especially useful in the early phase of computer-aided molecular design. This justifies our effort to analyze several measures of similarity (MS) quantifying molecular graphs, and to develop a method for quantitative treatment of similarity and dissimilarity of molecular graphs, TSDM. MS and TSDM have been applied to the study of detoxification process performed in human bodies within cytochrome P-450.

The cytochrome P450 family of heme-containing proteins is the most important enzyme system in terms of phase I-catalyzed oxidative biotransformations that result in the formation of biologically reactive metabolites<sup>2</sup>. For comprehensive analyses of specific aspects of the P450 enzymes, see refs.<sup>3-5</sup>.

The P450 enzymes are hemoproteins, which catalyze the monooxidation of a wide variety of structurally unrelated compounds, including endogenous steroid and fatty acids and an essentially unlimited number of lipophilic xenobiotics. The drug-metabolizing P450 enzymes are located primarily in the cells' endoplasmic reticulum.<sup>6</sup>

The amino acid sequences of more than 300 P450 enzymes have been deduced by recombinant DNA techniques and additional sequences; particularly those of bacterial, insect and plant origin are reported monthly.<sup>7</sup> These sequences are now the basis for classifying and naming P450 enzymes.<sup>8</sup>

In this paper we are only interested in quantifying the topological structure of alcohol molecules by means of similarity / dissimilarity descriptors and analyzing their capabilities to model the experimentally determined bioactivity.

## METHODS

### Theory

Let  $\mathbf{M}$  be a bioactive molecule (the ligand L), and  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$  the  $m$  features (called parameters) of  $\mathbf{M}$ . Each of these parameters has an interpretation in the chemical and biological universe  $\mathbf{U}$ . They may be *quantitative*, *i.e.*  $\mathbf{x}$  is isomorphic to a real number (for example, electronic or steric structural constants), *qualitative*, when  $\mathbf{x}$  may materialize as one of the elements of a finite set  $\{\mathbf{x}\}$  and the interpretation in  $\mathbf{U}$  is usually a name or a qualifier (for instance a colour, a quality of a drug such as the intensity of bioactivity, etc.), and *binary*; in this last case the variable parameter may take two states, usually called 0 or 1. If an arithmetic operation may be performed on  $\mathbf{x}$ , they are called *binary quantitative* (*e.g.*, topological description of a molecule by its chemical graph, using the adjacency matrix as invariant), otherwise *binary qualitative* (for

example, the variables of indicators used in some QSAR studies).

The measurements, *i.e.* structural and other descriptors, may be performed on  $n$  ligand molecules  $\mathbf{M}_i$  ( $1 = i = n$ ). Let  $\mathbf{X}$  be the "space" of these variables  $\mathbf{M}$  and of these objects or points  $M$ . Let  $E = \{M_1, \dots, M_n\}$  be the set of  $n$  of these molecules.

Let  $\mathbf{Y}$  be the corresponding variable. It is the Cartesian product of  $m$  elementary variables  $\mathbf{x}_j$ . The  $Y_j$  value of  $\mathbf{Y}_j$  is a list of the measurement on the  $n$  objects:

$$Y_j = (x_{1j}, x_{2j}, \dots, x_{nj})$$

		$Y_1$	$Y_2$	$\dots$	$Y_j$	$\dots$	$Y_m$
$X_1$		$x_{11}$	$x_{12}$	$\dots$	$x_{1j}$	$\dots$	$x_{1m}$
$X_2$		$x_{21}$	$x_{22}$	$\dots$	$x_{2j}$	$\dots$	$x_{2m}$
$\dots$		.....					
$X_i$		$x_{i1}$	$x_{i2}$	$\dots$	$x_{ij}$	$\dots$	$x_{im}$
$\dots$		.....					
$X_n$		$x_{n1}$	$x_{n2}$	$\dots$	$x_{nj}$	$\dots$	$x_{nm}$

The above table gives the complete results of  $m$  measurements (structural descriptors) on  $n$  objects (L molecules). Each elementary variable  $\mathbf{x}_j$  ( $j=1,m$ ) may take a value in a set of possible values. In our case this set is finite, *i.e.* the number of topological features of chemical graphs equals the number of non-hydrogen atoms. Similarity measure, SM, (or dissimilarity measure, DM) gives a numerical value to the most general notion of closeness (or farness) between two molecules  $M_p$  and  $M_q$  based on their topological descriptors.  $SM(M_p, M_q)$  or  $DM(M_p, M_q)$  is a real valued, symmetric function, whose domain is the set of possible  $M \times M$ . Usually a high value of SM (or DM) indicates high similarity or closeness (or dissimilarity and farness).<sup>2</sup>

A distance  $d$  is a real valued, symmetric function, obeying three axioms and whose domain is again  $M \times M$ . Let  $d(M_p, M_q)$  be such a function. It satisfies the following three axioms:

(i) Reflexivity

$$d(M_p, M_p) = 0$$

(ii) Symmetry

$$d(M_p, M_q) = d(M_q, M_p)$$

(iv) Triangle inequality

$$d(M_p, M_q) = d(M_p, M_j) + d(M_j, M_q)$$

A distance is a particular dissimilarity function for which a value of  $d$  close to 0 implies a high similarity (or closeness).

Similarity and distance measures may be normalized, by dividing the numerical function by the maximum value. Then, the range of variation is [0,1].

The fundamental purpose of a distance as similarity (or dissimilarity) measure is to induce an order on the set of couples  $(M_i, M_j)$  for any  $i$  or  $j$ , with  $i \neq j$ . An infinite number of distances (or similarities) may be used because they may induce the same order. In fact, simplicity and computational capability is the unique guide in this field. In general, one tries to choose a function, which seems to be reasonable, according to what one knows about the properties of  $\mathbf{U}$ .<sup>9</sup> Obviously, many such choices can be made.

The measurements  $x_i$  may be interpreted in  $\mathbf{U}$  as numbers. Among several existing dissimilarity measures, we mention here only

$$\text{Minkowski metric} \quad d(M_p, M_q) = \left[ \sum_{j=1}^m |x_{pj} - x_{qj}|^{\lambda} \right]^{\frac{1}{\lambda}} \quad (1)$$

For  $\lambda = 2$ ,  $d$  from relation (1) represents the Euclid distance.

### Binary distance measures

The binary distance measures are useful for developing reliable SM and DM descriptors when molecules are treated as chemical graphs, namely hydrogen depleted graphs. In this case, the universe  $\mathbf{U}$  is topological, and the relevant properties are described by the adjacency matrix  $\mathbf{A}(\mathbf{M})$ . Some transformations in  $\mathbf{U}$  should not change the similarity measure.

$M_i$  and  $M_q$  are binary lists (or vectors, if one prefers this term). Let  $a, b, c, e$ , be integers such as:

$a$  is equal to the number of occurrence

of  $x_{ij} = 1$  and  $x_{qj} = 1$

$b$  is equal to the number of occurrence

of  $x_{ij} = 0$  and  $x_{qj} = 1$

$c$  is equal to the number of occurrence

of  $x_{ij} = 1$  and  $x_{qj} = 0$

$e$  is equal to the number of occurrence

of  $x_{ij} = 0$  and  $x_{qj} = 0$

These definitions are symbolically described in Table 1.

**Table 1.** Definitions for computation of binary distance measures

$X_q \backslash X_i$	1	0
1	a	b
0	c	e

For quantitative evaluation of similarity (or dissimilarity) among bioactive molecules ( $\mathbf{L}$ ) we tested the following distance measures:<sup>9</sup>

$$\text{RUSSEL and RAO} \quad d_1(X_i, X_q) = \frac{a}{a+b+c+e} \quad (2)$$

$$\text{JACCARD and NEEDHAM} \quad d_2(X_i, X_q) = \frac{a}{a+b+c} \quad (3)$$

$$\text{DICE} \quad d_3(X_i, X_q) = \frac{a}{2a+b+c} \quad (4)$$

$$\text{SOKAL and SNEATH} \quad d_4(X_i, X_q) = \frac{a}{a+2(b+c)} \quad (5)$$

$$\text{SOKAL and MICHENER} \quad d_5(X_i, X_q) = \frac{a+e}{a+b+c+e} \quad (6)$$

$$\text{KULZINSKY} \quad d_6(X_i, X_q) = \frac{a}{b+c} \quad (7)$$

$$\text{ROGERS and TANIMOTO} \quad d_7(X_i, X_q) = \frac{a+e}{a+e+2(b+c)} \quad (8)$$

$$\text{YULE} \quad d_8(X_i, X_q) = \frac{ae-bc}{ae+bc} \quad (9)$$

$$\text{CORRELATION} \quad d_9(X_i, X_q) = \frac{ae+bc}{[(a+b)(c+e)(a+c)(b+e)]^{\frac{1}{2}}} \quad (10)$$

$$\text{HAMMING} \quad d_{10} = b+c \quad (11)$$

### Structural description

Graph theory is largely applied to characterization and to quantitative treatment of chemical structure by means of so-called topological indices. It is also used for structure – property (QSPR) and structure – activity (QSAR) correlations.<sup>10</sup> Among different approaches to QSPR/QSAR, the structure-explicit approach based on graph theoretical methods is especially interesting and attractive due to its effectiveness and simplicity.<sup>11, 12</sup>

A graph  $\mathbf{G}$  is a mathematical structure consisting of points (vertices) connected by lines (edges). The constitutional formula of a chemical compound may be viewed as a molecular graph, where the vertices represent atoms and the edges represent the covalent bonds. Commonly, for quantitative description one uses hydrogen-depleted graphs.

There are some topological invariants; the adjacency matrix is the most important for our purpose.

The adjacency matrix of the molecular graph  $\mathbf{G}$ ,  $\mathbf{A}(\mathbf{G})$ , is a basic mathematical structure which maps a certain molecule  $M$ . For a molecule having  $m$  atoms,  $\mathbf{A}(\mathbf{G})$  is a square  $m \times m$  matrix whose entries  $a_{ij}$  have only different values – 1 or 0 – due to the fact that two atoms in a molecule are in binary relation, being either connected or not connected.<sup>10</sup>

We used the adjacency matrix  $\mathbf{A}(\mathbf{G})$  to perform the superposition of bioactive molecules of the alcohol series under QSAR study.

### The Minimal Steric Difference method (MSD) and the superposition procedure

The *MSD* method was introduced as a quantitative measure for steric misfit, taking into account the fact that steric fit depends on the shape of both interacting molecules, the biological receptor *R* and the drug molecule *L*.<sup>13</sup> The fundamental of this approach is that the affinity of *L* molecules for a receptor is a closely fit linearly decreasing function of the steric misfit of the *L* with *R*.

The molecule of highest activity, denoted as standard *S*, models the receptor site. The shape of the standard is considered complementary to the receptor cavity. One seeks for the maximal superposition of each drug molecule *L* upon the standard *S*. Hydrogen atoms are neglected in order to simplify the problem, since their van der Waals volumes and covalent radii are rather small. Thus, in the *MSD* procedure one uses the hydrogen-depleted graphs. The small differences in bond lengths and bond angles are neglected, and various conformations of flexible molecules are considered to be degenerated. One supposes that the *L* molecule  $M_i$  interacts with the receptor by its conformation which better fits the cavity of *R*; that is, one must select the *minimal steric difference (MSD)*.

The *MSD* parameter was defined as the number of non-superposed non-hydrogen atoms of the two compared molecules,  $M_i$  and *S*; obviously, the value of *MSD* for *S* is 0.

### Topological Similarity/Dissimilarity Measure (TSDM)

Ciubotariu et al. have introduced a similarity/dissimilarity measure (*SDM*) on the basis of *MSD* procedure, by normalizing the Hamming distance (equation (10)).<sup>14</sup> The values of *SDM* can be calculated with the following formula:

$$SDM = \frac{b+c}{a+b+c} \quad (12)$$

The *SDM* parameter models these two concepts – similarity and dissimilarity – on the same scale: [0,1]. *SDM*=0 implies a high similarity, and *SDM*=1 – a high dissimilarity.

Similarity, dissimilarity and complementarity are concepts frequently used in drug design. To avoid ambiguity, they have to be defined and expressed mathematically, so that we can distinguish precisely between them. For example, these three concepts can be expressed analogously to the extrema of a correlation, where perfect similarity between two objects is given a value of +1, dissimilarity has a value of 0 and complementarity has a value of -1.<sup>15</sup>

Therefore, the *SDM* parameter can be transformed into a new one (denoted by *TSDM* – topological similarity/dissimilarity measure), so that the maximum degree of similarity corresponds to a value of 1, and the highest dissimilarity has a value of 0. Consequently,

$$TSDM = 1 - SDM = \frac{a}{a+b+c} \quad (13)$$

We can see that the *TSDM* parameter is right  $d_2$  binary distance measure (see the formulae (3) of Jaccard and Needham). The superposition procedure is that of the *MSD* and *MTD* techniques.<sup>16</sup>

The molecules from the series under study are ranking in descending order of biological activities. We describe the topography of the receptor by indicating the presence or absence of atoms from the considered  $M_i$  molecule ( $i=1,n$ ) in the vertices of the hypermolecule **H**. The hypermolecule is obtained by an approximately atom-by-atom superposition of all  $n$   $M_i$  molecules, neglecting the hydrogen atoms. The hypermolecule **H** can be considered as a topological network in which the vertices correspond to atoms and edges may be viewed as chemical bonds. Let  $m$  be the number of vertices of **H**. The vector  $\mathbf{X}=\{x_{ij}\}$  ( $i=1,n; j=1,m$ ), obtained from the adjacency matrix, describes the relation between the molecule  $M_i$  and the hypermolecule **H**. The entries  $x_{ij}$  is taken to be 1 if the vertex  $j$  of **H** is occupied by a non-hydrogen atom of the molecule  $M_i$  and  $x_{ij}=0$  if it is not occupied.

This superposition procedure was also applied in the QSAR analysis of the  $d_1 - d_{10}$  binary distance measures (relations 2 – 11).

## RESULTS AND DISCUSSION

The topological similarity/dissimilarity indicators (relations 2 – 10) and *TSDM* (relation 13, which is identical with relation 3) have been applied in the study of a series of saturated alcohols, which competitively inhibit microsomal *p*-hydroxylation of aniline. This biological process takes place within the cytochrome P-450, a part of the cell's detoxification system, in which many foreign substances are oxidized prior to their elimination.

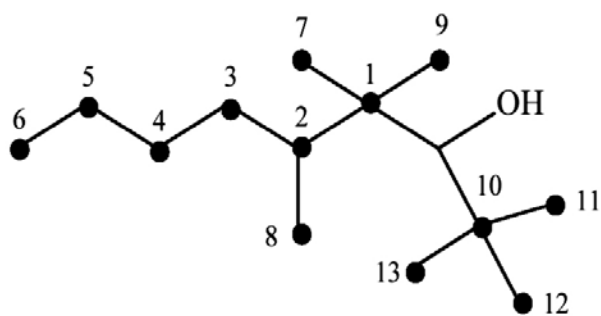
The experimental data<sup>17</sup> are listed in Table II. As bioactivity,  $A_p$ , we used the logarithm of reciprocal concentrations (millimolar) that causes a 50% inhibition:  $A = pIC_{50} = \log(1/C_{50})$ .

The most active ligand of this series has been chosen as standard, *S*; this is the first compound from Table II, 1-heptanol. The geometries of all compounds have been optimized by means of molecular mechanics force fields. These optimized geometries have been used in the superposition procedure. The

hypermolecule (**H**) obtained by approximate atom per atom superposition of all alcohols is depicted in Figure 1. It has  $m=13$  vertices.

**Table 2.** Biological activities ( $A$ ) of saturated alcohols inhibiting the microsomal  $p$ -hydroxylation of aniline.  $A=\log(1/C_{50})$ , where  $C_{50}$  is the concentration that produces 50% inhibition

No.	Alcohol	$A=-\log C_{50}$
1.	1-heptanol (S)	0.68
2.	1-hexanol	0.54
3.	1-pentanol	0.27
4.	2-heptanol	0.25
5.	2-hexanol	0.15
6.	1-butanol	-0.05
7.	2-pentanol	-0.07
8.	2-Me-1-butanol	-0.15
9.	3-Me-1-butanol	-0.19
10.	2-butanol	-0.35
11.	3-pentanol	-0.37
12.	2-Me-1-propanol	-0.39
13.	3-hexanol	-0.47
14.	2-propanol	-0.47
15.	1-propanol	-0.48
16.	2,2-diMe-1-propanol	-0.67
17.	2-Me-3-pentanol	-0.89



**Figure 1.** Hypermolecule **H** obtained by superposition of the saturated alcohols (see Table II).

**H** was used as a topological network to describe each of the molecules  $i=1, n$  ( $n=17$ ) involved in its building with the aid of the vector  $\mathbf{X}=\{x_{ij}\}$ ;  $x_{ij}$  takes the value 1 if the vertex  $j$  of **H** is occupied by a non-

hydrogen atom of the molecule  $i$ , and  $x_{ij}=0$  otherwise. The binary descriptor  $\mathbf{X}$  is presented in Table 3.

On the basis of  $\mathbf{X}$  we calculated with the relations (2) – (11) the values of similarity/dissimilarity indicators  $d_1, d_2, \dots, d_{10}$  (see Table 4).

From Table IV we can see that the alcohols 2-pentanol, 2-metil-1-butanol and 3-metil-1-butanol may be considered similar. This fact is due to the calculated optimized geometries of these alcohols, which have been used for the construction of the hypermolecule of Figure 1. Though **H** is a 2D topological network, it contains some 3D information. The superposition procedure and the all similarity indicators prove the equivalence of the vertices 7, 8 and 10. The nodes 7 – 13 produce a decrease of the inhibition process, if the atoms of alcohols occupy them.

The inter-correlation matrix of the similarity indicators is presented in Table 5. In general, these parameters are inter-related, with some exceptions:  $d_6, d_8$  and  $d_9$ . They do not measure the topological volume, as expressed by the number of carbon atoms,  $N$ .

**Table 3.** The structural descriptor  $X = \{x_{ij}\}$ ,  $i=1,17$ ,  $j=1,13$ , for the alcohols of Table 2 (see also Figure 1)

No.	$x_{ij}$	1	2	3	4	5	6	7	8	9	10	11	12	13
1.	$X_{11}$	1	1	1	1	1	1	0	0	0	0	0	0	0
2.	$X_{21}$	1	1	1	1	1	0	0	0	0	0	0	0	0
3.	$X_{31}$	1	1	1	1	0	0	0	0	0	0	0	0	0
4.	$X_{41}$	1	1	1	1	1	0	0	0	0	1	0	0	0
5.	$X_{51}$	1	1	1	1	0	0	0	0	0	1	0	0	0
6.	$X_{61}$	1	1	0	0	0	0	0	1	0	0	0	0	0
7.	$X_{71}$	1	1	1	0	0	0	0	0	0	1	0	0	0
8.	$X_{81}$	1	1	1	0	0	0	1	0	0	0	0	0	0
9.	$X_{91}$	1	1	1	0	0	0	0	1	0	0	0	0	0
10.	$X_{101}$	1	1	0	0	0	0	0	0	0	1	0	0	0
11.	$X_{111}$	1	1	0	0	0	0	0	0	0	1	0	0	1
12.	$X_{121}$	1	1	0	0	0	0	1	0	0	0	0	0	0
13.	$X_{131}$	1	1	1	0	0	0	0	0	0	1	1	0	0
14.	$X_{141}$	1	0	0	0	0	0	0	0	0	1	0	0	0
15.	$X_{151}$	1	1	0	0	0	0	0	0	0	0	0	0	0
16.	$X_{161}$	1	1	0	0	0	0	1	0	1	0	0	0	0
17.	$X_{171}$	1	1	0	0	0	0	0	0	1	1	0	1	0

**Table 4.** The values of similarity/dissimilarity measures  $d_1, d_2, \dots, d_{10}$  (and of TSDM= $d_2$ ) for the alcohols from Table 2.

No.	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$	$d_8$	$d_9$	$d_{10}$
1.	0.462	1.000	0.500	1.000	*	1.000	1.000	42.00	-	0
2.	0.385	0.833	0.455	0.714	5.000	0.923	0.857	35.00	35.000	1
3.	0.308	0.667	0.400	0.500	2.000	0.846	0.733	28.00	28.000	2
4.	0.385	0.714	0.417	0.556	2.500	0.846	0.733	29.80	30.024	2
5.	0.308	0.571	0.364	0.400	1.333	0.769	0.625	23.00	24.049	3
6.	0.231	0.500	0.333	0.333	1.000	0.769	0.625	21.00	21.000	3
7.	0.231	0.429	0.300	0.273	0.750	0.692	0.529	15.00	18.077	4
8.	0.231	0.429	0.300	0.273	0.750	0.692	0.529	15.00	18.077	4
9.	0.231	0.429	0.300	0.273	0.750	0.692	0.529	15.00	18.077	4
10.	0.154	0.286	0.222	0.167	0.400	0.615	0.444	4.00	12.113	5
11.	0.154	0.250	0.200	0.143	0.333	0.538	0.368	-2.00	10.206	6
12.	0.154	0.286	0.222	0.167	0.400	0.615	0.444	4.00	12.113	5
13.	0.231	0.375	0.273	0.231	0.600	0.615	0.444	11.00	15.146	5
14.	0.077	0.143	0.125	0.077	0.167	0.538	0.368	-19.00	6.164	6
15.	0.154	0.333	0.250	0.200	0.500	0.692	0.529	14.00	14.000	4
16.	0.154	0.250	0.200	0.143	0.333	0.538	0.368	-2.00	10.206	6
17.	0.154	0.222	0.182	0.125	0.286	0.462	0.300	-4.00	8.293	7

**Table 5.** The inter-correlation matrix of similarity/dissimilarity indicators  $d_1 - d_{10}$  and  $N$  – the number of carbon atoms from alcohol molecules. The figures are the values of correlation coefficients

	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$	$d_8$	$d_9$	$d_{10}$	$N$
$d_1$	1.000	0.982	0.977	0.958	0.919	0.848	0.933	0.940	0.814	-0.917	0.774
$d_2$		1.000	0.987	0.984	0.967	0.894	0.982	0.953	0.847	-0.967	0.656
$d_3$			1.000	0.945	0.968	0.835	0.964	0.983	0.779	-0.968	0.644
$d_4$				1.000	0.936	0.942	0.971	0.898	0.915	-0.936	0.651
$d_5$					1.000	0.825	0.992	0.951	0.787	-1.000	0.461
$d_6$						1.000	0.875	0.763	0.891	-0.825	0.490
$d_7$							1.000	0.937	0.846	-0.992	0.505
$d_8$								1.000	0.739	-0.951	0.586
$d_9$									1.000	-0.789	0.565
$d_{10}$										1.000	-0.461
$N$											1.000

**Table 6.** Correlation results: the correlation coefficients ( $r$ ) corresponding to the least-squares equation of the form  $A = \alpha + \beta \cdot d_j, j = 1, 11, d_{11} = N$  where  $A$  represents the calculated biological activities and  $d_j$  are the indicators analyzed here

$d_j$	$r$	$a$	$b$	$t^*$
$d_1$	0.907	-1.032	3.722	8.3
$d_2$	0.944	-0.931	1.706	11.1
$d_3$	0.936	-1.292	3.830	10.3
$d_4$	0.919	-0.682	1.604	9.0
$d_5$	0.961	-2.083	2.765	13.5
$d_6$	0.829	-0.481	0.255	5.5
$d_7$	0.958	-1.350	2.152	12.9
$d_8$	0.886	-0.481	0.024	7.4
$d_9$	0.765	-0.486	0.015	4.6
$d_{10}$	0.961	+0.682	-0.212	13.5
$d_{11}=N$	0.496	-	-	-

\* Student t-test

The results of correlations of the biological activities of the alcohols,  $A_i, i=1, 17$  (see Table 2), with the values of these indicators (from Table 6).

The correlation with  $d_6, d_8$  and  $d_9$  indices, and also with the number of carbon atoms  $N$ , fails (see the values of  $b$  for  $d_j, j=6, 8, 9$ ). The best results have been obtained for  $d_4$  and  $d_{10}$ . As a matter of fact, the values of MSD parameter are exactly the Hamming distances,  $d_{10}$ .

We have adopted as topological similarity/dissimilarity measure (TSDM) the relation (13), together with the superposition procedure described above. Hence, if TSDM=1, the similarity degree is maximum:  $b=c=0$  and  $a=0$ , that is, from the topological point of view, the two compared molecules are identical. If  $a=0$ , i.e., there are no common vertices of two compared molecules, the degree of topological dissimilarity is maximum and, consequently, TSDM=0 (see relation (13) or (3)).

We have applied the TSDM method to the study of the same series of alcohols from Table 2. By using the least-squares method, we have obtained the following correlation equation:

$$A = -0.931(\pm 0.193) + 1.707(\pm 0.327)SDM; (EV = 0.892; s = 0.144) \quad (14)$$

$$(t = 8.5; F = 73.0) \quad (t = 11.1; F = 123.7); r = 0.944; r_{CV}^2 = 0.736;$$

where  $r$  is the correlation coefficient, and the other statistical parameters have the following meanings:  $EV$  – explained variance,  $s$  – standard error,  $t$  – Student test,  $F$  – Fisher test,  $r_{CV}^2$  – cross-validation coefficient, and ( $t=, F=$ ) refers to the statistical corresponding indicators of the two coefficients ( $a=-0.931$  and  $b=1.706$  - see Table 6) from equation (14), respectively.

The 95% confidence limits for the parameters of the linear model (14) (0.776 and 1.707) have also been calculated. They offer important information about the reliability of the model.

Equation (14) accounts for about 90% of the variation in the  $pIC_{50}$  data, and the standard error of the estimate is only 4% of the range of the  $pIC_{50}$  data. The statistical significance of equation (14) is very good, as observed from the Fisher and Student's statistics.

The present QSAR analysis shows that the degree of inhibition of aniline p-hydroxylation by alcohols is greatly influenced by the degree of similarity between them. The positive slope of the line leads to the conclusion that inhibition increases with the growth of similarity degree or, equivalently, decreases with the growth of dissimilarity degree between n-heptanol and the other alcohols.

Our QSAR study is in accordance with the results reported for the interaction between cytochrome P-450 and alcohols, which competitively inhibit the microsomal p-hydroxylation of aniline.<sup>10</sup> Thus, the mechanism of interaction between alcohols and the enzymatic receptor can be viewed as a "zipper" mechanism, and it has two stages. The first stage consists in a (possibly hydrogen or dipole-dipole) bond formation between the hydroxyl group of the alcohols and some site of the enzyme. This interaction is responsible for the inhibitory activity of the alcohols. The second stage implies a hydrophobic interaction between the alkyl chain and a hydrophobic zone that is close to the enzyme site involved in the first stage. This stage proceeds step by step for different long chains with conformational flexibility, and it may account for the quantitative differences in activity for

the studied alcohols. Thus, the initial interaction from first stage is followed by a series of conformational rearrangements of both alcohol and enzyme, leading to the binding of the remaining segments of the alkyl chain to their appropriate position. This mechanism is strongly supported by the good correlation between experimental values,  $pIC_{50}$ , and TSDM model (see Equation (14)). The decrease of similarity among n-heptanol and the other alcohols from Table 2, accompanied by the increase of dissimilarity, also decrease the possibility of interaction with the hydrophobic walls of the (possibly) cleft enzyme site. Smaller chains realize this situation, the threshold between similarity and dissimilarity being for 2-hexanol ( $SDM=0.5$ ) when biological activity has still positive value.

## CONCLUSIONS

TSDM method, reported in this paper, allows quantitative treatment of the topological similarity and dissimilarity between two bioactive molecules on the same scale. The range of variation is from  $TSDM=1$ , which means the highest similarity, meaning that the two molecular graphs are identical for any type of atoms, to  $SDM=0$ , for which one obtains the lowest similarity, i.e. the highest dissimilarity. In this way, it is not necessary to define an arbitrary threshold between similarity and dissimilarity; the threshold has the value of 0.5.

As a quantitative measure of both topological similarity and dissimilarity, the TSDM parameter can be used for QSAR studies. If the correlation equations are statistically validated, it can be used to explain biological interactions, in terms of similarity or dissimilarity, between bioactive molecules (called also effectors or ligands) and biological receptor or to predict the biological activity for unknown compounds.

The potential of TSDM method for QSAR studies is revealed by the results obtained in the analysis of competitive enzyme inhibition of microsomal p-hydroxylation of aniline by a series of alcohols. Thus, it is evident that the binding of alcohols to cytochrome P-450, which causes the competitive inhibition mentioned above, must take place at the active site of the free enzyme. The size and the degree of the alkyl chains branching of the alcohols, which leads to a higher degree of dissimilarity with respect to the most active compound, n-heptanol, which complementary mod-

els the site of the biological receptor, account for the quantitative differences in the inhibition. The TSDM structural parameter used in our QSAR study adequately describes these quantitative differences. The predictive power of the linear model proposed above (Equation (14)) is also very good, if one takes into account the difficulty of obtaining the accurate experimental data. The introduction of other structural parameters may improve these results.

The TSDM index is also useful for designing test series of bioactive compounds from the general screening procedures, which lead to too many compounds (commonly, more than 10,000 ligand molecules). In this case the molecules, two by two, must be reciprocal dissimilar, for optimization of the information/expenses ratio.

## REFERENCES

1. Dean PM. *Molecular Similarity in Drug Design*, Blackie Academic and Professional, 1995, London.
2. Gonzalez FJ, Gelboin HV. Role of human cytochromes P-450 in the metabolic activation of chemical carcinogens and toxins. *Drug Metab Rev* 1994;26:165.
3. Waterman M, Johnson E. *Methods in enzymology*. Academic Press, 2001, New York.
4. Poulos T. Cytochrome P450. *Curr Opin Struct Biol* 1995;5:767.
5. Woolf TF. *Handbook of Drug Metabolism*. Marcel Dekker Inc 1999, New York, Basel.
6. Hall P. Cytochrome P450 and the regulation of steroid synthesis. *Steroids* 1986;48:131.
7. Nelson D, Koymans L, Kamataki K, et al. P 450 superfamily: Update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics* 1996;6:1.
8. Nebert DW, Nelson DR. P450 gene nomenclature based on evolution, in: Waterman R, Johnson EF. *Cytochrome P450. Methods in Enzymology*, 1991 Academic Press, New York.
9. Diday E, Simon JC. Clustering Analysis, in: Fu KS. *Digital Pattern Recognition*, Springer Berlin, 1980,47–57.
10. Trinajstić N, Randić M, Klein DJ, *Acta Pharm Jugosl* 1986;36:267–279.
11. Plavšić D, Graovac A. A calculation of molecular descriptors Based on various graphical bond orders, in: Diudea MV. *QSPR/QSAR Studies by Molecular Descriptors*. Nova Science Inc, Huntington, 2000, New York, p.39–61.
12. Balaban AT, Motoc I, Bonchev D, et al. Topological indices for structure – activity correlations, in: Charton M, Motoc I. *Steric effects in drug design*, Springer, 1983, Berlin, p.21–55.
13. Simon S, Szabadai Z. *Studia Biophys* 1973;39:123–7.
14. Ciubotariu D, Muresan S, Gogonea V, et al. *Roum Biotechnol Lett* 1997;2:114–130.
15. Dean PM. *Molecular Similarity*, in: Kubinyi H. *3D QSAR in drug design*, ESCOM, 1993, Leiden, p.151.
16. Ciubotariu D, Gogonea V, Medeleanu M. Van der Waals molecular descriptors. Minimal steric difference, in: Diudea MV. *QSPR/QSAR studies by molecular descriptors*, Nova Science Inc, Huntington, 2000, New York, p.281–361.
17. Sabljic A, Protic-Sabljić M. *Mol Pharmacol* 1983;23:213–8.