J Math Chem (2011) 49:2330–2344
DOI 10.1007/s10910-011-9889-1

ORIGINAL PAPER

A new manner to use application of Shannon Entropy in similarity computation

Laszlo Tarko

Received: 20 June 2011 / Accepted: 22 July 2011 / Published online: 6 August 2011

Abstract The paper proposes a new manner to use application of Shannon Entropy in similarity computation for any objects and any objects groups. In computation of the chemical similarity cases the proposed formulas use the values of one or more molecular descriptors, divided into classes (categories) by using a suitable criterion. The paper proposes original criteria to made difference between 'saturated', 'non-saturated non-aromatic' and 'aromatic' molecular fragments and between 'hydrogen-acceptor' and 'hydrogen-donor-acceptor' fragments for the purpose of classifying fragments into classes. According to the proposed formula two molecules A and B are similar enough if the value of Shannon Entropy of A + B aggregate is close to the value of Shannon Entropy of A molecule *and* close to the value of Shannon Entropy of B molecule. The proposed similarity formula can be used as statistical correlation index, useful if the number of values of two analyzed variables is unequal. The proposed formula is useful in the quantitative evaluation of the 'representative sample' character of any 'sample'. The paper presents the chemical similarity computation in Zimelidine/Fluoxtine/Chloramphenicol/Crufomate/Phoxim group. For comparison purposes, the paper also presents a Tanimoto coefficient calculation for the same molecules group. In addition, the paper presents two non-chemical examples regarding 'Representative Sample Index' calculation.

Keywords Shannon Entropy · Tanimoto · Chemical similarity · Representative sample

L. Tarko (✉)
Organic Chemistry Center 'C.D. Nenitzescu', Romanian Academy, 202B Spl. Independentei, 6th Sector, PO Box 35-108, Bucharest, MC 060023, Romania
e-mail: ltarko@cco.ro

1 Introduction

Similarity searching is one of the most widely applied approaches in chemical and pharmaceutical research to select compounds with desired properties from large databases. This approach is based on the hypothesis 'similar properties are the effect of similar chemical structures'. If the 'properties' are 'biochemical activities' this assertion is named 'the QSAR axiom' (QSAR means Quantitative Structure-Activity Relationship). Frequently, this 'axiom' is challenged because some similar structures present non-similar properties and some non-similar structures present similar properties.

Different methods have been developed for similarity searching when the molecules are small enough. All similarity search approaches depend on the representation of molecular structures and on the quantification of molecular similarity.

Molecular structure can be represented, for instance, by 'fingerprint' which is, in fact, a matrix that records the number of occurrences of specific features in molecular graph [1,2]. Fingerprints are usually compared by formula (1) of Tanimoto coefficient [3].

$$T = n_{AB} / (n_A + n_B - n_{AB}) \tag{1}$$

In formula (1) n_A is number of occurrences in the A molecule, n_B is the number of occurrences in the B molecule and n_{AB} is the number of occurrences in the A molecule *and* the B molecule. There are many other similarity metrics based on fingerprints [4–6].

Despite of a large number of methods for the evaluation of molecular shape there are only few methods which compute chemical similarity as 'similarity of shape'. These methods evaluate structural properties at each grid point of an orthogonal grid placed around the molecules [7], evaluate common-overlap steric volume between pairs of molecules [8], measure match between the electron densities of two analyzed molecules [9,10] etc.

After computation of certain number of descriptors D_i values, for two molecules A and B, one can calculate the similarity of the analyzed molecules using Euclidian distance ED.

$$ED = \left[\sum \left(D_i^A - D_i^B \right)^2 \right]^{1/2} \tag{2}$$

The statistical correlation of descriptors used in formula (2) should be low and the statistical distribution of descriptors values should be Gaussian. Mahalanobis distance [11] MD is less sensitive to these conditions because it includes the correlation between descriptors. The term S in formula (3) is the covariance matrix.

$$MD = \left[(D_A - D_B)^T S^{-1} (D_A - D_B) \right]^{1/2} \tag{3}$$

There are many papers in literature that describes the use of Shannon Entropy SE [12] (usually named 'Information Content') as molecular descriptor, computed using

the values of certain other descriptors and formula (4).

$$\mathrm{SE} = -\sum_{i=1}^{k} n_i/N \cdot \mathrm{Log}\,(n_i/N) \tag{4}$$

To apply formula (4) one must use an adequate criterion for putting a certain value into certain class (category). In formula (4) N is the total number of the descriptor values and n_i is the number of the values included in class i. The number k_{def} of defined classes can be large, but formula (4) uses only non-empty classes ($n_i > 0, k \leq k_{def}$). The base of the logarithm is 2, Euler's number e or 10.

If $k_{def} < N$ the value of SE is within [0, Log k_{def}] range, otherwise the value of SE is within [0, Log N] range. Consequently, the inequalities (5) are true.

$$0 \leq \mathrm{SE} \leq \min\left(\mathrm{Log}\ k_{def}, \mathrm{Log}\ N\right) \tag{5}$$

If $k = k_{def}$ and each class includes the same number of values then SE in formula (4) has the maximum value. If $k = 1$ (all values are included within the same class) then SE = 0, because $n_k = N$.

If $m = \min(\mathrm{Log}\ k_{def}, \mathrm{Log}\ N)$ the value of weighted Shannon Entropy WSE is within [0, 1] range.

$$\mathrm{WSE} = \mathrm{SE}/m \quad (m > 0) \tag{6}$$

We can compute SE for any group of objects according to the values of certain measured common feature of the analyzed objects. For instance, we can compute SE for a group of people using the values of height, weight, age, blood pressure etc. In addition, we observe that any object is, in fact, a collection of objects. A human body is a collection of cells, a state is a collection of districts, an electronic device is a collection of electronic pieces, a molecule is a collection of atoms/chemical bonds/topological paths/electric charges etc. Consequently, we can compute SE for a human body, state, electronic device, molecule etc.

According to this approach, Shannon Entropy measures, in chemistry field, the 'diversity' of atomic numbers, net charges, bond orders, length of topological paths, atomic distances, atomic volumes etc. in the analyzed molecule. Recent papers propose using of Shannon Entropy for computing a certain 'aromaticity descriptor' [13] and distributions of atom-centered feature pairs [14].

To compute 'chemical similarity' one can use the value of Shannon Entropy descriptors in above Euclidian / Mahalanobis distance formulas (2) and (3).

Another manner [15] to use Shannon Entropy in evaluation of chemical similarity is computation of 'Differential Shannon Entropy' DSE.

$$\mathrm{DSE} = \mathrm{SE_{AB}} - (\mathrm{SE_A} + \mathrm{SE_B})/2 \tag{7}$$

In formula (7) $\mathrm{SE_A}$ and $\mathrm{SE_B}$ are computed for molecules A and B and $\mathrm{SE_{AB}}$ is computed for virtual A + B aggregate. If $\mathrm{SE_{AB}}$ is within [$\mathrm{SE_A}$, $\mathrm{SE_B}$] range the value of DSE can be small, even null, despite of large difference $|\mathrm{SE_A} - \mathrm{SE_B}|$.

Using SE/Log N ratio (not SE/m ratio) instead SE one compute [15] the 'weighted' WDSE and the 'Reciprocal Differential Shannon Entropy' RDSE.

$$\mathrm{RDSE} = 1/\mathrm{WDSE} \tag{8}$$

Similarity searching usually provides a ranking of compounds relative to chosen reference molecule(s) [16–18].

This paper proposes a new manner to use Shannon Entropy in computation of similarity, applicable to chemical similarity.

2 Methods and formulas

The number k_{def} of ascertained classes can be large. Using the same descriptor and the same criterion, the descriptor values of the molecule A will be placed into k_A non-empty classes, the descriptor values of the molecule B will be placed into k_B non-empty classes and the descriptor values of the A + B aggregate will be placed into k_{AB} non-empty classes.

The inequalities in formula (9) are satisfied.

$$k_A + k_B \geq k_{AB} \geq \max\ (k_A, k_B) \tag{9}$$

Using the values (divided in classes) of certain molecular descriptor the chemical similarity SIM of two molecules A and B is computed by proposed formula (10), which is a product of two ratios.

$$\mathrm{SIM} = \mathrm{R_A}\cdot\mathrm{R_B} \tag{10}$$

where

if $\mathrm{SE_A} \leq \mathrm{SE_{AB}}$ then $\mathrm{R_A} = \mathrm{SE_A}/\mathrm{SE_{AB}}$ else $\mathrm{R_A} = \mathrm{SE_{AB}}/\mathrm{SE_A}$

if $\mathrm{SE_B} \leq \mathrm{SE_{AB}}$ then $\mathrm{R_B} = \mathrm{SE_B}/\mathrm{SE_{AB}}$ else $\mathrm{R_B} = \mathrm{SE_{AB}}/\mathrm{SE_B}$

$\mathrm{SE_A}$ is computed by formula (4) using $\mathrm{N_A}$ descriptor values of the molecule A

$\mathrm{SE_B}$ is computed by formula (4) using $\mathrm{N_B}$ descriptor values of the molecule B

$\mathrm{SE_{AB}}$ is computed by formula (4) using $\mathrm{N_A} + \mathrm{N_B}$ descriptor values of the A + B aggregate.

The value of SIM is within [0, 1] range.

Two molecules are similar enough (high value of SIM) if the value of Shannon Entropy $\mathrm{SE_{AB}}$ of the A + B aggregate is close to the value of Shannon Entropy $\mathrm{SE_A}$ of the A molecule *and* Shannon Entropy $\mathrm{SE_B}$ of the B molecule.

If difference $|\mathrm{SE_A} - \mathrm{SE_B}|$ is large, the value of SIM is always low enough. If difference $|\mathrm{SE_A} - \mathrm{SE_B}|$ is small, the value of SIM can be high or low.

The SIM value for two isomers is, as a rule, high.

One can use formula (10) as statistical correlation index. Unlike Pearson [21], Spearman [22] and Kendall [23] correlation indices, SIM can be computed even if two analyzed variables don't have the same number of values.

If two or more descriptors are used the 'total' similarity SIM_{total} should be evaluated, in our opinion, by criterion (11).

$$SIM_{total} = \min\ (SIM_1,\ SIM_2,\ SIM_3, \ldots, SIM_k) \tag{11}$$

If the classes are defined beforehand, one can divide the molecules (objects) in classes (categories) according to value of a certain descriptor, computed for each molecule (object). In this case the similarity **of two objects groups**, G_A and G_B, should be computed using the same formula (10).

If G_A is an extracted (selected) 'sample' from the G_B population, it is interesting to compare G_A and G_B. In this case a version of the formula (10) can be used as 'Representative Sample Index' RSI.

$$RSI = R_A \cdot R_B \tag{12}$$

where

if $SE_{sample} \leq SE_{sample+population}$ then $R_A = SE_{sample}/SE_{sample+population}$ else $R_A = SE_{sample+population}/SE_{sample}$

if $SE_{population} \leq SE_{sample+population}$ then $R_B = SE_{population}/SE_{sample+population}$ else $R_B = SE_{sample+population}/SE_{population}$

SE_{sample} is computed by formula (4) using N_A objects in sample

$SE_{population}$ is computed by formula (4) using N_B objects in population

$SE_{sample+population}$ is computed by formula (4) using $N_A + N_B$ objects in sample + population aggregate.

The value of RSI is within [0, 1] range.

Another manner of analysis involves computing of SIM for each **pair** of molecules (objects). One can find empirically a limit value for similarity, which can be used to decide whether two molecules (objects) should be included or not into the same class. Consequently, one can divide the analyzed group into '(chemical) clusters' which are not defined beforehand. In this case, we propose to compute the similarity **of two clusterized groups** using the same formula (10).

Clusterization of any group of objects is a difficult task [24] and it is not the subject of this paper. After (chemical) clusterization there are clusters which include many molecules (objects) and other clusters which include only few molecules (objects). The molecules (objects) included in very small '(chemical) clusters' are similar with only few other molecules (objects) and can be considered '(chemical) outliers'.

The 'number of molecular fragments' is a specific descriptor. The fingerprints of this descriptor are usually compared by formula (1). Here we used formula (10).

We used as virtual fragmentation procedure a previously presented method [19]. According to quoted method, two bonded (by a chemical bond with computed *B* bond order) heavy atoms (and bonded hydrogen atoms) are included within the same fragment if *B* exceeds a limit value, empirically established. In order to compute *B* we have used the PM6 method [20], after geometry optimization.

The identified molecular fragments were placed into classes (categories) according to the proposed criteria in Table 1. As a rule, these criteria differentiate

Table 1 Criteria to place molecular fragments into classes

A[a]	B[b]	C[c]	Class (category)	Examples
0	0	0	I	F, Br
1	0	0	II	C, CH, CH_3
2–4	0	0	III	C=C, C≡CH
>4	0	0	IV	C_6H_5, C_6H_4
1	>0	0	V	O, N, S
2–4	>0	0	VI	C=O, C≡N, NCS, NO_2, SO_2, PO, N=N, N_3
>4	>0	0	VII	C_4H_3O (furyl), C_5H_4N (pyridinyl)
1	>0	>0	VIII	OH, NH_2, SH
2–4	>0	>0	IX	NHCO, CH=NH
>4	>0	>0	X	C_4H_3NH (2-pyrrolyl)

[a] A is number of fragment atoms different from hydrogen
[b] B is number of fragment atoms different from hydrogen, carbon and halogen
[c] C is number of Z–H bonds in fragment, where Z is a non-carbon atom

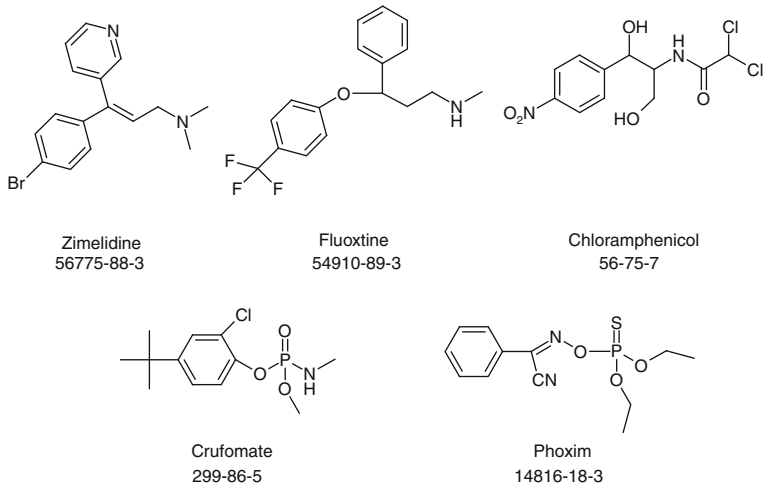

Fig. 1 The chemical structure of the analyzed molecules

'saturated'/'non-saturated non-aromatic'/'aromatic' fragments and 'hydrogen-acceptor'/'hydrogen-donor-acceptor' fragments.

3 Commented results

Present section includes an example of chemical similarity computation.

In formula (4) we used the natural logarithm, because it is default option for lot most software. However, this aspect is unimportant because the formula (10) uses ratios. Figure 1 presents the analyzed molecules and it's their Register Numbers.

Table 2 Some basic features of the analyzed molecules

Mol ID	Name	Chemical formula	Number of atoms	Number of edges in kenograph	Number vertices in kenograph
#1	Zimelidine	$C_{16}H_{17}BrN_2$	36	20	19
#2	Fluoxtine	$C_{17}H_{18}F_3NO$	40	24	22
#3	Chloramphenicol	$C_{11}H_{12}Cl_2N_2O_5$	32	20	20
#4	Crufomate	$C_{12}H_{19}ClNO_3P$	37	18	18
#5	Phoxim	$C_{12}H_{15}N_2O_3PS$	34	19	19

Table 3 Shannon Entropy of atomic numbers

Class	Mol ID				
	#1	#2	#3	#4	#5
Carbon	16	17	11	12	12
Hydrogen	17	18	12	19	15
Fluorine		3			
Chlorine			2	1	
Bromine	1				
Nitrogen	2	1	2	1	2
Oxygen		1	5	3	3
Phosphorus				1	1
Sulphur					1
SE[a]	0.9748	1.1017	1.3715	1.2039	1.3169

[a] Computed by formula (4)

Identifying of the most suitable molecular descriptors for the evaluation of chemical similarity is a difficult task and it is not subject of this paper. Here we have used the 'atomic number', 'chemical bond type in kenograph' and 'vertex degree in kenograph' because the values of these descriptors can be easily placed into classes. In addition, we have used the values of the 'number of molecular fragments' descriptor, according to the classes in Table 1.

According to the 'atomic number' value, the atoms in the analyzed molecule were placed into several classes (carbon, hydrogen, nitrogen, fluorine etc.).

Kenograph is the molecular graph including heavy atoms (different from hydrogen) only. On this graph the vertices are atoms (irrespective of type) and the edges are chemical bonds (irrespective of type).

According to the computed bond order value, the chemical bonds in the analyzed kenographs were placed into four classes (single, aromatic, double and triple).

The degree of certain vertex is the number of vertices bonded to the vertex. According to the 'vertex degree in kenograph' value (1, 2, 3 or 4), the atoms were placed into four classes.

Some basic features of molecules in Fig. 1 are presented in Table 2.

Table 4 Similarity of molecules aggregates

Class	Pair				
	(#1,#2)	(#1,#3)	(#1,#4)	(#1,#5)	(#2,#3)
(a)					
Carbon	33	27	28	28	28
Hydrogen	35	29	36	32	30
Fluorine	3				3
Chlorine		2	1		2
Bromine	1	1	1	1	
Nitrogen	3	4	3	4	3
Oxygen	1	5	3	3	6
Phosphorus			1	1	
Sulphur				1	
SE[a]	1.0884	1.2545	1.1548	1.2050	1.3035
SIM_1[b]	0.8847	0.7108	0.8096	0.7403	0.8034
	(#2,#4)	(#2,#5)	(#3,#4)	(#3,#5)	(#4,#5)
(b)					
Carbon	29	29	23	23	24
Hydrogen	37	33	31	27	34
Fluorine	3	3			
Chlorine	1		3	2	1
Bromine					
Nitrogen	2	3	3	4	3
Oxygen	4	4	8	8	6
Phosphorus	1	1	1	1	2
Sulphur		1		1	1
SE[a]	1.2077	1.2612	1.3095	1.3916	1.2824
SIM_1[b]	0.9094	0.8367	0.8778	0.9326	0.9141

[a] Computed by formula (4) for each aggregate
[b] Computed by formula (10)

Table 5 Shannon Entropy of bond orders

Mol ID	Single	Aromatic	Double	Triple	SE[a]
#1	7	12	1		0.8237
#2	10	13			0.6846
#3	10	10			0.6931
#4	11	7			0.6682
#5	10	7	1	1	1.0156

[a] Computed by formula (4)

Table 6 Similarity of molecules aggregates

Pair	Single	Aromatic	Double	Triple	SE[a]	SIM_2[b]
(#1, #2)	17	25	1		0.7697	0.8312
(#1, #3)	17	22	1		0.7847	0.8414
(#1, #4)	18	19	1		0.7962	0.8113
(#1, #5)	17	19	2	1	0.9586	0.8111
(#2 #3)	20	23			0.6907	0.9876
(#2, #4)	21	20			0.6928	0.9532
(#2, #5)	20	20	1	1	0.8846	0.6741
(#3, #4)	21	17			0.6876	0.9641
(#3, #5)	20	17	1	1	0.8923	0.6824
(#4, #5)	21	14	1	1	0.8844	0.6579

[a] Computed by formula (4) for each aggregate
[b] Computed by formula (10)

Table 7 Shannon Entropy of vertices degrees

Mol ID	Degree 1	Degree 2	Degree 3	Degree 4	SE[a]
#1	3	11	5		0.9592
#2	4	13	4	1	1.0713
#3	7	6	7		1.0961
#4	7	6	3	2	1.2763
#5	4	12	2	1	1.0102

[a] Computed by formula (4)

The values of the Shannon Entropy and the values of the similarity, computed using the values of 'atomic number' descriptor, are presented in Tables 3 and 4a, b. Table 3 includes the number of atoms having certain atomic number.

The values of the Shannon Entropy and the values of the similarity, computed using the values of 'chemical bond type in kenograph' descriptor, are presented in Table 5 and Table 6. Table 5 includes the number of bonds having a certain type, according to the computed bond order. The chemical bonds Ar–O in Fluoxtine, N–O in NO_2, N–C in amide group, C–O in amide group, P–O in Crufomate and P–S in Phoxim were computed as 'aromatic', according to the TOPAZ algorithm [25].

The values of the Shannon Entropy and the values of the similarity, computed using the values of 'vertex degree in kenograph' descriptor, are presented in Table 7 and Table 8. Table 7 includes the number of vertices having certain degree.

Figure 2 presents the identified molecular fragments in the analyzed molecules.

The values of the Shannon Entropy and the values of the similarity, computed using the values of 'number of molecular fragments' descriptor, are presented in Tables 9 and 10a, b. Table 9 includes the number of fragments in each molecule, in each class.

The values of the 'total' similarity are presented in Table 11.

Table 8 Similarity of molecules aggregates

Pair	Degree 1	Degree 2	Degree 3	Degree 4	SE[a]	SIM_3[b]
(#1, #2)	7	24	9	1	1.0387	0.8953
(#1, #3)	10	17	12		1.0736	0.8752
(#1, #4)	10	17	8	2	1.1998	0.7515
(#1, #5)	7	23	7	1	1.0229	0.9260
(#2 #3)	11	19	11	1	1.1496	0.8885
(#2, #4)	11	19	7	3	1.2079	0.8394
(#2, #5)	8	25	6	2	1.0991	0.8959
(#3, #4)	14	12	10	2	1.2382	0.8588
(#3, #5)	11	18	9	1	1.1462	0.8429
(#4, #5)	11	18	5	3	1.1853	0.7916

[a] Computed by formula (4) for each aggregate
[b] Computed by formula (10)

Table 9 Shannon Entropy of molecular fragments

Class	Mol ID				
	#1	#2	#3	#4	#5
I	1	3	2	1	
II	3	5	4	6	4
III	1				
IV	1	1	1	1	1
V	1			2	3
VI			1	1	3
VII	1	1			
VIII		1	2	1	
IX			1		
X					
SE[a]	1.6675	1.3667	1.6417	1.4735	1.2945

[a] Computed by formula (4)

According to the values of similarity in Table 11, the analyzed pairs of molecules can be ordered in several ways. We observe that the order is quit sensitive to the utilized set of descriptors.

Using SIM_4 similarity only:

$(\#3, \#4) > (\#4, \#5) > (\#1, \#4) > (\#2, \#3) > (\#2, \#4) > (\#1, \#2) > (\#1, \#5) > (\#1, \#3) > (\#3, \#5) > (\#2, \#5)$.

Using SIM_1 and SIM_3 non-similarities:

$(\#1, \#2) > (\#3, \#4) > (\#3, \#5) > (\#2, \#4) > (\#2, \#5) > (\#2, \#3) > (\#4, \#5) > (\#1, \#4) > (\#1, \#5) > (\#1, \#3)$.

Using SIM_1, SIM_2 and SIM_3 non-similarities:

$(\#3, \#4) > (\#2, \#4) > (\#1, \#2) > (\#2, \#3) > (\#1, \#4) > (\#1, \#5) > (\#1, \#3) > (\#3, \#5) > (\#2, \#5) > (\#4, \#5)$.

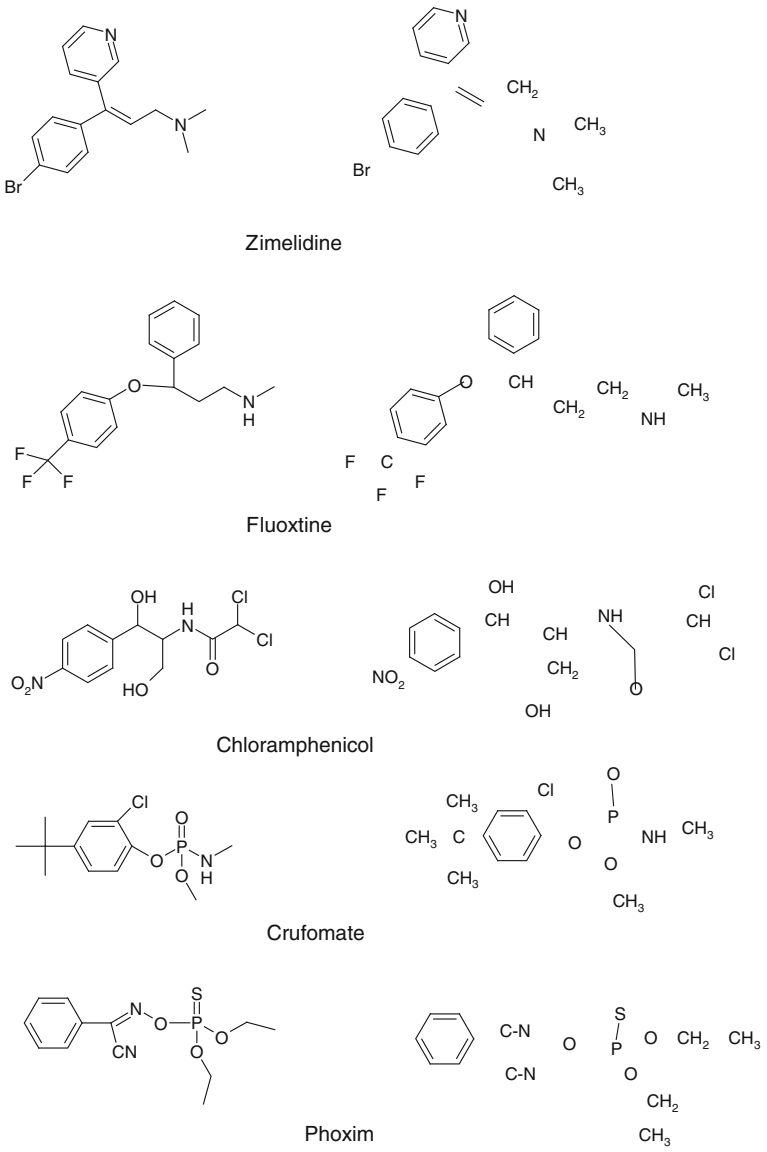

Fig. 2 Identified molecular fragments in the analyzed molecules

Using SIM_1, SIM_2, SIM_3 and SIM_4 non-similarities: $(\#3, \#4) > (\#2, \#4) > (\#1, \#2) > (\#2, \#3) > (\#1, \#4) > (\#1, \#5) > (\#1, \#3) > (\#3, \#5) > (\#4, \#5) > (\#2, \#5)$

Table 12 includes data useful for computing the Tanimoto coefficient by formula (1), with a view to compare. The values of n_A, n_B and n_{AB} are inferred from data in Table 9. The number n_A is, in fact, the number of non-empty classes in molecule A, n_B is the number of non-empty classes in molecule B and n_{AB} is the number of

Table 10 Similarity of molecules aggregates

Class	Pair				
	(#1,#2)	(#1,#3)	(#1,#4)	(#1,#5)	(#2,#3)
(a)					
I	4	3	2	1	5
II	8	7	9	7	9
III	1	1	1	1	
IV	2	2	2	2	2
V	1	1	3	4	
VI		1	1	3	1
VII	2	1	1	1	1
VIII	1	2	1		3
IX		1			1
X					
SE[a]	1.6311	1.9081	1.7036	1.6892	1.6136
SIM_4[b]	0.8196	0.7519	0.8466	0.7566	0.8325
	(#2,#4)	(#2,#5)	(#3,#4)	(#3,#5)	(#4,#5)
(b)					
I	4	3	3	2	1
II	11	9	10	8	10
III					
IV	2	2	2	2	2
V	2	3	2	3	5
VI	1	3	2	4	4
VII	1	1			
VIII	2	1	3	2	1
IX			1	1	
X					
SE[a]	1.5668	1.6797	1.6670	1.7440	1.4831
SIM_4[b]	0.8203	0.6271	0.8705	0.6987	0.8671

[a] Computed by formula (4) for each aggregate
[b] Computed by formula (10)

non-empty classes in A molecule *and* molecule B. Used in this manner Tanimoto coefficient compares the presence/absence of fragments classes, not the presence/absence of molecular fragments themselves.

Based on the values of Tanimoto similarity in Table 12 the analyzed pairs of molecules can be ordered.

$(\#3, \#4) > (\#4, \#5) > (\#1, \#2) \sim (\#2, \#3) \sim (\#2, \#4) > (\#1, \#4) > (\#1, \#5) \sim (\#3, \#5) > (\#1, \#3) > (\#2, \#5)$.

Further on the paper presents two non-chemical examples of SIM computation as 'Representative Sample Index'.

Table 11 The values of the 'total' similarity

Pair	SIM_1	SIM_2	SIM_3	SIM_4	SIM_{total}^{13a}	SIM_{total}^{123b}	SIM_{total}^{1234c}
(#1, #2)	0.8847	0.8312	0.8953	0.8196	0.8847	0.8312	0.8196
(#1, #3)	0.7108	0.8414	0.8752	0.7519	0.7108	0.7108	0.7108
(#1, #4)	0.8096	0.8113	0.7515	0.8466	0.7515	0.7515	0.7515
(#1, #5)	0.7403	0.8111	0.9260	0.7566	0.7403	0.7403	0.7403
(#2 #3)	0.8034	0.9876	0.8885	0.8325	0.8034	0.8034	0.8034
(#2, #4)	0.9094	0.9532	0.8394	0.8203	0.8394	0.8394	0.8203
(#2, #5)	0.8367	0.6741	0.8959	0.6271	0.8367	0.6741	0.6271
(#3, #4)	0.8778	0.9641	0.8588	0.8705	0.8588	0.8588	0.8588
(#3, #5)	0.9326	0.6824	0.8429	0.6987	0.8429	0.6824	0.6824
(#4, #5)	0.9141	0.6579	0.7916	0.8671	0.7916	0.6579	0.6579

[a] Identified by criterion (11) using SIM_1 and SIM_3 non-similarities
[b] Identified by criterion (11) using SIM_1, SIM_2 and SIM_3 non-similarities
[c] Identified by criterion (11) using SIM_1, SIM_2, SIM_3 and SIM_4 non-similarities

Table 12 The value of Tanimoto coefficient

Pair	n_A	n_A	n_{AB}	T
(#1, #2)	6	5	4	0.571
(#1, #3)	6	6	3	0.333
(#1, #4)	6	6	4	0.500
(#1, #5)	6	4	3	0.429
(#2 #3)	5	6	4	0.571
(#2, #4)	5	6	4	0.571
(#2, #5)	5	4	2	0.286
(#3, #4)	6	6	5	0.714
(#3, #5)	6	4	3	0.429
(#4, #5)	6	4	4	0.667

Table 13 includes imaginary data regarding a number of objects in each group, in each class.

According to the formula (12) Representative Sample Index for Sample #1 is $RSI_1 = 1.6120/1.89081.8908/1.9294 = 0.8355$. The Representative Sample Index for Sample #2 is $RSI_2 = 1.7946/1.93051.9294/1.9305 = 0.9291$. Therefore, we can say that the Sample #2 is "more representative" than Sample #1 for the analyzed population. In addition, we observe that, according to the formula (10), the similarity of Sample #1 and Sample #2 is low enough ($SIM_{12} = 1.6120/1.92941.7946/1.9294 = 0.7771$).

Table 14 includes data taken from literature [26], regarding age, sex and mean weight (kilograms) in groups of children and teenagers from USA.

The size of G_A and G_B groups is slightly different ($N_A = 4, 247$ male and $N_B = 4, 119$ female). To divide the above group of 36 values of mean weight into classes

Table 13 Imaginary data regarding objects groups

Groups	Classes											SE[a]
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	
Population	1	5	4	12	9	32	0	5	5	8	3	1.9294
Sample #1	1	2	0	5	0	17	0	3	4	8	0	1.6120
Sample #2	0	3	4	7	9	15	0	2	1	0	3	1.7946
Sample #1 + population	2	7	4	17	9	49	0	8	9	16	3	1.8908
Sample #2 + population	1	8	8	19	18	47	0	7	6	8	6	1.9305

[a] Computed by formula (4)

Table 14 Data regarding two young people groups

Age	Male group size	G_A mean weight	Female group size	G_B mean weight
2	262	13.7	248	13.3
3	216	15.9	178	15.2
4	179	18.5	191	17.9
5	147	21.3	186	20.6
6	182	23.5	171	22.4
7	185	27.2	196	25.9
8	214	32.7	184	31.9
9	174	36.0	183	35.4
10	187	38.6	164	40.0
11	182	43.7	194	47.9
12	299	50.4	316	52.0
13	298	53.9	321	57.7
14	266	63.9	324	59.9
15	283	68.3	266	61.1
16	306	74.4	273	63.0
17	313	75.6	256	61.7
18	284	75.6	243	65.2
19	270	78.2	225	67.9

we identified the maximum value 78.2, the minimum value 13.3 and calculated the difference 78.2 – 13.3 = 64.9. Using a tenth of this difference we divided [13.3, 78.2] range in ten classes ($k_{def} = 10$).

Class i includes the values within $[13.3+(i-1)\cdot 6.49,\ 13.3+i\cdot 6.49]$ range. Table 15 presents the number of youngsters in each class, in each group and the computed value for the Shannon Entropy.

Using the computed values of SE and formula (10) we calculated the similarity of groups G_A and G_B ($SIM = 0.8746$). Within the analyzed population, according to formula (12), the 'Representative Sample Index' of the male group G_A ($RSI_A = 0.8448$)

Table 15 The computed value of Shannon Entropy for youngster groups

Class	Groups				
	G_A male sample	G_B female sample	Population	G_A + population	G_B + population
1	657	617	1,274	1,931	1,891
2	329	553	882	1,211	1,435
3	399	184	583	982	767
4	361	183	544	905	727
5	182	164	346	528	510
6	299	510	809	1,108	1,319
7	298	321	619	917	940
8	266	1,362	1,628	1,894	2,990
9	283	225	508	791	733
10	1,173		1,173	2,346	1,173
SE[a]	2.0620	1.8034	1.9980	2.2084	2.1652

[a] Computed by formula (4)

is higher than the 'Representative Sample Index' of the female group G_B($RSI_B = 0.7686$), from the point of view of mean weight.

References

1. P. Baldi, R.W. Benz, D.S. Hirschberg, S.J. Swamidass, J. Chem. Inf. Model. **47**, 2098 (2007)
2. M.R. Landon, S.E. Schaus, Mol. Divers. **10**, 333 (2006)
3. N. Nikolova, J. Jaworska, QSAR Comb. Sci. **22**, 1006 (2003)
4. P. Willett, J.M. Barnard, G.M. Downs, J. Chem. Inf. Model. **38**, 983 (1998)
5. A. Tversky, Psychol. Rev. **20**, 1 (1977)
6. N. Salim, J. Holliday, P. Willett, J. Chem. Inf. Comput. Sci. **43**, 435 (2003)
7. A.M. Meyer, W.G. Richards, J. Comput.Aided Mol. Des. **5**, 426 (1991)
8. D.E. Walters, A.J. Hopfinger, J. Mol. Struct. THEOCHEM **134**, 317 (1986)
9. R. Carbo, L. Leyda, M. Arnau, Int. J. Quantum Chem. **17**, 1185 (1980)
10. P.G. Mezey, J. Chem. Inf. Comput. Sci. **36**, 1076 (1996)
11. P.C. Mahalanobis, Proc. Natl. Inst. Sci. India **2**, 49 (1936)
12. C.A. Shannon, Bell Syst. Tech. J. **27**, 623 (1948)
13. S. Noorizadeh, E. Shakerzadeh, Phys. Chem. Chem. Phys. **12**, 4742 (2010)
14. E. Gregori-Puigjané, J. Mestres, J. Chem. Inf. Model. **46**, 1615 (2006)
15. J. Godden, J.J. Bajorath, J. Chem. Inf. Comput. Sci. **41**, 1060 (2001)
16. S.K. Lin, Molecules **1**, 57 (1996)
17. M. Hô, V.H. Smith, D.F. Weaver, C. Gatti, R.P. Sagar, R.O. Esquivel, J. Chem. Phys. **108**, 5469 (1998)
18. Y. Wang, H. Geppert, J. Bajorath, J. Chem. Inf. Model. **49**, 1687 (2009)
19. L. Tarko, Rev. Chim. (Bucuresti) **55**, 539 (2004)
20. J.J.P. Stewart, J. Mol. Model. **13**, 1173 (2007)
21. J.L. Rodgers, W.A. Nicewander, Am. Stat. **42**, 59 (1988)
22. J.L. Myers, W.D. Arnold, *Research Design and Statistical Analysis*, 2nd edn. (Lawrence Erlbaum, 2003), p. 508, ISBN 0805840370
23. M. Kendall, Biometrika **30**, 81 (1938)
24. http://en.wikipedia.org/wiki/Clustering_algorithm Accessed 03 June 2011
25. L. Tarko, ARKIVOC **11**, 24 (2008)
26. C.L. Ogden, C.D. Fryar, M.D. Carroll, K.M. Flegal, Advance Data, **347**, October 27 (2004)