

Computation of the molecular shapes' similarity and diversity using USR method and General Shape Index

Laszlo Tarko¹

Received: 10 November 2014 / Accepted: 9 April 2015 / Published online: 15 May 2015
© Springer International Publishing Switzerland 2015

Abstract Regarding the molecular shape, shapes' similarity and shapes' diversity the paper presents (1) a new molecular descriptor, (2) using of the new descriptor together with the previous Ultrafast Shape Recognition (USR) formula, (3) a quantitative method to verify the observance of the 'QSAR axiom', (4) a formula for identification of the activity and shape 'cliffs', (5) a method to divide in classes any group of molecules, (6) a criterion to identify the 'atypical' molecules and (7) a method, based on Shannon entropy formula, for computation of the molecules' diversity and the similarity of two groups of molecules. The proposed formulas/procedures are simple and suggestive. The algorithm which uses the proposed descriptor *and* USR formula describes correctly enough the molecular similarity in three analyzed groups of molecules.

Keywords Molecular shape · Molecular similarity · Shannon entropy · USR method · QSAR axiom

1 Introduction

It is believed that 3D molecular shape is a valuable pattern for biological activity because the shape is related to the electrostatic interactions between the low molecular mass molecule ('effector') and active site of the target macromolecule ('receptor'). Accordingly, many studies emphasized the importance of shape, 'electrical shape'

✉ Laszlo Tarko
ltarko@cco.ro

¹ Centre of Organic Chemistry, Romanian Academy, Sector 6, Spl. Independentei 202B,
PO Box 35-108, MC 060023 Bucharest, Romania

(shape *and* atomic net charges) or shape description based on the electronic density of molecules as indicator for molecular bio-activity [1–24].

Flexible molecules can adopt different shapes [25, 26]. In fact, the ‘shape’ of effector is the shape of the conformer having lowest potential energy. Searching and identification of this conformer is called ‘geometry optimization’. The ‘false minimal energy’ conformers should be carefully avoided because the value of many descriptors depends on geometry. After all, the correctness of obtained geometry depends on the correctness of parameterization of used software for various atoms and chemical bonds.

The ‘effector’ interacts with the active site of the ‘receptor’ by a ‘key-lock’ mechanism. The effector and active site are flexible, due to the presence of rotatable chemical bonds. When the ‘key’ approaches the ‘lock’ the geometry of the two systems starts to change gradually. The final shape’s complementarity of the two components of the ‘key-lock’ system is going to be the complementarity of two *modified* geometries. An alternative approach is geometry optimisation of the effector molecule-active site aggregate, if the initial geometry of the active site is known. Docking of the effector molecule in the active site can be done using computer-generated structures [27–32].

As a rule, ‘similar structures have similar properties’. If the ‘properties’ are ‘bio-chemical activities’ this statement is named ‘QSAR axiom’. Frequently, this ‘axiom’ is challenged because some similar structures present non-similar values of activities and some non-similar structures present similar values of activities. Despite this, available databases, including a great number of low molecular mass molecules, are screened to find compounds similar from the point of view of shape with a given molecule having known biological activity. The goal of screening is finding a novel drug leads.

There are two classes of methods used in shapes’ similarity computations:

- (a) superposition methods
- (b) comparing the value of shape’s molecular descriptors

The methods in class (a) require previous alignment of the molecules. Then, one computes the distances between atoms in 3D space. The precision of these methods is high but the computation time is great.

The methods in class (b) require previous calculation of the value of certain arguable shape’s molecular descriptors. Then, one calculates the Manhattan or Euclidian distance of the values. The precision of these methods is lower but the computation time is much smaller. This advantage is important if the screened database include quite great number of molecules.

There are a huge number of molecular topological indices [33–38] which describe the molecular shape *and* size. Calculated as mathematical functions of values in distance and/or adjacency matrices these 2D molecular descriptors are useful in similarity computations and QSAR (*Quantitative Structure-Activity Relationship*) studies.

A widely used method in class (b) is patented USR (*Ultrafast Shape Recognition*) algorithm [39–43]. This method calculates three statistical functions (mean, standard deviation and skewness) of atomic distances, called *moments*, related to four different points in analyzed molecule (geometric center, the closest atom to geometric center, the farthest atom to geometric center and the farthest atom to the farthest atom to the geometric center). The shape similarity of two certain molecules depends on the mean of Manhattan distances of the twelve μ computed moments. Actually, many

moments of USR measure *size* (not shape) of molecules. Consequently, if the shape is similar enough and the size is different enough, the USR shapes' similarity is quite underestimated. On the other hand USR method seems to be adequate to describe the unevenness (rugosity) of molecular surface.

Regarding the mathematical definition of shape and shapes' similarity the paper

- proposes a new molecular descriptor
- describes using of this descriptor together with USR formula
- proposes a quantitative method to verify the observance of 'QSAR axiom'
- proposes a formula for identification of 'activity cliffs'
- proposes a method to divide in classes the analyzed group of molecules
- propose a criterion to identify 'atypical' molecules
- describes a method for computation of molecules' diversity
- propose a formula for calculation of similarity of two molecules groups

2 Methods and formulas

The virtual building of the molecules and the geometry optimization were done using the molecular mechanics program PCModel [44]. A more rigorous geometry optimization was subsequently performed by MOPAC software [45], semi-empirical quantum-mechanics PM6 method included [46].

MOPAC computes the 'molecular dimensions' D_1 , D_2 and D_3 ($D_1 \geq D_2 \geq D_3$) of analyzed molecule. In computation of dimensions MOPAC disregards the atomic diameters. Consequently, in 1D molecules, such as (halogenated) acetylene, $D_1 \gg D_2 \sim D_3 \sim 0$, in 2D molecules, such as PAHs or $C_6H_nX_{6-n}$ halogenated benzene, $D_1 \sim D_2 \gg D_3 \sim 0$ and in 3D almost spherical molecules, such as C_{60} fullerene, $D_1 \sim D_2 \sim D_3 \gg 0$.

The proposed General Shape Index of analyzed molecule is the value of function gsi , having value within [1, 3] range.

$$gsi = (1 \cdot SIM1 + 2 \cdot SIM2 + 3 \cdot SIM3) / (SIM1 + SIM2 + SIM3) \quad (1)$$

where

$$SIM1 = [1 - (1/3 \cdot \sum R_{i1}^2)^{0.5}]^3$$

$$SIM2 = [1 - (1/3 \cdot \sum R_{i2}^2)^{0.5}]^3$$

$$SIM3 = [1 - (1/3 \cdot \sum R_{i3}^2)^{0.5}]^3$$

$$i = 1, 2, 3$$

$$R_{11} = (D_2 + 1) / (D_1 + 1)$$

$$R_{12} = R_{13} = R_{11} - 1$$

$$R_{21} = R_{22} = (D_3 + 1) / (D_1 + 1)$$

$$R_{23} = R_{21} - 1$$

$$R_{32} = (D_3 + 1) / (D_2 + 1)$$

$$R_{31} = R_{33} = R_{32} - 1$$

Table 1 Three cases of the similarity SIM_{shape} as SIM_{gsi}

Case	gsi_1	gsi_2	D_2/D_1 or D_3/D_2 in first molecule	D_2/D_1 or D_3/D_2 in second molecule	Comments
#1	<1.47	Any	Any	Any	At least one ~1D molecule
#2	>1.99	<2.01	>0.88	>0.88	Two ~2D and symmetrical molecules
#3	>2.81	Any	Any	Any	Two ~3D molecules

The R_{ij} ratios compare the dimensions. The factor +1 in ratios R_{ij} allows using of zero dimensions computed by MOPAC. The value of R_{ij} is within [-1, 1] range. The similarities SIM_1 , SIM_2 and SIM_3 , having value within [0, 1] range, are the similarities with 'ideal' 1D, 2D and 3D molecules. The value of gsi is weighted sum of similarities SIM .

If $gsi < 1.5$ the shape of circumscribed ovoid is very elongated. If $1.7 < gsi < 2.2$ the molecular shape is somehow planar. If $gsi > 2.7$ the shape of circumscribed ovoid is almost spherical.

The similarity SIM_{gsi} of two molecules is calculated by proposed formula (2), where $gsi_1 \leq gsi_2$ and the value of x factor is 4/3, empirically established.

$$SIM_{\text{gsi}} = (gsi_1/gsi_2)^x \quad (2)$$

We remind the similarity formula of USR algorithm [39].

$$SIM_{\text{usr}} = 1/(1 + M) \quad (3)$$

where M is mean of the Manhattan distances of moments, $M = 1/12 \cdot \sum |\mu_{1i} - \mu_{2i}|$, $i = 1, \dots, 12$

The similarity of shapes SIM_{shape} of two molecules depends on similarity SIM_{gsi} and/or similarity SIM_{usr} . Here, the similarity $SIM_{\text{shape}} = SIM_{\text{gsi}}$ in cases in Table 1. In other conditions $SIM_{\text{shape}} = SIM_{\text{usr}}$. The limit values 0.88, 1.47, 1.99, 2.01 and 2.81 in Table 1 are empirically established after analysis of many databases and tens thousands different molecules pairs.

The value of SIM_{shape} is within [0, 1] range.

If one analyzes a group of N molecules one calculates $N(N-1)/2$ values of SIM_{shape} . Therefore, the pairs of molecules having the maximum/minimum value of SIM_{shape} can be identified.

For each pair of molecules having known value of bio-activity, the similarity SIM_{act} of bio-activities A_i and A_j is calculated by formula (4), where A_{max} and A_{min} are the maximum and minimum value of bio-activity in analyzed group of molecules.

$$SIM_{\text{act}} = 1 - |A_i - A_j|/(A_{\text{max}} - A_{\text{min}}) \quad (4)$$

We propose here the Kendall rank correlation [47,48] of the values of SIM_{shape} and SIM_{act} as quantitative measure of QSAR axiom's observance. The value of this statistical function is within $[-1, +1]$ range.

We propose here the ratios (5) as signal of presence of activity and shape 'cliffs'.

$$C_{act} = SIM_{shape}/(1 + SIM_{act}) \quad (5a)$$

$$C_{shape} = SIM_{act}/(1 + SIM_{shape}) \quad (5b)$$

The pairs of molecules having high value of C_{act} within $[0, 1]$ range present a high value of SIM_{shape} vs. a low value of SIM_{act} . The pairs of molecules having high value of C_{shape} present a high value of SIM_{act} vs. a low value of SIM_{shape} . After all, the presence of 'cliffs' having high value emphasizes the violation of QSAR axiom for some molecules.

If the similarity is high enough ($SIM_{shape} \geq k$) the two analyzed molecules should be considered 'high similar' or 'in the same class', from the point of view of shape. We propose for k the value 0.9, empirically established.

The N molecules in analyzed group should be included in classes (types, categories) called 'shape clusters', according to similarity. Each pair of molecules in certain cluster fulfils the condition $SIM_{shape} \geq k$. Here, we propose a very intuitive clusterization procedure including five steps.

Step #1 identification of the first 'seed', i.e. the object having minimum sum of similarities ΣS_{ij} with the other $N - 1$ objects; the first seed is included into first class

Step #2 identification of the next 'seeds', i.e. objects having similarity (with each seed) smaller than k and minimum sum of similarities ΣS_{ij} (with the other 'seeds')

Each 'seed' is included into new class. After n times running of Step #2 there are $n + 1$ classes, each class includes 1 object, the number of 'seeds' becomes, as a rule, zero and the number of non-classified objects is $N - n - 1$.

Step #3 identification of the object having maximum sum of similarities ΣS_{ij} with the objects included in classes

Step #4 identification of the class having features (a) and (b)

(a) all similarities of included objects with the object identified in Step #3 fulfil the condition (2)

(b) greatest mean value of similarities of included objects with the object identified in Step #3

The object identified in Step #3 is the most suitable to be classified. The class identified in Step #4 is the most suitable to include the object identified in Step #3. After $N - n - 1$ times running of Step #3 + Step #4 there are $n + 1$ classes also, each class includes few objects and number of the non-classified objects becomes, as a rule, zero. However, sometimes, the last analyzed object remains non-classified, because each class includes one or more object(s) which have too low (i.e. smaller than k) similarity with the last object.

Step #5 the non-classified object, if it exist, becomes the last 'seed' of a new (last) class

After clusterization one calculates the entropy of the analyzed group of molecules, from the point of view of molecular shapes, using the discontinuous formula of Shannon entropy [49], see formula (6). Here, we used the natural logarithm. The value of entropy SE is within $[0, \text{Log}(N)]$ range.

$$SE = -\sum p_i \cdot \text{Log}(p_i) \quad (6)$$

where i is number of clusters, $p_i = n_i/N$, n_i is number of molecules in cluster i , N is number of molecules in analyzed group.

If $i = 1$ (all molecules are very similar with all molecules) then $SE = 0$.

If number n_i of molecules in cluster i is small, see proposed criterion (7), the molecules in cluster i are considered ‘atypical (outliers) from the point of view of shape’. These molecules are similar with a very small number of other molecules in analyzed group.

$$n_i < 0.5 \cdot N^{1/2} \quad (7)$$

The diversity D_{shape} of molecules is weighted value of Shannon entropy, it is calculated by formula (8) and his value is within $[0, 1]$ range.

$$D_{\text{shape}} = SE/\text{Log}(N) \quad (8)$$

Frequently, in QSAR studies, one uses two groups of molecules. The calibration set includes molecules having known values of activity. The prediction set includes new, not yet synthesized molecules, having unknown values of activity.

If the observance of QSAR axiom, from the point of view of shapes, is high, the similarity of calibration and prediction sets from this point of view seems to be the decisive factor for a correct estimation of activities of the molecules in prediction set [50].

Here, the similarity SIM_{12} of two groups of molecules (each considered as a whole) is calculated by proposed formula (9) and his value is within $[0, 1]$ range.

$$\text{SIM}_{12} = R_1 \cdot R_2 \quad (9)$$

where

If $SE_1 \leq SE_{12}$ then $R_1 = (0.1 + SE_1)/(0.1 + SE_{12})$ else $R_1 = (0.1 + SE_{12})/(0.1 + SE_1)$

If $SE_2 \leq SE_{12}$ then $R_2 = (0.1 + SE_2)/(0.1 + SE_{12})$ else $R_2 = (0.1 + SE_{12})/(0.1 + SE_2)$

SE_1 is Shannon entropy of group #1

SE_2 is Shannon entropy of group #2

SE_{12} is Shannon entropy of aggregate group #1 + group #2

The factor 0.1 allows using of zero values of entropies SE. The formula (9) is a simplified version of previously proposed formulas [50,51].

3 Results and discussion

General Shape Index describes only the general shape of circumscribed ovoid, not other complex features of molecular shape as ruggedness of molecular surface or local shape of certain fragments in analyzed molecule. This descriptor is not viewed as a substitute of other shape descriptors and it is used here to increase the precision of USR method in computation of molecular similarity.

The molecules in Table 2, having high diversity of shapes, were analyzed to emphasize the difference between USR method and algorithm proposed here.

Table 3 includes, for comparison, the value of *gsi* and some usual topological indices, i.e. Randic (R) [52], Wiener (W) [53] and Balaban (J) [54], calculated for molecules in Table 2.

The Pearson square linear correlations of *gsi* with Randic index R in Table 3 ($r^2 = 0.0092$), with Wiener index W ($r^2 = 0.0028$) and with Balaban index J ($r^2 = 0.0036$) are very low. Therefore, as was to be expected, *gsi* and topological indices R, W and J describe different aspects of shape.

We remind that the topological indices describe the molecular shape *and* size. To decrease effect of size, Table 4 includes the value of the same topological indices referred to the number *g* of heavy atoms (different from hydrogen).

The correlations of *gsi* with ratios R/*g* in Table 4 ($r^2 = 0.1943$), W/*g* ($r^2 = 0.0509$) and J/*g* ($r^2 = 0.1773$) are higher than correlations with topological indices in Table 3.

According to USR method there are no pairs in Table 2 which fulfill the condition $SIM_{USR} \geq 0.9$. The maximum similarity of molecules is computed for pair **12–13**, $SIM_{USR} = 0.8767$. Consequently, the number of clusters is 32, each cluster includes one molecule, all molecules are ‘atypical’ (outliers) and the diversity D_{shape} of molecules has maximum value 1.

For molecules which are, intuitively, similar enough from the point of view of shape but different enough from the point of view of size USR method computes very low similarities. For instance, for pair **27–28** $SIM_{USR} = 0.4634$, for pair **15–24** $SIM_{USR} = 0.3862$, for pair **17–19** $SIM_{USR} = 0.2898$, for pair **22–23** $SIM_{USR} = 0.2275$ and for pair **10–32** $SIM_{USR} = 0.1040$. In these pairs the values of *gsi* are very close.

In pairs *n-alkane-32* the pair **8–32** possess the highest similarity, $SIM_{USR} = 0.7825$, because of similarity of sizes.

In pairs *any-29* the pair **15–29** possess the highest similarity, $SIM_{USR} = 0.8116$, although **15** is a 2D molecule and **29** is a 3D molecule.

Accordingly, the USR method *alone* is not suitable for description of shape similarity of molecules in Table 2.




Using proposed algorithm we obtained very different results. Many molecules are included in the same cluster because $SIM_{shape} \geq 0.9$.

The maximum similarity of molecules in Table 2 is computed for pair **11–24**, $SIM_{shape} = 0.9997$.

The number of clusters is 17 and the diversity D_{shape} of molecules in Table 2 is lower, $D_{shape} = 0.7151$.

Nine polyhedral 3D molecules **1–4**, **25–29** are included in the same cluster. Five 2D molecules **11–14**, **24** are included in the same cluster. Elongated 3D molecules **9**,

Table 2 The structure of analyzed molecules

Index	Type	Structure
1	Methane derivatives CX ₄	X = F
2		X = Cl
3		X = Br
4		X = I
5	<i>n</i> -Alkane CH ₃ (CH ₂) _n CH ₃	n = 1
6		n = 2
7		n = 4
8		n = 7
9		n = 11
10		n = 16
11	Benzene derivatives C ₆ X ₆	X = F
12		X = Cl
13		X = Br
14		X = I
15		X = H
16	 Linear PAHs	c = 2
17	 Linear PAHs	c = 3
18		c = 4
19		c = 5
20	 Polyphenylenes	c = 2
21		c = 3
22		c = 4
23		c = 5
24	Miscellanea	Coronene
25		Adamantane
26		Twistane
27		C ₂₀ fullerene
28		C ₆₀ fullerene
29		Cubane
30		1,3,5-triiodo-benzene
31		1,4-dibutyl-benzene
32		1,6-dibromo-acetylene

10, 23 are included in the same cluster. Surprisingly, the maximum similarity in group **15–19** is low, $SIM_{shape} = 0.5495$ for pair **15–16**.

There are 15 ‘atypical’ molecules. For instance, in pairs **any–30** the maximum similarity is $SIM_{shape} = 0.7890$ (for pair **14–30**) and so, the molecule **30** is one of ‘atypical’ molecules in analyzed group.

Table 3 The value of *gsi* and topological indices

Index in Table 2	<i>gsi</i>	R	W	J
1	2.89	2.00	16	3.02
2	2.87	2.00	16	3.02
3	2.87	2.00	16	3.02
4	2.86	2.00	16	3.02
5	2.68	1.41	4	1.63
6	2.34	1.91	10	1.98
7	1.74	2.91	35	2.34
8	1.32	4.41	120	2.60
9	1.17	6.41	364	2.76
10	1.12	8.91	969	2.86
11	2.01	5.46	174	2.76
12	2.00	5.46	174	2.76
13	2.00	5.46	174	2.76
14	2.00	5.46	174	2.76
15	2.01	3.00	27	2.00
16	1.98	4.97	109	1.93
17	1.94	6.93	279	1.68
18	1.89	8.90	569	1.47
19	1.84	10.87	1011	1.29
20	1.95	5.97	198	1.80
21	1.47	8.93	657	1.45
22	1.29	11.90	1548	1.19
23	1.20	14.87	3015	1.01
24	2.01	11.90	1002	1.41
25	2.98	4.90	96	1.90
26	2.87	4.93	95	1.94
27	2.95	10.00	500	1.50
28	2.98	30.00	8340	0.91
29	2.92	4.00	48	2.00
30	2.02	4.18	84	2.34
31	1.54	6.86	367	2.05
32	1.06	3.91	84	2.53

To present applicability of formulas (4)–(9) we analyzed the phenol derivatives in Table 5. The toxicity ($T = \text{Log}(1/\text{LD}_{50})$) of these molecules against *Tetrahymena pyriformis* protozoan is quoted in literature [55, 56].

From the point of view of *gsi* value, a large majority of molecules in Table 5 are very similar 2D molecules. Accordingly, the *gsi* molecular descriptor *alone* is not suitable for description of shape similarity. Using proposed algorithm we obtained next presented results.

There is only one pair of molecules having very high similarity i.e. $\text{SIM}_{\text{shape}} \geq 0.98$, more precisely, **41** and **44**, $\text{SIM}_{\text{shape}} = 0.9966$.

Table 4 The value of *gsi* and topological indices/g ratio

Index in Table 2	<i>gsi</i>	R/g	W/g	J/g
1	2.89	0.40	3.20	0.61
2	2.87	0.40	3.20	0.61
3	2.87	0.40	3.20	0.61
4	2.86	0.40	3.20	0.61
5	2.68	0.47	1.33	0.54
6	2.34	0.48	2.50	0.49
7	1.74	0.49	5.83	0.39
8	1.32	0.49	13.33	0.29
9	1.17	0.49	28.00	0.21
10	1.12	0.50	53.83	0.16
11	2.01	0.46	14.50	0.23
12	2.00	0.46	14.50	0.23
13	2.00	0.46	14.50	0.23
14	2.00	0.46	14.50	0.23
15	2.01	0.50	4.50	0.33
16	1.98	0.50	10.90	0.19
17	1.94	0.50	19.93	0.12
18	1.89	0.49	31.61	0.08
19	1.84	0.49	45.96	0.06
20	1.95	0.50	16.50	0.15
21	1.47	0.50	36.50	0.08
22	1.29	0.50	64.50	0.05
23	1.20	0.50	100.50	0.03
24	2.01	0.50	41.75	0.06
25	2.98	0.49	9.60	0.19
26	2.87	0.49	9.50	0.19
27	2.95	0.50	25.00	0.08
28	2.98	0.50	139.00	0.02
29	2.92	0.50	6.00	0.25
30	2.02	0.47	9.33	0.26
31	1.54	0.49	26.21	0.15
32	1.06	0.49	10.50	0.32

The molecules **1** and **48** present lowest similarity $SIM_{shape} = 0.2904$.

The observance of the QSAR axiom is 0.1776.

The molecules' pair **16–41** presents maximum Toxicity cliff, i.e. highest shape similarity / Toxicity similarity ratio, $SIM_{shape} = 0.8202$, $SIM_{tox} = 0.2074$, $C_{tox} = 0.6793$. The molecules' pair **44–48** presents maximum shape cliff, i.e. highest Toxicity similarity / shape similarity ratio, $SIM_{shape} = 0.3593$, $SIM_{tox} = 0.9951$, $C_{shape} = 0.7321$.

The number of shape clusters is 35 and 46 molecules are 'atypical'. There is only one cluster, i.e. (**26, 28, 30, 35**), including four non-atypical molecules. In addition,

Table 5 The structure and *gsi* value of analyzed phenol derivatives

No.	Substituent(s)	T	<i>gsi</i>
1	None	-0.431	1.997
2	2,6-difluoro	0.396	2.001
3	2-fluoro	0.248	2.002
4	4-fluoro	0.017	1.990
5	3-fluoro	0.473	1.997
6	4-methyl	-0.192	2.083
7	3-methyl	-0.062	2.153
8	2-chloro	0.277	2.012
9	2-bromo	0.504	2.013
10	4-chloro	0.545	1.975
11	3-ethyl	0.229	2.199
12	2-ethyl	0.176	2.325
13	4-bromo	0.681	1.968
14	2,3-dimethyl	0.122	2.207
15	2,4-dimethyl	0.128	2.117
16	2,5-dimethyl	0.009	2.105
17	3,4-dimethyl	0.122	2.174
18	3,5-dimethyl	0.113	2.177
19	3-chloro-4-fluoro	0.842	1.994
20	2-chloro-5-methyl	0.640	2.155
21	4-iodo	0.854	1.960
22	3-iodo	1.118	1.991
23	2- <i>iso</i> -propyl	0.803	2.634
24	3- <i>iso</i> -propyl	0.609	2.523
25	4- <i>iso</i> -propyl	0.473	2.293
26	2,5-dichloro	1.128	2.003
27	2,3-dichloro	1.271	2.015
28	2-methyl-4-chloro	0.700	2.185
29	3-methyl-4-chloro	0.795	2.155
30	2,4-dichloro	1.036	1.995
31	3- <i>tert</i> -butyl	0.730	2.524
32	4- <i>tert</i> -butyl	0.913	2.300
33	3,5-dichloro	1.562	2.016
34	2-phenyl	1.094	2.030
35	2,4-dibromo	1.403	1.996
36	2,3,6-trimethyl	0.418	2.575
37	3,4,5-trimethyl	0.930	2.332
38	2,4,6-trimethyl	1.695	2.485
39	3,5-dimethyl-4-chloro	1.203	2.162
40	2,6-dichloro-4-bromo	1.779	1.994
41	2,4,6-trichloro	2.100	2.008
42	2-methyl-4-bromo-6-chloro	1.277	2.168

Table 5 continued

No.	Substituent(s)	T	gsi
43	2,6-dimethyl-4-bromo	1.278	2.582
44	2,4,6-tribromo	2.050	2.003
45	2- <i>tert</i> -butyl-4-methyl	1.297	2.532
46	2- <i>iso</i> -propyl-4-chloro-5-methyl	1.862	2.509
47	2,4-dimethyl-6- <i>tert</i> -butyl	1.245	2.587
48	2,6-diphenyl	2.113	2.087
49	2,4-dibromo-6-phenyl	2.207	2.257
50	2,6-di- <i>tert</i> -butyl-4-methyl	1.788	2.392

there are only three clusters, i.e. (**40, 42, 43**), (**15, 16, 20**) and (**17, 19, 29**), including three molecules. Consequently, the diversity of molecules is high, $D_{\text{shape}} = 0.8816$.

The significance of similarity's value in formula (9) should be high because the observance of QSAR axiom for molecules in Table 5 is high enough. To verify this assumption we arranged the molecules according to value of toxicity T and then we made two tests.

In first test the Group #1 includes the 25 molecules having ranks 1, 2, 3, ..., 25, i.e. the molecules having lowest value of T and Group #2 include the 25 molecules having highest value of T. The similarity of toxicities is, intuitively, low, because the average toxicity in each group is very different. The shape similarity of these two groups (each considered as a whole) is $SIM_{12} = 0.6848$.

In second test the Group #1 includes the 25 molecules having ranks 1, 3, 5, ..., 49 and Group #2 includes the 25 molecules having ranks 2, 4, 8, ..., 50. The similarity of toxicities is, intuitively, high, because the average toxicity in each group is similar. The similarity of these two groups was $SIM_{12} = 0.8024$.

Indeed, for molecules in Table 5, the similarity of toxicities is correlated with the similarity of shapes and the computed value 0.1776 of QSAR axiom's observance seems to be 'significant'.

The same two tests were made for N-methyl-phenyl urethanes $\text{CH}_3\text{NHCOO-C}_6\text{H}_n\text{-Z}_{5-n}$ in Table 6, which are insecticides having toxicity quoted in literature [57–59]. The value of observance is much lower, -0.0012 . The value of similarity of groups in first test $SIM_{12} = 0.7499$ is much closer to similarity of groups in second test, $SIM_{12} = 0.7810$.

In Table 6 the molecules **32** and **67** present lowest similarity $SIM_{\text{shape}} = 0.2988$.

The molecules' pair (**35, 58**) presents maximum toxicity cliff $C_{\text{tox}} = 0.72$. The molecules' pair (**27, 67**) presents maximum shape cliff $C_{\text{shape}} = 0.76$.

The number of clusters is 52. There is only one cluster, i.e. (**9–12, 25**), including five non-atypical molecules. Consequently, the diversity of molecules is high, $D_{\text{shape}} = 0.8875$.

All computations are made by PRECLAV software [50,60,61] and a 3200 MHz Pentium4 computer. Input file in calculation of USR moments was MOPAC output file. Input table in similarity calculations included the values of USR moments and MOPAC dimensions. The computation time for the values of SIM_{shape} for molecules in Tables 5 and 6 was ~ 1100 values/s.

Table 6 The structure of analyzed urethanes

No.	Substituent(s) Z	T
1	3,5-di- <i>iso</i> -propyl	7.48
2	3-methyl	4.85
3	3- <i>iso</i> -propyl	6.47
4	3- <i>tert</i> -butyl	6.40
5	3,5-di-methyl	5.22
6	3-methyl-5- <i>iso</i> -propyl	7.25
7	2-methyl-5- <i>iso</i> -propyl	5.70
8	2- <i>O-iso</i> -propyl	6.17
9	2-fluoro	4.80
10	2-chloro	5.30
11	2-bromo	5.66
12	2-iodo	6.10
13	4-chloro	3.62
14	3- <i>N</i> (methyl) ₂	5.10
15	3,5-di-methyl-4- <i>S</i> -methyl	5.92
16	4- <i>N</i> (methyl) ₂	3.62
17	3- <i>N</i> (methyl) ₂ -5- <i>iso</i> -propyl	6.72
18	3- <i>iso</i> -propyl-4- <i>N</i> (methyl) ₂	6.82
19	3,5-di- <i>N</i> (methyl) ₂	5.59
20	3- <i>iso</i> -propyl-4- <i>N</i> (methyl) ₂ -6-methyl	6.41
21	2- <i>iso</i> -propyl-4- <i>N</i> (methyl) ₂ -5-methyl	5.89
22	3- <i>S</i> -methyl	5.16
23	4- <i>S</i> -methyl	4.47
24	None	3.70
25	2-methyl	3.85
26	4-methyl	4.00
27	2-ethyl	4.89
28	3-ethyl	5.32
29	4-ethyl	4.42
30	2- <i>iso</i> -propyl	5.22
31	4- <i>iso</i> -propyl	4.16
32	2- <i>tert</i> -butyl	5.22
33	4- <i>tert</i> -butyl	5.82
34	2- <i>sec</i> -butyl	5.96
35	3- <i>sec</i> -butyl	6.80
36	4- <i>sec</i> -butyl	5.75
37	3- <i>sec</i> -amyl	6.96
38	2-cyclopentyl	5.96
39	3-cyclopentyl	5.82
40	4-cyclopentyl	4.57
41	2-propyl	5.27
42	2- <i>iso</i> -butyl	5.64

Table 6 continued

No.	Substituent(s) Z	T
43	3-fluoro	4.07
44	4-fluoro	3.64
45	3-chloro	4.30
46	3-bromo	4.89
47	4-bromo	4.00
48	3-iodo	5.16
49	4-iodo	4.06
50	2,3-di-chloro	4.32
51	2,4-di-chloro	4.85
52	2,5-di-chloro	4.30
53	2,6-di-chloro	2.89
54	3,4-di-chloro	4.72
55	3,5-di-chloro	4.92
56	2-nitro	2.30
57	3-nitro	2.70
58	4-nitro	2.52
59	2-nitro-3-methyl	3.70
60	2-nitro-4-methyl	3.89
61	2-nitro-5-methyl	4.80
62	3-nitro-4-methyl	3.50
63	2-S-hexyl	5.40
64	3-ethyl-4-nitro	3.70
65	3- <i>iso</i> -propyl -4-nitro	5.55
66	4-S-ethyl	4.25
67	4-S-propyl	4.92
68	4-S- <i>iso</i> -propyl	5.05
69	4-S-butyl	5.43
70	2-cyclohexyl	5.85
71	3-cyclohexyl	5.70
72	4-cyclohexyl	5.05
73	3- <i>iso</i> -propyl-4-S-methyl	7.00
74	3- <i>iso</i> -propyl-6-S-methyl	6.75
75	2-S-allyl	5.44
76	4-S-allyl	5.07

4 Conclusions

The proposed shape descriptor *gsi* (General Shape Index) is simple and suggestive.

The USR formula and *gsi* descriptor *alone* cannot describe correct enough the molecular similarity. The algorithm which uses the *gsi* descriptor *and* USR formula describe correctly enough the molecular similarity.

The proposed method to verify the observance of ‘QSAR axiom’ is simple and suggestive.

The proposed formula to identify the activity and shape ‘cliffs’ is simple and suggestive.

The steps in clusterization procedure are intuitive and the computation time is short.

The proposed criterion to identify ‘atypical’ molecules is simple and suggestive.

After clusterization, the Shannon entropy formula seems to be most suitable to calculate diversity of molecules and the similarity of two groups of molecules.

References

1. T. Kotani, K. Higashiura, *J. Chem. Inf. Comput. Sci.* **42**, 58 (2002)
2. R.J. Zauhar, G. Moyna, L. Tian, Z. Li, W.J. Welsh, *J. Med. Chem.* **46**, 5674 (2003)
3. T.S. Rush, J.A. Grant, L. Mosyak, A. Nicholls, *J. Med. Chem.* **48**, 1489 (2005)
4. Y. Zyrianov, *J. Chem. Inf. Model.* **45**, 657 (2005)
5. V. Schnecke, J. Boström, *Drug Discov. Today* **11**, 43 (2006)
6. J.A. Wilson, A. Bender, T. Kaya, P.A. Clemons, *J. Chem. Inf. Model.* **49**, 2341 (2009)
7. Y.-S. Liu, Q. Li, G.-Q. Zheng, K. Ramani, W. Benjamin, *BMC Bioinformatics* **11**, 480 (2010)
8. D. Kihara, L. Sael, R. Chikhi, J. Esquivel-Rodriguez, *Curr. Prot. Pept. Sci.* **12**, 520 (2011)
9. A. Jennings, *Method. Mol. Biol.* **841**, 235 (2012)
10. A.S. Karaboga, F. Petronin, G. Marchetti, M. Souchet, B. Maigret, *J. Mol. Graph. Model.* **41**, 20 (2013)
11. L. Tarko, *Rev. Chim.* **45**, 395 (1994)
12. L. Tarko, *Rev. Chim.* **46**, 113 (1995)
13. L. Tarko, *Rev. Chim.* **47**, 238 (1996)
14. F. Berenger, A. Voet, X.Y. Lee, K.Y.J. Zhang, *J. Cheminform.* **6**, 23 (2014)
15. P.G. Mezey, *Shape in Chemistry: An Introduction to Molecular Shape and Topology* (Wiley VCH Publishers, New York, 1993)
16. J.A. Grant, B.T. Pickup, *J. Phys. Chem.* **99**, 3503 (1995)
17. P.G. Mezey, *J. Chem. Inf. Comput. Sci.* **36**, 1076 (1996)
18. N. Nikolova, J. Jaworska, *QSAR Comb. Sci.* **22**, 1006 (2003)
19. P.G. Mezey, C. Majdik, *Studia Universitatis Babes-Bolyai Chemia* **53**, 7 (2008)
20. P.G. Mezey, *J. Math. Chem.* **45**, 544 (2009)
21. R. Carbo-Dorca, P.G. Mezey, *Fundamentals of Molecular Similarity* (Springer, New York Inc, 2010)
22. A. Nicholls, G.B. McGaughey, R.P. Sheridan et al., *J. Med. Chem.* **53**, 3862 (2010)
23. P.G. Mezey, *J. Math. Chem.* **50**, 926 (2012)
24. Z. Antal, P.L. Warburton, P.G. Mezey, *Phys. Chem. Chem. Phys.* **16**, 918 (2014)
25. M. Hahn, *J. Chem. Inf. Comput. Sci.* **37**, 80 (1997)
26. P. Willett, *J. Med. Chem.* **48**, 4183 (2005)
27. A.C. Good, T.J. Ewing, D.A. Geschwend, I.D. Kuntz, *J. Comput. Aided Mol. Des.* **9**, 1 (1995)
28. M. Murcia, A. Morreale, A.R. Ortiz, *J. Med. Chem.* **49**, 6241 (2006)
29. A. Weber, M. Böhm, C.T. Supuran, A. Scozzafava, C.A. Sottriffer, G. Klebe, *J. Chem. Inf. Model.* **46**, 2737 (2006)
30. R. Huey, G.M. Morris, A.J. Olson, D.S. Goodsell, *J. Comput. Chem.* **28**, 1145 (2007)
31. T. Tucinardi, E. Nuti, G. Ortore, C.T. Supuran, A. Rossello, A. Martinelli, *J. Chem. Inf. Model.* **47**, 515 (2007)
32. A. Axenopoulos, P. Daras, G. Papadopoulos, E. Houstis, *Trans. Comput. Biol. Bioinform.* **8**, 1441 (2011)
33. L.H. Hall, L.B. Kier, *Molecular Connectivity in Chemistry and Drug Research* (Academic Press, Boston, 1976)
34. H. Timmerman, R. Todeschini, V. Consonni, R. Mannhold, H. Kubinyi, *Handbook of Molecular Descriptors* (Wiley-VCH, Weinheim, 2002)
35. H. González-Díaz, S. Vilar, L. Santana, E. Uriarte, *Curr. Top. Med. Chem.* **7**, 1015 (2007)
36. M.I. Trofimov, *J. Math. Chem.* **8**, 327 (1991)
37. D. Bonchev, O. Mekenyan, N. Trinajstić, *J. Comput. Chem.* **2**, 127 (1981)

38. J.C. Dearden, *J. Comput. Aided Mol. Des.* **17**, 119 (2003)
39. P.J. Ballester, W.G. Richards, *Proc. R. Soc. A* **463**, 1307 (2007)
40. E.O. Cannon, F. Nigsch, J.B.O. Mitchell, *Chem. Cent. J.* **2**, 3 (2008)
41. P.J. Ballester, W.G. Richards, *J. Comput. Chem.* **28**, 1711 (2007)
42. P.J. Ballester, *Future Med. Chem.* **3**, 65 (2011). doi:10.4155/fmc.10.280
43. P. J. Ballester, Patent US 8244483 B2, US 20090006395. 14 Aug 2012
44. J.J. Gajewski, K.E. Gilbert, PCModel; Serena Software, Box 3076, Bloomington, IN
45. <http://www.openmopac.net/>. Accessed Oct 2014
46. J.J.P. Stewart, *J. Mol. Model.* **13**, 1173 (2007)
47. M. Kendall, *Biometrika* **30**(1), 81 (1938)
48. T. Andrei, S. Stancu, *Statistica, Bucuresti, Ed. ALL*, p. 341 (1995)
49. C.E. Shannon, *Bell. Syst. Tech. J.* **27**, 379 (1948)
50. L. Tarko, *J. Math. Chem.* **52**, 948 (2014)
51. L. Tarko, *J. Math. Chem.* **49**, 2330 (2011)
52. M. Randic, *J. Am. Chem. Soc.* **97**, 6609 (1975)
53. H. Wiener, *J. Am. Chem. Soc.* **69**, 17 (1947)
54. A. Balaban, *Chem. Phys. Lett.* **89**, 399 (1982)
55. L.H. Hall, T.A. Vaughn, *Med. Chem. Res.* **7**, 407 (1997)
56. K. Roy, G. Gosh, *Int. Electron. J. Mol. Des.* **2**, 599 (2003)
57. R.L. Metcalf, T.R. Fukuto, *J. Agric. Food Chem.* **13**, 220 (1965)
58. R.L. Metcalf, T.R. Fukuto, C.F. Wilkinson, M.H. Fajmy, A.E. Azziz, E.R. Metcalf, *J. Agric. Food Chem.* **15**, 555 (1967)
59. R.L. Metcalf, T.R. Fukuto, *J. Agric. Food Chem.* **15**, 1022 (1967)
60. PRECLAV program is available from author
61. L. Tarko, C.T. Supuran, *Bioorg. Med. Chem.* **21**, 1404 (2013)