Orthosimilarity

Milan Randić

Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311

Received February 22, 1996[⊗]

Because of the interdependence of descriptors, the traditional approach to estimate the similarity for a set of structures is biased. We describe use of orthogonalized descriptors for determining the degree of similarity among molecules. The approach is illustrated for octane isomers when path numbers are used as molecular descriptors. Use of orthogonalized descriptors removed the high degree of the degeneracy that characterized the similarity/dissimilarity table for octane isomers. The method outlined in the article is general. It allows one to eliminate the inherent bias caused by the interdependence of molecular descriptors (mathematical or not). The approach equally applies to compounds viewed as three-dimensional objects characterized by invariants based on three-dimensional molecular data.

INTRODUCTION

Similarity plays an important role in the discussion of structure−property relationship, particularly in the discussions of biological activity of chemicals in chemistry.[1] Much of the work in the quantitative structure−activity relationship (QSAR) studies rests on the paradigm that similar structures have similar properties.

In quantitative applications of molecular similarity three distinct problems have to be addressed and resolved:[2,3]

(1) the choice of molecular descriptors
(2) the selection of the measure of similarity
(3) the choice of the clustering technique

Since early 1970s there was revived interest in the design of mathematical molecular descriptors, usually referred to as topological indices.[4] The topological indices have been used in simple regression analysis,[5] in multiple regression analysis,[6] in the PCA (the Principal Component Analysis),[7] in similarity studies,[8] for partial ordering,[9] for characterization of local features,[10] and even in graph isomorphism studies.[11]

The degree of similarity between molecule is usually derived from a characterization of a molecule by sequence $M = (M_1, M_2, M_3, ...)$. Hence, quantitative measure of similarity between molecules parallels closely the comparison of sequences. The most commonly used measure of similarity is the Euclidean distance:

$$D = \{(M_1 - M'_1)^2 + (M_2 - M'_2)^2 + (M_3 - M'_3)^2 + ...\}^{1/2}$$

where M and M' are viewed as vectors in n-dimensional space. Once the similarity/dissimilarity table is constructed one can extract the ranking of similar compounds using one of several available clustering patterns.[12] The result, the quantitative measure of similarity, depends on the choices made in each step of the analysis. Hence the first step, the selection of molecular descriptors adopted, is the most critical step. The result gives the similarity characterization of the structures as reflected by the chosen descriptors.

In this report we will consider the bias that is introduced in the standard quantitative approach to molecular similarity by hitherto overlooked effects of the inherent interdependence of molecular descriptors. The intercorrelation of the descriptors distorts the outcome of the similarity analyses. This is due to the fact that interdependent descriptors duplicate information. To illustrate how serious this problem is we will consider the 18 octane isomers and the molecular path numbers as the mathematical descriptors for the characterization of the molecules. It will become clear that the traditional similarity/dissimilarity studies are to a greater or lesser degree biased. A nonbiased approach requires use of orthogonal descriptors.[13] Because of considerable differences in the results of the traditional and the novel approach we will refer to the later as orthosimilarity. The octane isomers and the molecular paths serve here only to *illustrate* the problem. The approach is quite general and applies to other structures and other descriptors, including three-dimensional descriptors, mathematical or not.

PATH NUMBERS

It has been recognized for some time that shorter paths dominate many physicochemical molecular properties. In the case of isomeric variations, the paths of length two (p_2) and paths of length three (p_3) play the dominant role.[9] If, however, we are interested in similarity among molecules we should consider also the longer paths as they better discriminate among similar structures. Hence, we will use the sequence of path numbers: $P = (p_1, p_2, p_3, ...)$ as a characterization for each molecule. In Table 1 we list the path numbers for 18 isomers of octane shown in Figure 1. A characterization (i.e., representation of a structure by invariants) is typically associated with a loss of information. The path counts give only the information on the *number* of the neighbors certain distance away from the atom considered, not on the relative *distribution* of the neighbors. Thus the path count 1, 2, 2 represents two distinct distributions of the third neighbors:

The count of paths for a molecule is given simply by half of the sum of the paths count for all the atoms. It should not be surprising therefore that different molecules may have

[⊗] Abstract published in *Advance ACS Abstracts,* October 15, 1996.

S0095-2338(96)00021-2 CCC: $12.00

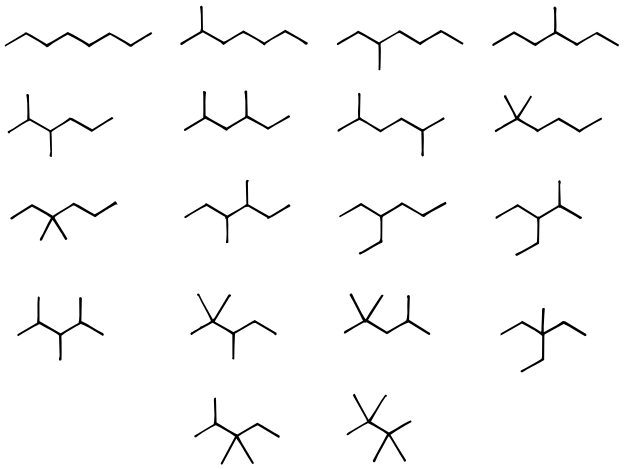
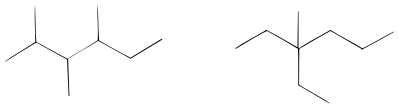


Figure 1. Carbon skeletons of the 18 isomers of octane.

Table 1. Path Numbers for the 18 Isomers of Octane Illustrated in Figure 1

		p_1	p_2	p_3	p_4	p_5	p_6	p_7
1	*n*-octane	7	6	5	4	3	2	1
2	2-M	7	7	5	4	3	2	0
3	3-M	7	7	6	4	3	1	0
4	4-M	7	7	6	5	2	1	0
5	3-E	7	7	7	5	2	0	0
6	2,2-MM	7	9	5	4	3	0	0
7	2,3-MM	7	8	7	4	2	0	0
8	2,4-MM	7	8	6	5	2	0	0
9	2,5-MM	7	8	5	4	4	0	0
10	3,3-MM	7	9	7	4	1	0	0
11	3,4-MM	7	8	8	4	1	0	0
12	2-M, 3-E	7	8	8	5	0	0	0
13	3-M, 3-E	7	9	9	3	0	0	0
14	2,2,3-MMM	7	10	8	3	0	0	0
15	2,2,4-MMM	7	10	5	6	0	0	0
16	2,3,3-MMM	7	10	9	2	0	0	0
17	2,3,4-MMM	7	9	8	4	0	0	0
18	2,2,3,3-MMMM	7	12	9	0	0	0	0

the same molecular path count. This occurs already for 2,3,4-trimethylhexane and 3-methyl-3-ethylhexane, two isomers of nonane[22] with the path count: 8, 10, 10, 6, 2.



It is not unusual that structures have several invariants identical.[14] Slater[15] and Balaban[16] have constructed pairs of nonisomorphic graphs that have all atomic path counts identical. It is generally believed that a final list of invariants need not be unique for a structure. In other words, for any given finite list of invariants one cannot exclude the possibility that there are two structures that will possess the same set of invariants. However, equally we can conjecture that given any pair of structures there will be at least a single invariant in which they will differ.

SIMILARITY BASED ON PATH NUMBERS

In order to derive a quantitative measure of molecular similarity we will use the paths of Table 1 as molecular descriptors. In Table 2 we show the similarity/dissimilarity table for the octane isomers derived when using the Euclidean distance as the measure of similarity. The small entries in the Table 2 indicate pairs of molecules found similar. The large entries in the Table 2 then point to the least similar pairs of isomers.

Table 2. Similarity/Dissimilarity Table for Octane Isomers Based on Path Numbers of Table 1

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0	1.414	2	2.449	3.464	3.742	3.464	3.162	4.690
2		0	1.414	2	3.162	2.828	3.162	2.828	2.449
3			0	1.414	2	2.449	2	2	2
4				0	1.414	2.828	2	1.414	2.828
5					0	3.162	1.414	1.414	1.414
6						0	2.449	2	2.414
7							0	1.414	2.828
8								0	2.449
9									0

	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	4.690	4.690	5.292	6.325	6.325	5.831	7.071	5.657	9.022
2	4	4.243	4.899	5.831	5.657	5.099	6.481	5.099	8.367
3	3.162	3.162	4	4.899	4.899	4.899	5.657	4.243	7.746
4	2.828	2.828	2.828	3.162	4.690	4.690	4	5.657	3.742
5	2.449	2	2.449	4	4.243	4.243	5.099	3.162	7.616
6	2.828	3.742	4.472	5.099	4.472	3.742	5.477	4.243	7.071
7	1.414	1.414	2.449	3.162	3.162	4	4	2.449	6.325
8	2	2.449	2.828	4.243	4	3.162	5.099	3.162	7.348
9	3.742	4.242	5.099	5.831	5.477	4.899	6.325	5.099	8

	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
10	0	1.414	2	2.449	2	3.162	3.162	1.414	5.477
11		0	1.414	2	2.449	4.243	3.162	1.414	5.831
12			0	2.449	2.828	3.742	3.742	1.414	6.481
13				0	1.414	5.099	1.414	1.414	4.243
14					0	4.243	1.414	1.414	3.742
15						0	5.657	3.742	7.483
16							0	2.449	2.828
17								0	5.099
18									0

The least similar structures can often be identified by an inspection. For example, *n*-octane and 2,2,3,3-tetramethylbutane appear as the least similar among the 18 isomers of octane. However, even to identify the most different molecules, particularly most different conformations, may often be a challenging problem.[17] To select the most similar pair of structures is generally even more difficult. Is *n*-octane and 2-methylheptane the most similar pair of isomers? Or should 2-methyloctane and 3-methylheptane be considered as the most similar? Or is there yet another pair that is even more similar?

The smallest entry in the similarity/dissimilarity table is 1.414, which appears whenever two path sequences differ in a single entry by one. From Table 2 we see that within the set of 18 isomers of octane somewhat disappointingly there are 20 pairs of isomers that qualify as the most similar. Moreover, similar degeneracy occurs also for intermediate values of the similarity/dissimilarity. The values 3.162, 2.000, and 2.828 appear in Table 2 a dozen times and more. Hence, the path numbers, even though they discriminate the individual isomers, do not discriminate sufficiently well among the pairs of isomers. One is prone to assume that this lack of discrimination among pairs of isomers is due to inherent limitations of the path numbers as molecular descriptors. But is this so? We will see that the high degree of the degeneracy observed in the similarity/dissimilarity table for octanes is an artifact of the interdependence of the path numbers as descriptors.

One way to narrow down the list of the most similar structures is to use additional descriptors for the characteriza-

Table 3. Correlation Matrix between Different Path Numbers

	p_2	p_3	p_4	p_5	p_6	p_7
p_2	1.000	0.543	−0.610	−0.685	−0.625	−0.417
p_3	0.543	1.000	−0.601	−0.783	−0.513	−0.304
p_4	−0.610	−0.601	1.000	0.286	0.108	0.021
p_5	−0.685	−0.783	0.286	1.000	0.517	0.281
p_6	−0.625	−0.513	0.108	0.517	1.000	0.606
p_7	−0.417	−0.304	0.021	0.281	0.606	1.000

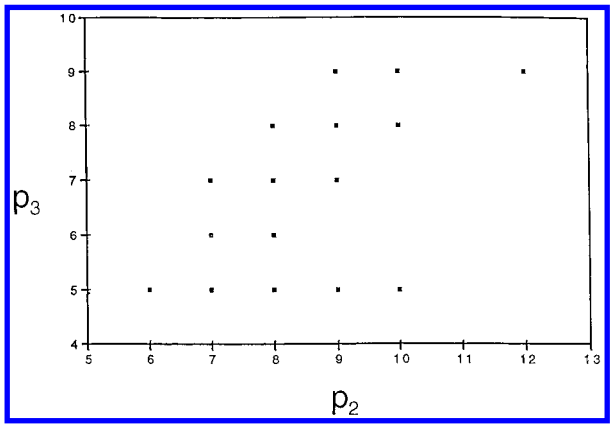
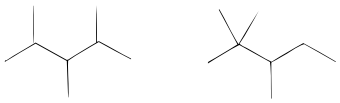


Figure 2. The interrelation of p_2 and p_3 for octane isomers.

tion of the molecules. For example, when the count of paths is extended to the count of disjoint paths we obtain a characterization of octane isomers that leads to a considerable reduction of the degeneracy observed earlier in the similarity/dissimilarity table.[18] Then as the most similar pair of octane isomers emerges 2,3,4-trimethylpentane and 2,2,3-trimethylpentane.



Are the path numbers inherently limited as molecular descriptors when studying molecular similarity? As will be seen, the answer is "No." We will see that the degeneracy in the similarity/dissimilarity table for the path numbers is a consequence of the *bias* caused by the high interdependence of the path numbers. In Table 3 we show the correlation matrix for the path numbers for the 18 isomers of octane. The table clearly points to the contamination of the structural information each of the descriptors is carrying by the other descriptors. The interdependence of path numbers is reflected in that often p_2 parallels p_3. This effectively gives to p_2 a greater weight than to p_3. In Figure 2 we show the simple regression of p_3 against p_2. Although the correlation coefficients is low, the pattern clearly indicates the relatedness of the two descriptors. In fact, the pattern seen in Figure 2 underlines the partial ordering of octane isomers and has lead to a template for the "periodic table of isomers", on which one can display many regularities in isomeric variations of molecular properties.[9,19]

INTERDEPENDENCE OF PATH NUMBERS

The degeneracy in the similarity table can be reduced *without* introducing additional descriptors, or different weights for the bonds of different type, as will be outlined here. One way of doing this is to follow the general procedure for comparison of sequences. Scottish mathematician Muirhead[20] already at the beginning of the century outlined a scheme that allows comparisons of sequences by considering their partial sums. Hence, instead of considering the original sequences $P = (p_1, p_2, p_3, ...)$ one considers the sequences of the partial sums $S = (p_1, p_1 + p_2, p_1 + p_2 + p_3, ...)$. This procedure tacitly increases the weight of the leading members of the sequence (short paths), since when evaluating the Euclidean distance for the higher members of the sequences the differences among the leading members for two sequences are augmented by their repetitive use. The process can be even further generalized by considering the partial sums of so derived novel sequences of partial sums.[21]

In Table 4 we listed the sequences of partial sums for the 18 isomers of octane, and in Table 5 we show the revised similarity/dissimilarity table based on the characterization

Table 4. Sequences of the Partial Sums for the Path Numbers of Table 1

		s_1	s_2	s_3	s_4	s_5	s_6	s_7
1	*n*-octane	7	13	18	22	25	27	28
2	2-M	7	14	19	23	26	28	28
3	3-M	7	14	20	24	27	28	28
4	4-M	7	14	20	25	27	28	28
5	3-E	7	14	21	26	28	28	28
6	2,2-MM	7	16	21	25	28	28	28
7	2,3-MM	7	15	22	26	28	28	28
8	2,4-MM	7	15	21	26	28	28	28
9	2,5-MM	7	15	20	24	28	28	28
10	3,3-MM	7	16	23	27	28	28	28
11	3,4-MM	7	15	23	27	28	28	28
12	2-M, 3-E	7	15	23	28	28	28	28
13	3-M, 3-E	7	16	25	28	28	28	28
14	2,2,3-MMM	7	17	25	28	28	28	28
15	2,2,4-MMM	7	17	22	28	28	28	28
16	2,3,3-MMM	7	17	26	28	28	28	28
17	2,3,4-MMM	7	16	24	28	28	28	28
18	2,2,3,3-MMMM	7	19	28	28	28	28	28

Table 5. Revised Similarity/Dissimilarity Table for Octane Isomers Based on Partial Sums of Table 4

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0	2.236	3.742	4.359	6	6.083	6.782	6.245	4.690
2		0	1.732	2.449	4.123	4	4.796	4.243	2.646
3			0	1	2.449	2.646	3.162	2.646	1.414
4				0	1.732	2.449	2.646	2	1.732
5					0	2.236	1.414	1	2.449
6						0	1	2.828	1.732
7							0	2.236	2.449
8								0	4.359
9									0

	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	8.307	8	8.660	10.20	10.54	8.832	11.22	9.539	13.49
2	6.325	6.083	6.782	8.307	8.602	6.856	9.327	7.616	11.62
3	4.796	4.472	5.196	6.782	7.141	5.477	7.874	6.083	10.30
4	4.243	3.873	4.472	6.245	6.633	4.796	7.416	5.477	9.950
5	3	2.449	3	4.899	5.385	3.742	6.164	4.123	8.832
6	2.828	3	3.742	5	5.099	3.317	5.916	4.243	8.185
7	1.732	1.414	2.236	3.742	4.123	2.828	4.899	3	7.483
8	2.449	2.236	2.828	4.583	4.899	3	5.745	3.742	8.307
9	4.359	4.243	5	6.481	6.708	4.899	7.483	5.745	9.798

	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
10	0	1	1.414	2.236	2.449	1.732	3.137	1.414	5.916
11		0	1	2.449	3	2.449	3.742	1.732	6.481
12			0	2.236	2.828	2.236	3.606	1.414	6.403
13				0	1	3.162	1.414	1	4.243
14					0	3	1	1.414	3.606
15						0	4	2.236	6.326
16							0	2.236	2.828
17								0	5
18									0

Table 6. Orthogonalize Path Numbers

	Ω_0	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5
1	8.4444	−2.4444	−0.4695	−1.3954	−1.0780	−0.3765
2	8.4444	−1.4444	−1.0274	−1.0308	−0.52034	0.3932
3	8.4444	−1.4444	−0.0274	−0.6952	0.2075	0.1628
4	8.4444	−1.4444	−0.0274	0.3048	−0.2174	0.1804
5	8.4444	−1.4444	0.9726	0.6404	0.5104	−0.0500
6	8.4444	+0.5556	−2.1433	−0.3016	0.5949	−0.0674
7	8.4444	−0.4444	0.4146	0.0050	0.4929	−0.0463
8	8.4444	−0.4444	−0.5854	0.6694	0.3402	−0.0499
9	8.4444	−0.4444	−1.5854	−0.6662	1.0372	−0.0887
10	8.4444	+0.5556	−0.1433	0.3696	0.0505	−0.0250
11	8.4444	−0.4444	1.4146	0.3406	0.2208	−0.0251
12	8.4444	−0.4444	1.4146	1.3406	−0.2041	−0.0075
13	8.4444	+0.5546	1.8567	0.0408	−0.0689	−0.0001
14	8.4444	+1.5546	0.2988	0.0698	−0.2391	0.0000
15	8.4444	+1.5546	−2.7012	2.0630	−0.6974	−0.0109
16	8.4444	+1.5546	1.2988	−0.5947	−0.0863	0.0036
17	8.4444	+0.5546	0.8567	0.7052	−0.2216	−0.0038
18	8.4444	+3.5546	0.1829	−1.8654	−0.1213	0.0111

of octanes given by Table 4. Clearly we lifted most of the degeneracy observed in Table 2, that is, most of the coincidental entries of Table 2 disappeared. While we have thus on one side resolved the problem of the high degeneracy of the similarity/dissimilarity table for octanes, on the other side we have even more obscured the role of the individual descriptors used in the characterization of the molecules.

We would like to propose a procedure that does just the opposite: It eliminates in the calculations of the Euclidean distances the repeated occurrence of the contributions of the initial members (short paths) when calculating the contributions from subsequent members (longer paths). This is accomplished by use of orthogonalized path numbers.[13] The orthogonalized descriptors are constructed in the following way: The first descriptor, here p_2 (since p_1 is the same for all isomers) is selected as the first orthogonal descriptor. Instead of p_2 we, without a loss of the generality, can use as Ω_1 the mean deviation of p_2 from the average p_2 for the whole set (which is 8.4444, shown in Table 6 as Ω_0). Hence, $p_2 = \Omega_0 + \Omega_1$. To obtain the second descriptor, here orthogonalized p_3, we first consider the regression of p_3 against p_2 (already shown in Figure 2). The residual of that regression, listed in Table 6 as Ω_2, is taken as the second orthogonal descriptor. Clearly, the part of p_3 that correlates with p_2 can be evaluated from the regression, hence it is redundant. However, the part of p_3 that cannot be determined from p_2, the residual, contains the information independent of p_2. Hence, the residual of the correlation between the two descriptors, by definition, is the second orthogonal descriptor. Clearly, by definition, the correlation of the residual p_3/p_2 and p_2 is zero. The process of orthogonalization continues by considering the regression of p_4 against p_2, which gives the residual p_4/p_2, the part of the original p_4 count that does not parallel p_2. The residual p_4/p_2, however, may show a correlation with the second orthogonal descriptor Ω_2 or the residual p_3/p_2. We have therefore to regress the residual p_4/p_2 against the residual p_3/p_2 and use the residual of that correlation, i.e., the residual of the correlation of the residuals, as our third orthogonal descriptor Ω_3. Alternatively, as has been recently outlined by Šoškić *et al.*,[22] one can obtain directly the *n*th orthogonal descriptor as the residual in a multiple regression of *n*th descriptor as dependent variable against $(n-1)$ descriptors.

As we can see the magnitudes of the orthogonalized path in Table 6 decrease after each successive step of the orthogonalization, so that in the column that would correspond to Ω_6 we would have all the entries being zero. Apparently the paths of increasing length contribute gradually less and less novel information. In the case of octanes the paths $p_2 - p_6$ (or their orthogonal counterparts) practically have all the information that is relevant for comparison of these structures. This is not surprising since except for a single molecule, *n*-octane, p_7 is equal (zero) for all the other molecules.

Table 7. Similarity/Dissimilarity Table for Octane Isomers Based on Orthogonalized Path Numbers of Table 6

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0	1.532	1.905	2.266	3.139	3.986	3.053	3.224	3.215
2		0	1.302	1.709	2.838	2.689	2.318	2.241	2.025
3			0	1.087	1.709	2.972	1.346	1.799	2.044
4				0	1.302	3.093	1.357	1.345	2.453
5					0	3.821	1.310	1.859	3.087
6						0	2.765	2.106	1.281
7							0	1.210	2.179
8								0	1.809
9									0

	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	3.690	3.518	3.992	4.199	4.425	5.764	4.571	4.005	6.141
2	2.691	3.093	3.584	3.718	3.493	4.644	3.865	3.288	5.241
3	2.282	2.046	2.724	2.863	3.149	4.960	3.299	2.637	5.152
4	2.033	1.821	2.046	2.770	3.032	4.417	3.408	2.231	5.458
5	2.352	1.170	1.483	2.341	3.216	5.097	3.315	2.134	5.684
6	2.179	3.770	4.123	4.070	2.793	2.928	3.661	3.269	4.168
7	1.281	1.090	1.809	1.844	2.134	4.400	2.341	1.483	4.465
8	1.170	2.031	2.179	2.744	2.341	3.391	3.055	1.844	4.820
9	2.265	3.269	3.818	3.818	3.119	3.963	3.687	3.231	4.682

	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
10	0	1.859	2.106	2.031	1.170	3.312	2.007	1.090	3.759
11		0	1.087	1.170	2.352	4.975	2.232	1.281	4.744
12			0	1.704	2.619	4.659	2.788	1.310	5.273
13				0	1.859	5.124	1.310	1.210	3.929
14					0	3.631	1.210	1.310	2.788
15						0	4.841	3.966	5.299
16							0	1.704	2.619
17								0	4.009
18									0

ORTHOSIMILARITY

If we use the orthogonalized path numbers form Table 6 and construct the similarity/dissimilarity table we obtain an *unbiased* measure of the degree of similarity among the isomers of octane. The orthogonalization procedure requires a prior *ordering* of descriptors. The same is true for the orthogonalization of vectors in the Gram−Schmidt orthgonalization scheme. Just as in the case of vectors, different ordering will therefore results in different basis for representing structures. This adds some flexibility to the scheme since in different applications one can use different ordering of the descriptors.[13,23,24]

We have assumed the "natural" ordering of path numbers according to the length of the paths involved. We use the term "unbiased" to signify that all the descriptors used in calculating the similarity/dissimilarity are unrelated, hence no duplication of information between descriptors is involved. However, one can influence the outcome of the similarity/dissimilarity, or the regression equation, by changing the order in which the descriptors are orthogonalized. Moreover, one can use descriptors that have been made orthogonal to descriptors not used in the regression (or comparison) and in this way one can "compact" information

Table 8. Most Similar Pairs of Octane Isomers as Given by the Unbiased Similarity/Dissimilarity Table

rank			magnitude
1−2	3,4-dimethylhexane	2-methyl-3-ethylpentane	1.086 67
1−2	3-methylheptane	4-methylheptane	1.086 67
3−4	2,3-dimethylhexane	3,4-dimethylhexane	1.089 55
3−4	3,3-dimethylhexane	2,3,4-trimethylpentane	1.089 55
5	3-ethylhexane	3,4-dimethylhexane	1.170 38
6−9	3,3-dimethylhexane	2,2,3-trimethylpentane	1.170 39
6−9	3,4-dimethylhexane	3-methyl-3-ethylpentane	1.170 39
6−9	2,4-dimethylhexane	3,3-dimethylhexane	1.170 39

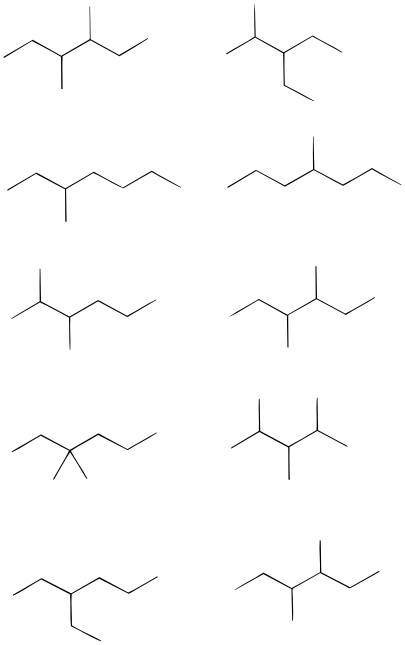


Figure 3. The most similar pairs of octane isomers based on orthogonal descriptors.

Table 9. Least Similar Pairs of Octane Isomers According to the Unbiased Similarity/Dissimilarity Table

rank			magnitude
1	*n*-octane	2,2,3,3-tetramethylbutane	6.141 01
2	*n*-octane	2,2,4-trimethylpentane	5.763 64
3	3-ethylhexane	2,2,3,3-tetramethylbutane	5.683 79
4	4-methylheptane	2,2,3,3-tetramethylbutane	5.458 22
5	2,2,4-trimethylpentane	2,2,3,3-tetramethylbutane	5.299 40
6	2-methyl-3-ethylpentane	2,2,3,3-tetramethylbutane	5.272 84
7	2-methylheptane	2,2,3,3-tetramethylbutane	5.240 89
8	3-methylheptane	2,2,3,3-tetramethylbutane	5.152 17
9	3-methyl-3-ethylpentane	2,2,4-trimethylpentane	5.124 40
10	3-ethylhexane	2,2,4-trimethylpentane	5.097 14

from several descriptors in few that are more effective for the problem considered.[25]

In Table 7 we show the similarity/dissimilarity matrix based on orthogonal descriptors of Table 6. Observe how the degeneracy of the similarity/dissimilarity table for path numbers so dominant in Table 2 has almost completely disappeared. The most similar pairs of isomers are listed in Table 8 and illustrated in Figure 3. Among these are the most similar the two pair of isomers that have the same both p_2 and p_3. These are the only pairs of octane isomers that have common sites on the (p_2, p_3) coordinate grid. This result should have been expected, since the role of longer path is decreasing in each successive orthogonalization step. The least similar octane isomers are listed in Table 9 and illustrated in Figure 4. That we found 2,2,3,3-tetramethylbutane to be the least similar to *n*-octane and other "long" isomers (2-methyl-, 3-methyl-, and 4-methylheptane) is not surprising. It is, however, surprising to see among the least similar the pair 2,2,4-trimethylpentane, 2,2,3,3-tetramethylbutane and the pair 2-methyl-3-ethylpentane, 2,2,3,3-tetramethylbutane. A visual inspection would fail to indicate this pair among the least similar, yet there is no doubt that molecules show large dissimilarities when their paths are compared. Indeed, even if we compare the dissimilarity based on nonorthogonal paths for the pair 2,2,4-trimethylpentane, 2,2,3,3-tetramethylbutane we see that they are accompanied by a large entry in the Table 2. That visual inspection can be misleading is well illustrated by the trio: *n*-octane (A), 2,2,4-trimethylpentane (B), and 2,2,3,3-tetramethylbutane (C). The dissimilarity table for these three isomers is

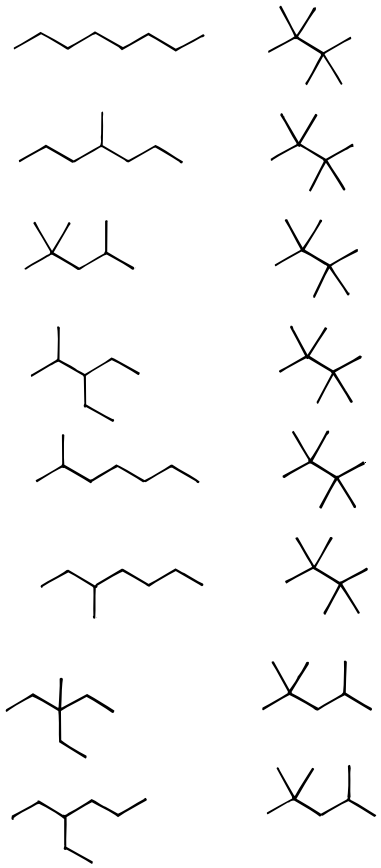


Figure 4. The least similar pairs of octane isomers based on orthogonal descriptors.

	A	B	C
A		5.764	6.141
B			5.299
C			

That A is very dissimilar from B and C agrees with our intuition, A is very "long," the other two isomers are highly branched structures. But similarity is not transitive property, and by A being so dissimilar from B and C it does not make B and C necessarily similar! As we see, B and C are almost as different among themselves as they differ from non branched *n*-octane.

CONCLUDING REMARKS

We have seen in this paper how interdependence among descriptors used to characterize structures can totally obscure

the outcome of the similarity/dissimilarity studies. Hence, besides generally recognized factors (1) the choice of descriptors, (2) the choice of similarity measure, and (3) the choice of clustering of data one ought to include in the analysis, also, orthogonalization of the descriptors, since as a rule molecular descriptors are intercorrelated, often highly intercorrelated. Orthosimilarity, i.e., the similarity based on orthogonalized descriptors, favors no descriptor and shows no bias toward any of the descriptors used. If in specific application one wants to use descriptors that are intercorrelated one can do this equally well, starting with orthogonal descriptors. Moreover, in that case one is fully in charge to choose the weights that may suit the particular application rather than be left at the mercy of the inherent interdependence of the descriptors of which one has no control.

ACKNOWLEDGMENT

The author thanks Professor Ch. Rucker (University of Freiburg, Germany) for examining the manuscript and suggesting many improvements in the presentation of the material.

REFERENCES AND NOTES

(1) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G., Eds.; Wiley: New York, 1990.
(2) Randić, M. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G., Eds.; Wiley: New York, 1990; pp 77−145.
(3) Randić, M. In *Mathematical Methods in Contemporary Chemistry*; Kuchanov, S., Ed.; Gordon & Breach: New York, 1996.
(4) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1992; Chapter 10, pp 225−273.
(5) Randić, M. Comparative regression analysis. Regressions based on a single descriptor. *Croat. Chem. Acta* **1993**, *66*, 289−312.
(6) Randić, M.; Trinajstić, N. Viewpoint 4 - Comparative structure−property studies: the connectivity basis. *J. Mol. Struct. (Theochem)* **1993**, *284*, 209−221. Needham, D. E. ; Wei, I. C; Seybold, P. G. Molecular modeling of the physical properties of alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186−4194.
(7) Basak, S. C.; Grunwald, G. D. Molecular similarity and risk assessment: analog selection and property estimation using graph invariants, SAR & QSAR. **1994**, *2*, 289−307. Randić, M. Similarity based on extended basis descriptors. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 686−692.
(8) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Application of graph theoretical parameters in quantifying molecular similarity and structure−activity studies. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 270−276. Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining structural similarity of chemicals using graph theoretic indices. *Discrete Appl. Math.* **1988**, *19*, 17−44. Hall, L. H.; Kier, L. B.; Brown, B. B. Molecular similarity based on novel atom-type electrotopological state indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1074−1080.
(9) Randić, M.; Wilkins, C. L. On graph theoretical basis for ordering of structures. *Chem. Phys. Lett.* **1979**, *63*, 332−336. Randić, M.; Wilkins, C. L. Graph theoretical ordering of structures as a basis for systematic searches for regularities in molecular data. *J. Phys. Chem.* **1979**, *83*, 1525−1540.
(10) Randić, M. Graph theoretical approach to structure-activity studies: search for optimal antitumor compounds. In Molecular Basis of Cancer, Part A: Macromolecular Structure, Carcinogens, and Oncogens; Rein, R., Ed.; Alan R. Liss: 1985; pp 309−318. Randić, M.; Jerman-Blažič, B.; Rouvray D. H.; Seybold, P. G.; Grossman, S. C. The search for active substructures in structure−activity studies. *Int. J. Quant. Chem.: Quant. Biol. Symp.* **1987**, *14*, 245−260. Randić, M.; Jurs, P. On a fragment approach to structure−activity correlations. *Quant. Struct.-Act. Relat.* **1989**, *8*, 39−48.
(11) Balaban, A. T.; Liu, X.; Klein, D. J., Babić, D.; Schmalz, T. G.; Seitz, W. A.; Randić, M. Graph invariants of fullerenes. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 396−404. Balaban, A. T. Lowering the intra- and intermolecular degeneracy of topological invariants. *Croat. Chem. Acta* **1993**, *66*, 447−458.
(12) Hartigan, J. A. *Clustering algorithms*; Wiley: New York, 1975.
(13) Randić, M. Orthogonal molecular descriptors. *New J. Chem.* **1991**, *15*, 517−525. Randić, M. Resolution of ambiguities in structure−property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311−320. Randić, M. Fitting of nonlinear regressions by orthogonalized power series. *J. Comput. Chem.* **1993**, *14*, 363−370. Randić, M. Curve-fitting paradox. *Int. J. Quant. Chem: Quant. Biol. Symp.* **1994**, *21*, 215−225.
(14) Quintas, L. V; Slater, P. J. Pairs of non-isomorphic graphs having the same path degree sequence. MATCH **1981**, *12*, 75−86. Razinger, M.; Chretian, J. R.; Dubois, J. E. Structural selectivity of topological indexes in alkane series, *J. Chem. Inf. Comput. Sci*. **1985**, *25*, 23. Szymanski, K.; Muller, W. R.; Knop, J.; Trinajstić, N. On Randić's molecular identification numbers. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 413−415. Szymanski, K.; Muller, W. R.; Knop, J.; Trinajstić, N. Molecular ID numbers. *Croat. Chem. Acta* **1986**, *59*, 719−723. Chalcraft, D. A. *J. Graph Theory* **1990**, *14*, 314. Ivanciouc, O.; Balaban, A. T. Nonisomorphic graphs with identical counts of self-returning walks: Isocodal graphs. *J. Math. Chem.* **1992**, *11*, 155−168. Hosoya, H.; Nagashima, U.; Hyugaji, S. Topological twin graphs. Smallest pair of isospectral polyhedral graphs with eight vertices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 428−431.
(15) Slater, P. J. *Graph Theory* **1982**, *6*, 89.
(16) Balaban, A. T.; Quintas, L. V. The smallest graphs, trees, and 4-trees with degenerate topological index J. MATCH **1983**, *14*, 213−233.
(17) Leach, A. R. An algorithm to directly identify a molecule's "most different" conformations. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 661−670.
(18) Randić, M. On the representation of molecular graphs by basis graphs. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 57−69.
(19) Randić, M.; Wilkins, C. L. Graph theoretical analysis of molecular properties. Isomeric variations in nonanes. *Int. J. Quant. Chem.* **1980**, 18, 1005−1027. Randić, M. Survey of structural regularities in molecular properties. I. Carbon-13 chemical shifts in alkanes. *Int. J. Quant. Chem.* **1983**, *23*, 1707−1722. Randić, M. Chemical structure - What is "She"? *J. Chem. Educ.* **1992**, *69*, 713−718. Randić, M.; Trinajstić, N. Isomeric variation in alkanes: boiling points of nonanes. *New J. Chem.* **1994**, *18*, 179−189.
(20) Muirhead, R. F. *Proc. Edinburgh Math. Soc.* **1901**, *19*, 36; **1903**, *21*, 144; **1906**, *24*, 45. Karamata, J. *Publ. Math., Univ. Belgrade* **1932**, *1*, 145. Beckenbach, E. F.; Bellman, R. *Inequalities*; Springer: Berlin, 1961.
(21) Randić, M. On comparability of structures. *Chem. Phys. Lett.* **1978**, *55*, 547−551.
(22) Šoškić, M.; Plavšić, D.; Trinajstić, N. 2-Difluoromethylthio-4,6-bis-(monoalkylamino)-1,3,5-triazines as inhibotors of Hill reaction: A QSAR study with orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 146−150.
(23) Lucić, B.; Nikolić, S.; Trinajstić, N.; Juretic, D. The structure−property models can be improved using orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532−538. Amic, D.; Davidovic-Amic, D.; Juric, A.; Lucić, B.; Trinajstić, N. Structure−activity correlation of flavone derivatives for inhibition of cAMP phosphodiesterase. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1034−1038.
(24) Randić, M. Compact QSAR descriptors. *Acta Pharm*. Submitted for publication.

CI9600216